Research Reports in Public Health Sciences | 2021:1

An applied guide to quantitative methods with Stata

Ylva B Almquist, Christoffer Åkesson & Lars Brännström

Department of Public Health Sciences

Stockholm
University

An applied guide to quantitative methods with Stata

Ylva B Almquist, Christoffer Åkesson &
Lars Brännström

Research Reports in Public Health Sciences 2021:1

ISBN 978-91-89107-03-8

ISSN 2003-0142

Cite this report:
Almquist, Y. B., Åkesson, C., Brännström L. (2021). An applied guide to quantitative methods with Stata. *Research Reports in Public Health Sciences*, Stockholm University, no:2021:1.

TABLE OF CONTENTS

1. INTRODUCTION

The purpose of this guide is to provide both a basic understanding of statistical concepts (know-why) as well as instructions for analysing quantitative data in Stata (know-how). The assumption is that you already have some data – therefore, there is only a limited discussion about study design and all the issues related to this. It should also be noted that this guide is positioned somewhere in the intersection between social sciences and medical sciences.

We wanted to keep the guide almost completely free of formulas (i.e. brain-freezing mathematical equations). In doing so, we have tried to explain everything at the most elementary level and only included aspects that we think are important for applied statistics. As such, this guide is pragmatic and research-oriented. Hopefully, you will find it useful.

This guide consists of three parts. In the first part, we introduce the guide (Chapter 1) and the Stata environment (Chapter 2), after which we discuss basic statistical concepts (Chapter 3) and different types of descriptive analysis (Chapter 4).

The second part starts with issues related to statistical significance (Chapter 5) and then continues with basic types of analysis, such as t-tests, ANOVA, chi-square, correlation analysis, and factor analysis (Chapters 6-8).

In the third part, we focus on more advanced statistical analysis: we discuss some theoretical dimensions of statistical analysis (Chapter 9) and briefly explore different extensions of ANOVA (Chapter 10), before continuing to regression analysis (Chapters 11-17). In addition to this, we guide you through mediation analysis (Chapter 18) and interaction analysis (Chapter 19).

You are welcome to contact ylva.almquist@su.se with any questions or suggestions for improvements or additions.

Happy exploration! And remember these words (Aaliyah, 2000):

> *If at first you don't succeed,*
> *Dust yourself off, and try again.*

The Authors
Stockholm, February 2021

Contributions and credits

The original idea behind this guide as well as its overall structure and content was conceived by Ylva B Almquist and Lars Brännström. Ylva B Almquist has authored Chapters 1-3, 5-8, 11-16, and 18-19. Lars Brännström has authored Chapter 9. Christoffer Åkesson has proof-read the guide, authored Chapter 10, and provided theoretical examples throughout Chapters 11-16 and 18-19.

The authors are also very grateful to Lauren Bishop who co-authored Chapters 4 and 17 together with Ylva B Almquist.

Data

The primary data material used in this guide is:

- “StataData1”, which is a dataset developed for the purposes of the guide. It contains data on a fictional cohort of 10,000 individuals born in 1970, who have been followed-up until 2020. In other words, the data are fake.

In Chapter 8, another dataset is used:

- “StataData2”, which is based on several waves of data collection related to the World Values Survey.

Additionally, a dataset called TestData1 will be created, and one called TestData2 will be used, in Chapter 2.

Versions

This version of the guide (1.1) is based on Stata/SE 16.1 for Windows.

The joys of do-files

Nowadays, Stata’s interface is quite user-friendly and, if you want, you can do most things through the menus. This is, however, not what we will practice in this guide; instead, we will learn how to use syntax. Syntax is basically how we tell Stata to do what we want it to do. We write our syntax in a so-called do-file, which is a text editor where it is possible to add comments and commands (see Section 2.1.2).

The reasons for why we strongly recommend everyone to use do-files are, among other things:

- It is a way of documenting and archiving everything you have done with the data material.
- It is easy to repeat parts or all of the analysis.
- Other people involved in the data material can easily understand what you have done and how you have done it.
- It saves an enormous amount of time.

File paths

To make full use of do-files, you need to be able to specify file paths. A file path is a specification of the location of a file on your computer. There are many ways of identifying a file path, of which a few will be outline below.

Note The following instructions are based on Windows.

A first option is to open the File Explorer and locate the file you want. Right click on the file and choose Properties. In the new window that opens, look for the line that says Location. Select the path that is specified here, right-click, and choose Copy. It could look something like this:

C:\Users\yerik\Stata Guide

Paste the path into your do-file. Note that you also need to add the file name (including the file extension) to the path. It could look something like this:

C:\Users\yerik\Stata Guide\StataData1.dta

Another option is to open the File Explorer and locate the file you want. In the upper area of the window, there is a search bar. Place the mouse pointer on the right-hand side of the path specification here and click. The path is now selected. Right-click and choose Copy. The subsequent instructions are then the same as for the first option.

General advice

- Keep all your files for the course/project in the same main folder and use sub folders to organise the files further.
- Save your files under appropriate names.
  Example: “Ericsson Data Lesson 1”
- Always keep a copy of the original files.
  Example: “Ericsson Data Lesson 1 Original”
- Save intermediate versions of your files.
  Example: “Ericsson Data Lesson 1 200629”
- Always double-check that you have spelled variable names, values, and labels correctly.

Note Stata is case-sensitive! Moreover, you must always use lower-case letters for your commands.

2. THE STATA ENVIRONMENT

Outline

Content

The Stata environment may come across as daunting at first, but it is actually quite simple once you have had some practice. In this chapter, we will review main file types and the most useful commands. It is far from being exhaustive – there is much more to explore!

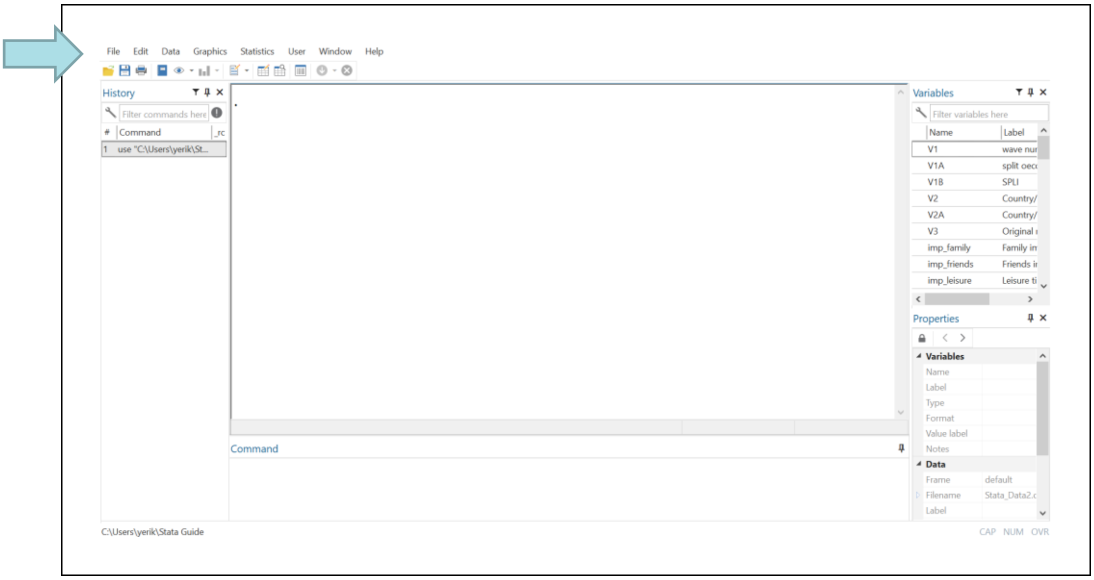
2.1 File types

The are many different types of Stata files, all with different functions and extensions. The most common ones are:

Type	**Extension**	**Content**
Dataset	.dta	Data and variables
Do-file	.do	Commands and comments
Log	.smcl	Results
Graph	.gph	Graphs
Package	.ado	User-written packages

2.1.1 Dataset

This is what Stata looks like. The menu bar (File, Edit, and so on) is located in the upper area.

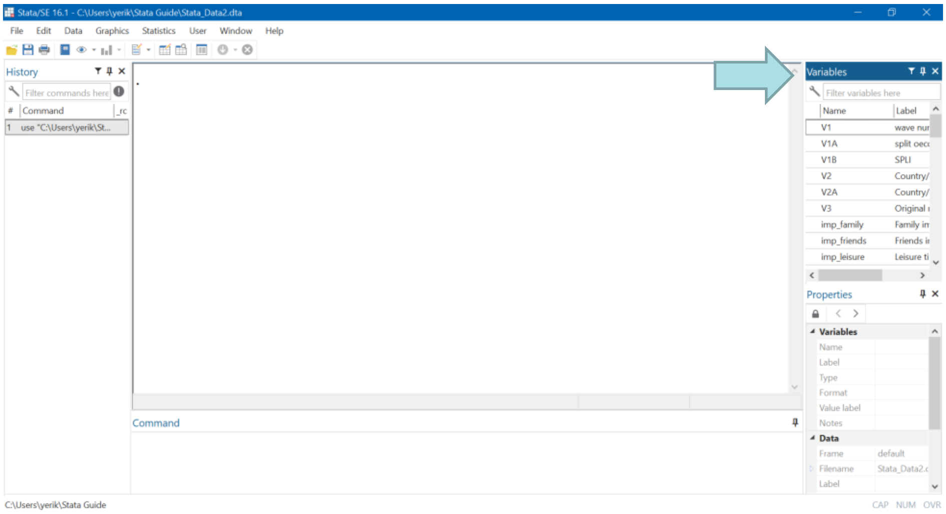


The Stata menu works similar to the menus in many other programs, such as Word or Excel. Some useful File options are listed below (most of these can be found also in the menu of e.g. do-files):

Option	Description
Open a file	Go to File\Open and browse your computer to locate the file that you want to open. Then click on Open.
	Keyboard shortcut: Ctrl+O
Save a file	Go to File\Save As. Type in a descriptive name and then click Save.
	Keyboard shortcut: Ctrl+Shift+S
Overwrite a file	Go to File\Save.
	Keyboard shortcut: Ctrl+S
Import data	Go to File\Import and choose what kind of format you want to import. Browse the file and click on Open.
Export data	Go to File\Export and choose what kind of format you want to export to, and what to call the new file. Then click on OK.

Variables

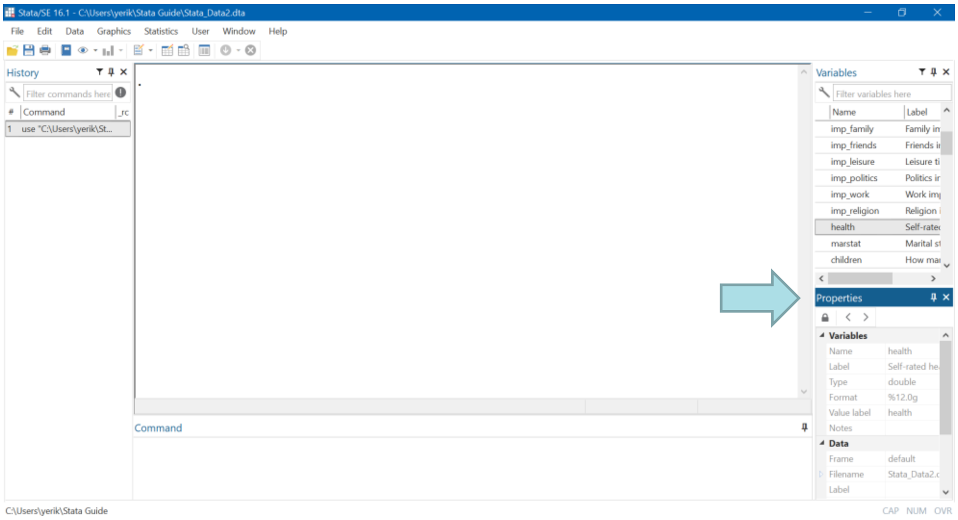
The Variables window shows a list of variables in the dataset, including the Name and the Label for each variable:



```
1   use "C:\Users\yerik\St...
```

Properties

Every time that you click on a variable in the list, some information for that variable appears in the Properties window, e.g. Name, Label, Type, Format, and Value Label:

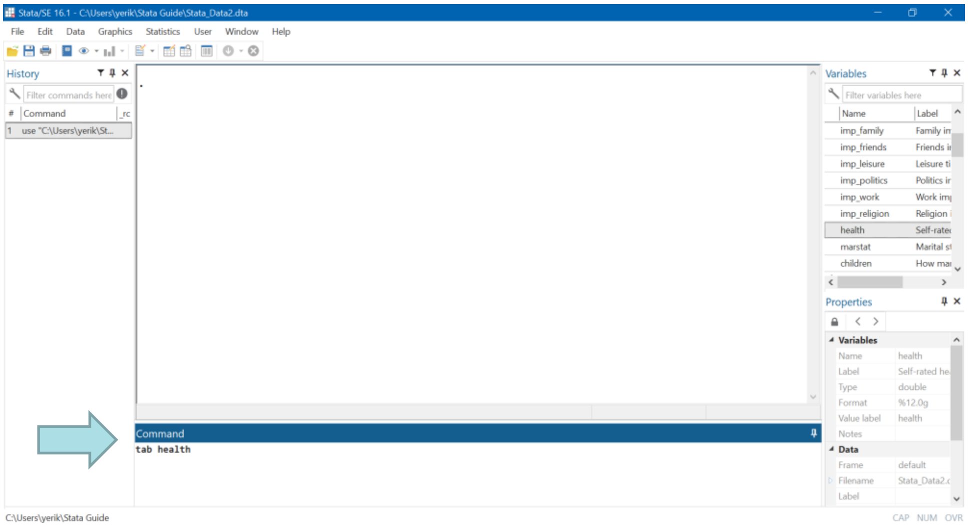


```
1   use "C:\Users\yerik\St...
```

Name: health
Label: Self-rated he
Type: double
Format: %12.0g
Value label: health
Notes:

Command

The Command window is where you can write your commands – although we highly recommend that you primarily write them in your do-file instead (see Section 2.1.2):

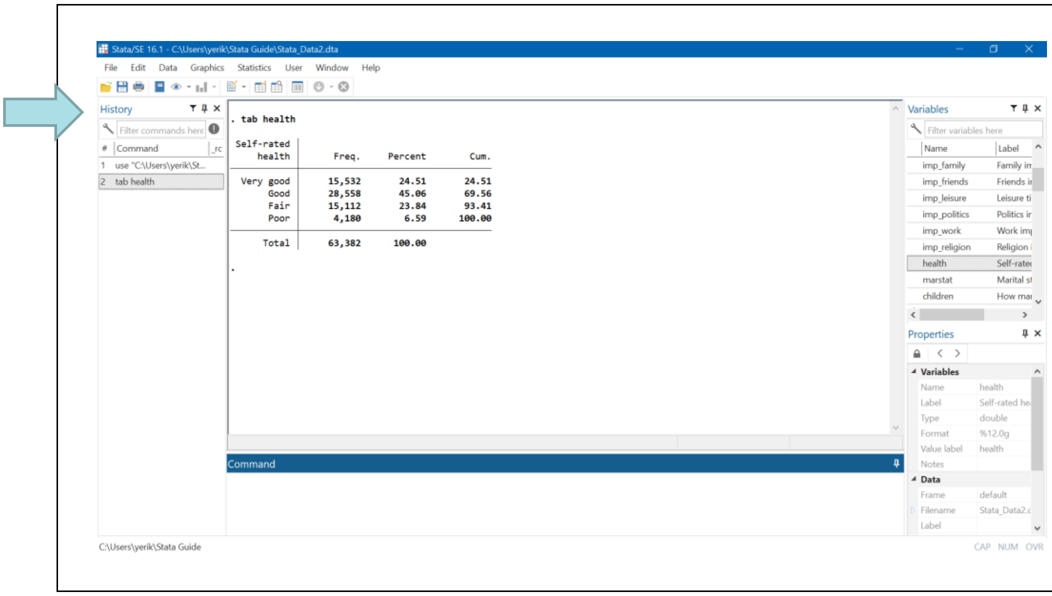


```
tab health
```

Note You can scroll among your previous commands by using the Page up and Page down keys on your keyboard.

History

All the commands that Stata performs, end up in the History window. However, if you use a do-file, the list will diminish considerably (a note will instead appear in the History window, stating that you have executed commands from a do-file):

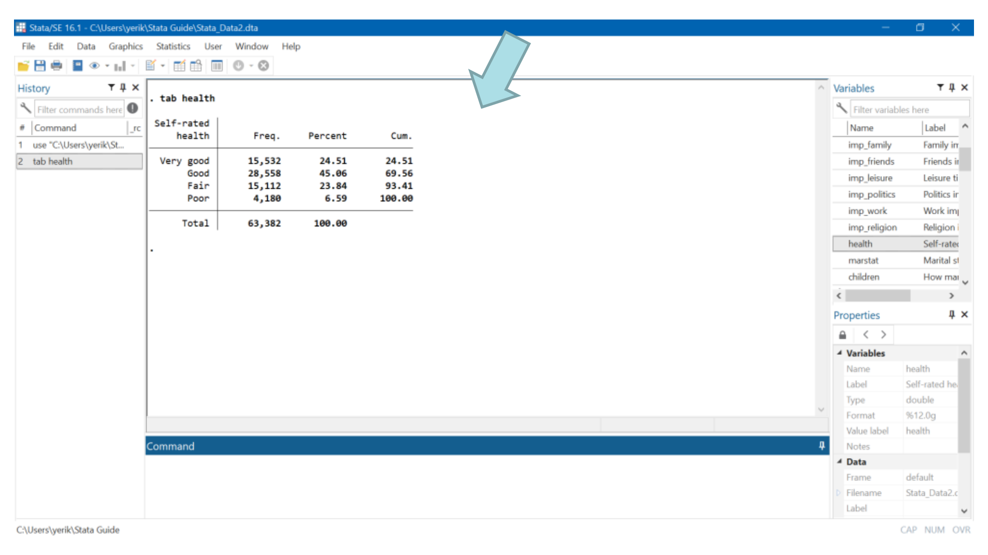


```
1   use "C:\Users\yerik\St...
2   tab health
```

```
. tab health

 Self-rated
     health |      Freq.     Percent        Cum.
------------+-----------------------------------
  Very good |     15,532       24.51       24.51
       Good |     28,558       45.06       69.56
       Fair |     15,112       23.84       93.41
       Poor |      4,180        6.59      100.00
------------+-----------------------------------
      Total |     63,382      100.00

.
```

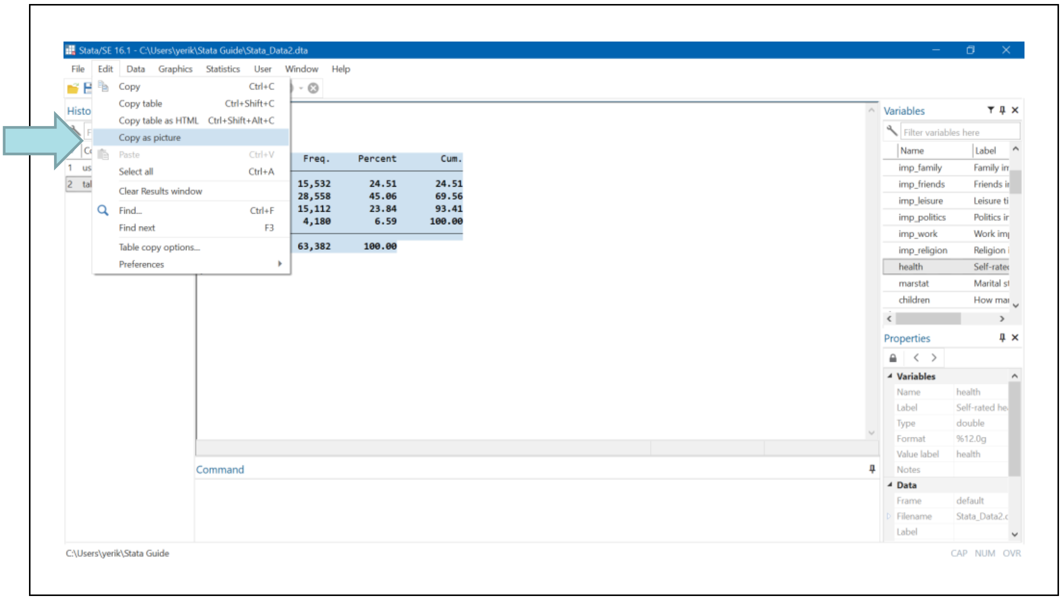
Results

All the results that Stata produces, end up in the Results window.



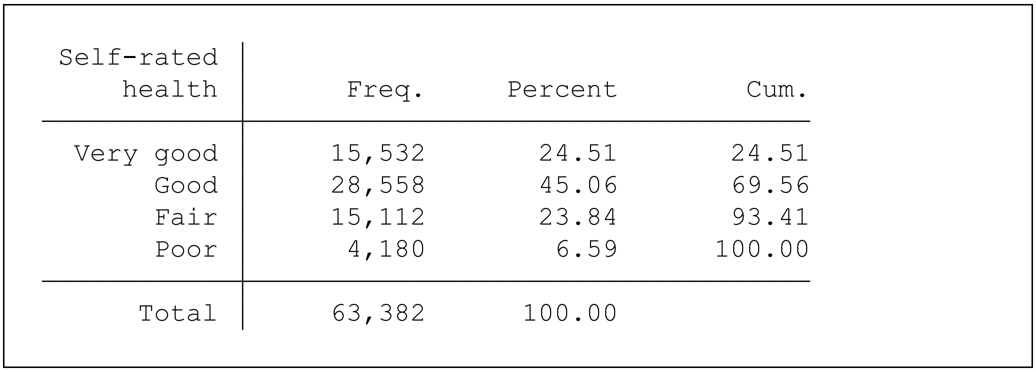
```
. tab health

   Self-rated |
       health |      Freq.     Percent        Cum.
--------------+-----------------------------------
    Very good |     15,532       24.51       24.51
         Good |     28,558       45.06       69.56
         Fair |     15,112       23.84       93.41
         Poor |      4,180        6.59      100.00
--------------+-----------------------------------
        Total |     63,382      100.00

.
```

One really practical thing is that you can copy the tables as pictures. Just highlight what you want, click on Edit\Copy as picture, and then paste it anywhere you like.



```
     Freq.     Percent        Cum.

    15,532       24.51       24.51
    28,558       45.06       69.56
    15,112       23.84       93.41
     4,180        6.59      100.00

    63,382      100.00
```

A table copied as a picture from Stata (note that the size of the picture depends on the width of your Results window; wider = smaller table):

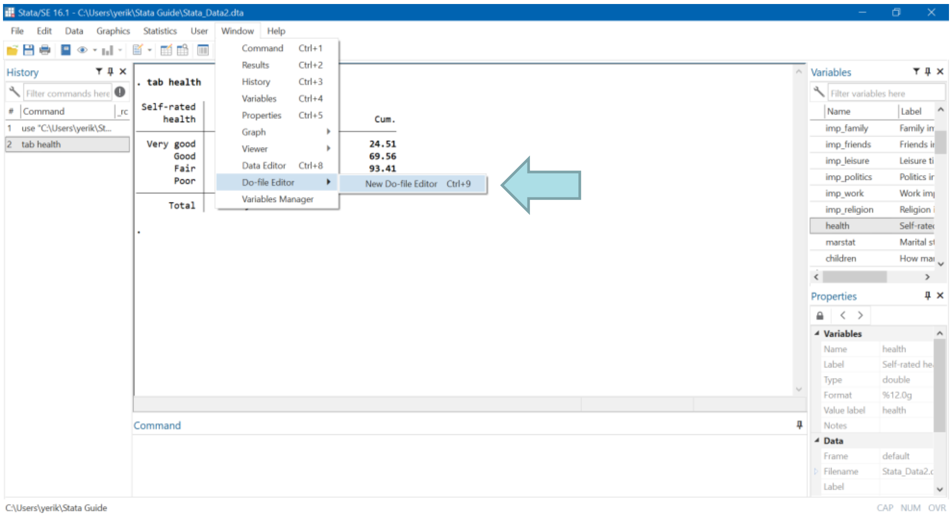


```
 Self-rated │
     health │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
  Very good │     15,532       24.51       24.51
       Good │     28,558       45.06       69.56
       Fair │     15,112       23.84       93.41
       Poor │      4,180        6.59      100.00
────────────┼───────────────────────────────────
      Total │     63,382      100.00
```

Alternatively, you can choose to copy the output as text and then insert it in another document, e.g. Word. The formatting will look terrible at first – make sure to change the font to Courier New (and choose a font size that fits the page).

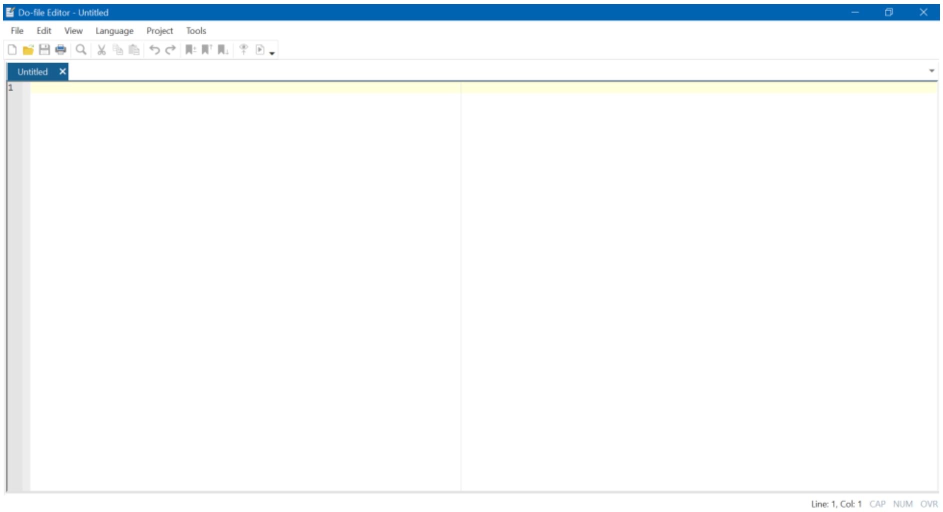
2.1.2 Do-file

As previously mentioned, we recommend that you use do-files to structure your work with Stata. Create a new do-file by clicking on Window\Do-file Editor\New Do-file Editor, or by using the keyboard shortcut Ctrl+9:



```
. tab health

Self-rated
    health                          Cum.

 Very good                         24.51
      Good                         69.56
      Fair                         93.41
      Poor

     Total

.
```

This is what a do-file looks like:


Do-file Editor - Untitled
File Edit View Language Project Tools
Untitled
1
Line: 1, Col: 1 CAP NUM OVR

Note You can increase/decrease the font size in the do-file by clicking on View\Zoom and choose Zoom in or Zoom out.

Add comments

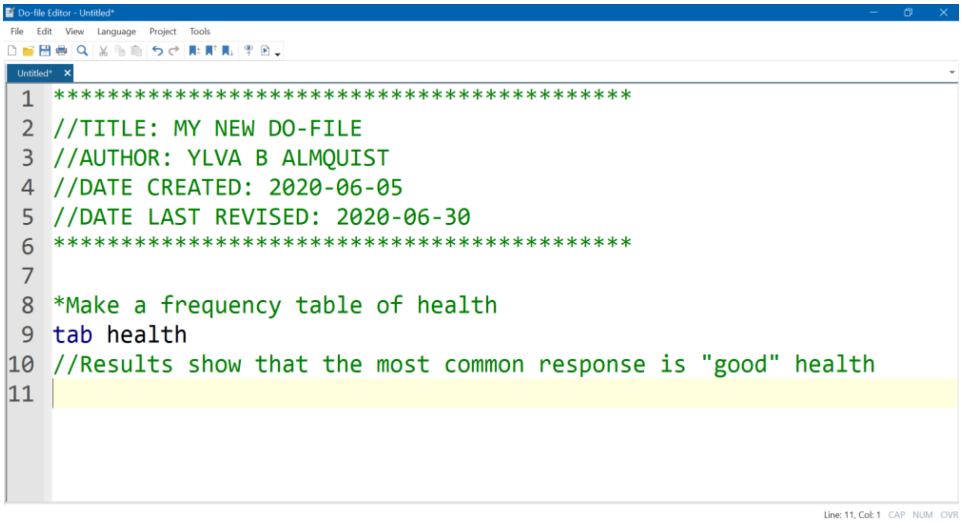
It is highly recommended that you comment your do-file. You may add a heading above each command, and also make notes of interesting findings, etc. There are four ways of adding comments in a do-file:

Alternative	Example
Start the comment with *	* This is a comment
Start the comment with //	// This is a comment
Start the comment with ///	/// This is a comment
Enclose the comment with /* */	/* This is a comment */

An advantage of the last option is that it allows you to include comments anywhere in the do-file, even in the middle of a command. The other need to be separated from the commands by using line breaks.

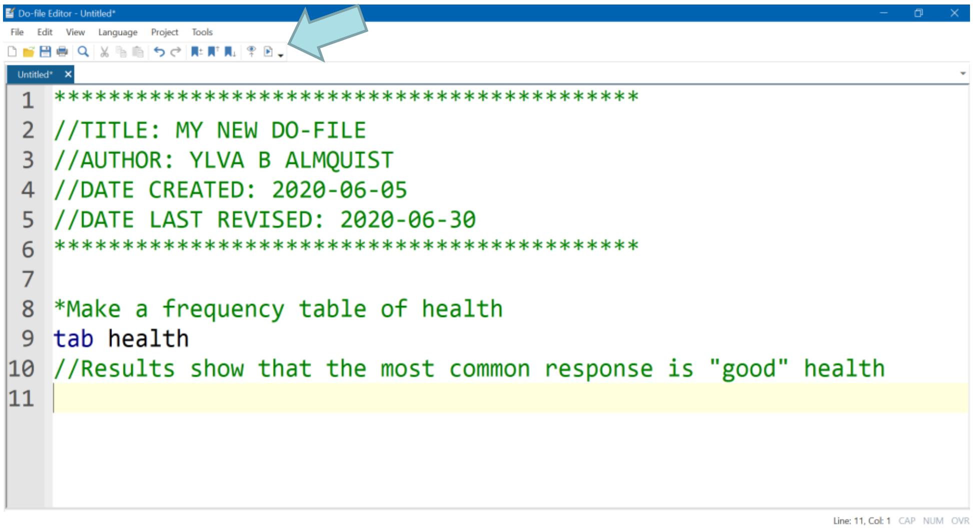
Note You can easily double-check that the comments are correctly entered, because they turn green if they are. It is also possible to mix different types of comments in the do-file.

Here is an example where we have started with an informative header for the do-file. After this, we can include the commands, along with headings and comments of the results:



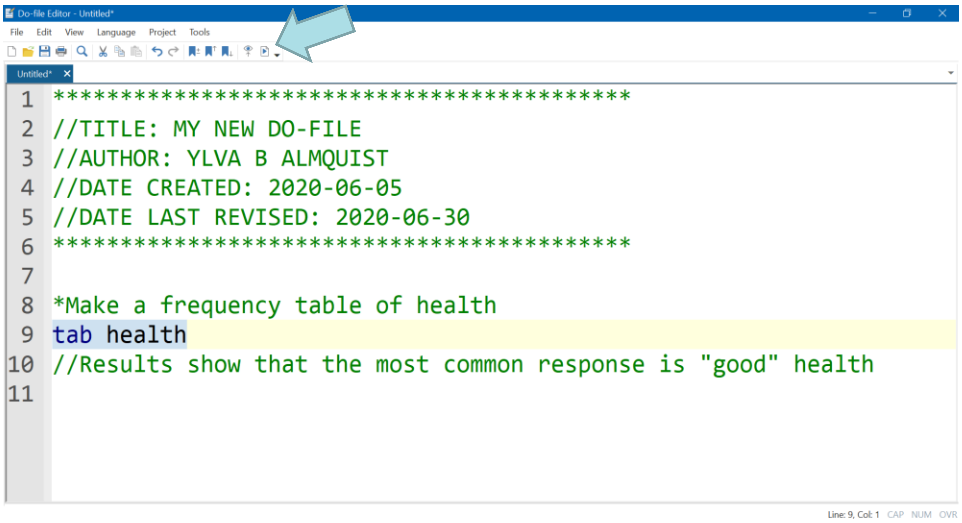
```
*******************************************
//TITLE: MY NEW DO-FILE
//AUTHOR: YLVA B ALMQUIST
//DATE CREATED: 2020-06-05
//DATE LAST REVISED: 2020-06-30
*******************************************

*Make a frequency table of health
tab health
//Results show that the most common response is "good" health
```

To execute the do-file, there are some different options. Either you can click on the button with a “play” arrow on it – Execute (do) – or you can use the keyboard shortcut Ctrl+D.



```
*****************************************
//TITLE: MY NEW DO-FILE
//AUTHOR: YLVA B ALMQUIST
//DATE CREATED: 2020-06-05
//DATE LAST REVISED: 2020-06-30
*****************************************

*Make a frequency table of health
tab health
//Results show that the most common response is "good" health

```

This executes the entire do-file. Most of the time, you only want to execute a specific command of part of the do-file. To manage this, you first need to highlight the part that you want to execute. Now you can click on the same button – this called Execute selection (do) – or use the keyboard shortcut Ctrl+D.



```
*****************************************
//TITLE: MY NEW DO-FILE
//AUTHOR: YLVA B ALMQUIST
//DATE CREATED: 2020-06-05
//DATE LAST REVISED: 2020-06-30
*****************************************

*Make a frequency table of health
tab health
//Results show that the most common response is "good" health

```

Error messages and return codes

If you execute a command that is incorrectly specified, Stata will return an error message in the Results window. It is not always evident why the error has occurred, even though Stata often provides some clues (sometimes, a specific return code is given, which you can learn more about through Google…).

Note In a majority of the cases, the error message is due to a missing symbol (e.g. a dot, a comma, or the inclusion a of a single equal sign instead of a double one). Review your do-file for mistakes like this.

Note Be careful – sometimes the command works even though it is not specified as you intended. This risk is particularly important to consider when you generate and recode variables.

Save a do-file

To save a do-file, you need to use the menu – it cannot be saved by a including a command in the do-file itself.

Open a do-file

To open an existing do-file, the easiest way is to do it through the menu (inside your dataset).

Note If you instead double-click on a do-file on your computer to open it up, an empty dataset will also open up – to which the do-file will be linked. This will cause problems. Therefore, it is suggested that you avoid this approach.

Load a dataset

In the above example, we already had a dataset open and then started a new do-file. But it is often the case that you want to open the dataset from within the syntax. You can do this with the command use:

```
use "path\filename.dta"
```

Change path\filename to the full path (i.e. the folder on your computer that contains the file), and specify the file name, such as:

```
use "C:\Users\yerik\Stata Guide\StataData2.dta"
```

More information	help use

Save a dataset

To save the dataset, type:

```
save "path\filename.dta"
```

or

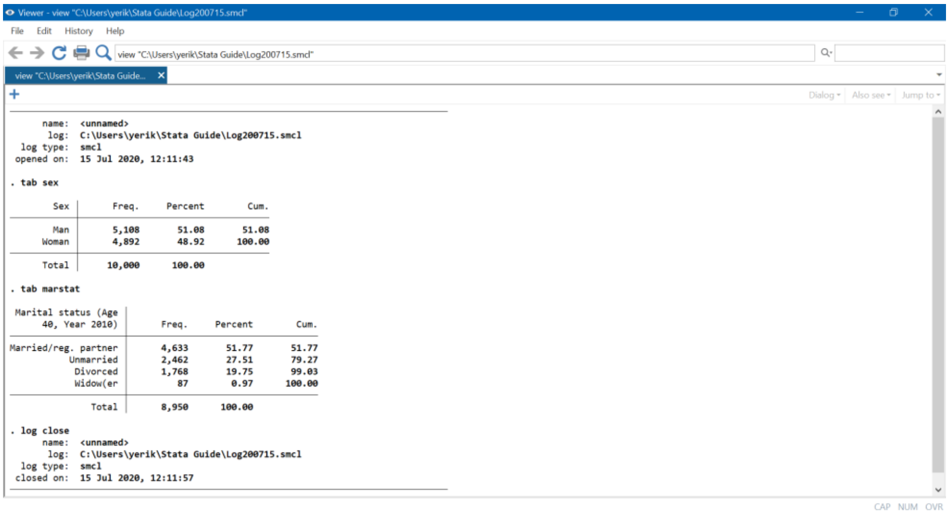
```
save "path\filename.dta", replace
```

More information	help save

2.1.3 Log

In Section 2.1.1, we demonstrated how to copy a table from the Results window. But sometimes you rather want to save a big chunk of output. To do this, you need to use a log file – this is a sort of a nicely formatted output file.

This is what a log can look like:


Viewer - view "C:\Users\yerik\Stata Guide\Log200715.smcl"
File
Edit
History
Help
view "C:\Users\yerik\Stata Guide\Log200715.smcl"
view "C:\Users\yerik\Stata Guide...
Dialog
Also see
Jump to
CAP
NUM
OVR

```
      name:  <unnamed>
       log:  C:\Users\yerik\Stata Guide\Log200715.smcl
  log type:  smcl
 opened on:  15 Jul 2020, 12:11:43

. tab sex

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
        Man |      5,108       51.08       51.08
      Woman |      4,892       48.92      100.00
------------+-----------------------------------
      Total |     10,000      100.00

. tab marstat

 Marital status (Age |
      40, Year 2010) |      Freq.     Percent        Cum.
---------------------+-----------------------------------
Married/reg. partner |      4,633       51.77       51.77
           Unmarried |      2,462       27.51       79.27
            Divorced |      1,768       19.75       99.03
            Widow(er |         87        0.97      100.00
---------------------+-----------------------------------
               Total |      8,950      100.00

. log close
      name:  <unnamed>
       log:  C:\Users\yerik\Stata Guide\Log200715.smcl
  log type:  smcl
 closed on:  15 Jul 2020, 12:11:57
```

More information	help log

Start a new log

Use the following command to start a new log (the path specifies where you want to log to be saved):

```
log using "path\filename.smcl"
```

Change path\filename to the full path (i.e. the folder on your computer that you want the log to end up in), and specify the log name of your choice, such as:

```
log using "C:\Users\yerik\Stata Guide\Log200715.smcl"
```

Close a log

Once you have produced the output you want, stop the log with this command:

```
log close
```

Continue with or replace a log

And here is how you open up an old log and add more stuff to it:

```
log using "path\filename.smcl", append
```

For example:

```
log using "C:\Users\yerik\Stata Guide\Log200715.smcl", append
```

Or if you prefer to open up the old log and overwrite it:

```
log using "path\filename.smcl", replace
```

For example:

```
log using "C:\Users\yerik\Stata Guide\Log200715.smcl", replace
```

View your log

Want to view your log? This is how:

```
view "path\filename.smcl"
```

For example:

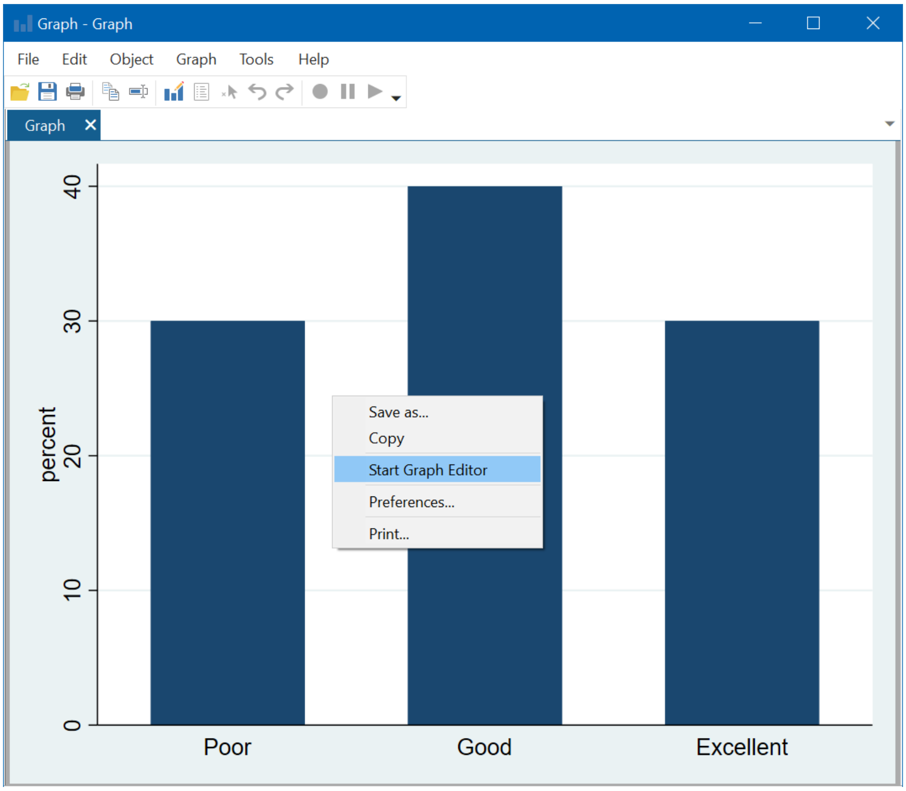
```
view "C:\Users\yerik\Stata Guide\Log200715.smcl"
```

2.1.4 Graphs

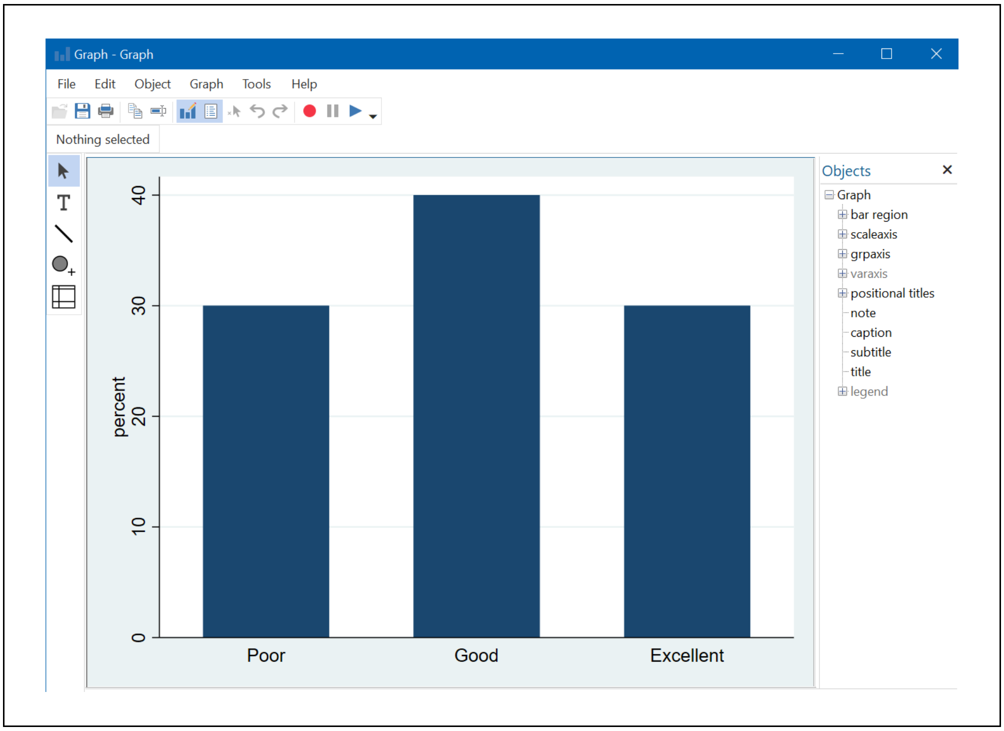
Any graph that you produce in Stata will appear in a new pop-up window.

Edit a graph

You can alter your graph at any time by right-clicking on it and choose Start Graph Editor.



This turns the pop-up window into the Graph Editor.



The remaining functions available here will not be covered in this guide – but we strongly encourage you to experiment! The bottom line is that you double-click on the things you want to change, in order to activate different options.

Note As long as the Graph Editor is open, you cannot do anything in your dataset or do-file (they are "locked").

To exit the Graph Editor, go to the menu, click on File and choose Stop Graph Editor. If you have made any changes, you will be asked whether you want to save those or not.

More information	help graph editor

Save a graph

Save your graph using the following command (note that this does not work if the Graph Editor is open):

```
graph save "path\filename.gph"
```

Change path\filename to the full path (i.e. the folder on your computer that you want the file to end up in), and specify the file name of your choice, such as:

```
graph save "C:\Users\yerik\Stata Guide\Srh.gph"
```

This only works the first time you save the graph. If you want to re-save a file, you need to specify the command a bit more:

```
graph save "C:\Users\yerik\Stata Guide\Srh.gph", replace
```

More information	help graph save

Open a graph

Here is how you open a graph:

```
graph use "path\filename.do"
```

Change path\filename to the full path (i.e. the folder on your computer that contains the file), and specify the file name, such as:

```
graph use "C:\Users\yerik\Stata Guide\Srh.gph"
```

More information	help graph use

2.1.5 Package

Stata is loaded with functions – but it is also possible to install clever user-written additions. Any specific packages will not be covered in this guide, but we will go through the installation process below.

New and popular

Want to see if which the most popular packages are?

```
ssc hot
```

Or do you want to see if there are any new, cool packages?

```
ssc new
```

Install

Most of the time, you know the name of the package that you want to install:

```
ssc install name
```

For example:

```
ssc install outreg2
```

Update

It might be good to check now and then whether your packages need updates.

```
ado update
```

And then to actually install the updates:

```
ado update, update
```

Uninstall

Want to uninstall a package?

```
ssc uninstall name
```

For example:

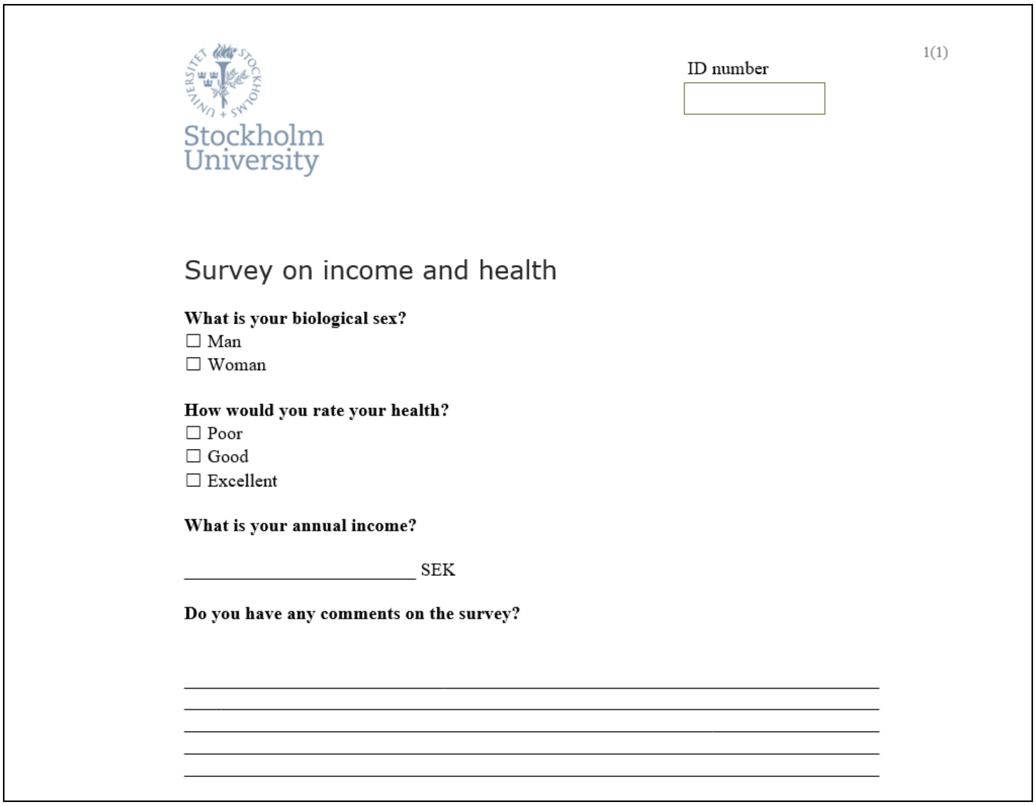
```
ssc uninstall outreg2
```

More information	help ssc
	help ado

2.2 Creating a new dataset

2.2.1 From questionnaire to dataset

Sometimes there is a need to create a dataset from scratch. This is, for example, the case when we have performed a survey (like the one below) and have a pile of filled in paper questionnaires that somehow need to be transferred into Stata.



1(1)

ID number

Stockholm University

Survey on income and health

What is your biological sex?
☐ Man
☐ Woman

How would you rate your health?
☐ Poor
☐ Good
☐ Excellent

What is your annual income?

______________________ SEK

Do you have any comments on the survey?

Before we can actually code the questionnaire responses, we need to create the variable structure in Stata. In the questionnaire shown above, there is a total of five variables:

- ID number
- What is your biological sex?
- How would you rate your health?
- What is your annual income?
- Do you have any comments on the survey?

In Stata, each of these variables should be specified according to its Name, Label, Type, Format, and Value Label.

Name

This is the name that you choose for a variable. Make it short, clear, and logical. Avoid any spaces or special symbols. Underscores can be useful. It is highly recommended that you use lower case letters.

Label

This a more elaborate description of your variable. If the variable is drawn from a questionnaire, it would be practical to use the question as the label.

Type

There are two different types of variables in Stata: numeric and string.

Numeric variables can only handle numeric data. Such variables are the basis of quantitative research – which is why we usually always “translate” categorical variables into numeric variables by assigning a number to each category. Numbers are stored as byte, int, long, float, or double. Among these, byte, int, and long can hold only integers (i.e. whole numbers). The default storage type when you create a new variable in Stata is float.

String variables can handle any data (i.e. any numbers and letters) but is more difficult to analyse. Therefore, they are often processed (“quantified”) in ways that make it possible to use them in statistical analysis. Either way, strings are stored as str#, for instance, str1, str2, str3, ..., or as strL. The # sign indicates the maximum length of the string, i.e. how many characters that the variable can store. For example, a str2 can hold the word "no", but not the word "yes". A strL can hold strings up to 2000000000 characters.

Note If you are worried about the size of your data files, it is good to read up on the different storage types. If not, just keep in mind the difference between numeric and string variables. And also make sure to know whether your variable is an integer (whole numbers) or not (has decimals).

Format

The format of the variable is a function of its type. Default is:

Storage type	Format
Byte	%8.0g
Int	%8.0g
Long	%12.0g
Float	%9.0g
Double	%10.0g
str#	%#s
strL	%9s

Value label

This is where you specify the labels for any categories that the variable might have (thus, only useful for categorical variables, not continuous ones).

2.2.2 Variable structure

We will now use some syntax to create the variables specified in the table:

Name	Label	Type	Format	Value label
id	ID number	Int	%8.0g	-
sex	What is your biological sex?	Int	%8.0g	0=Man 1=Woman
srh	How would you rate your health?	Int	%8.0g	1=Poor 2=Good 3=Excellent
income	What is your annual income?	Int	%9.0g	-
comment	Do you have any comments on the survey?	strL	%9s	-

Step 1. Generate variables and specify type

As a first step, we generate the variables and specify their type. We also need to tell Stata what the values of the new variable should be. For the numeric variables, we go with missing (denoted by ".").

```
gen int id=.
```

```
gen int sex=.
```

```
gen int srh=.
```

```
gen int income=.
```

```
gen strL comment=""
```

Note For our string variable, it is slightly different. We need to use double quotes here – but nothing actually has to be specified within the double quotes.

More information	help generate

Step 2. Add labels

We can now add labels for the variables:

```
label variable id "ID number"
```

```
label variable sex "What is your biological sex?"
label variable srh "How would you rate your health?"

label variable income "What is your annual income?"

label variable comment "Do you have any comments on the survey?"
```

More information	help label

Step 3. Add value labels

The final step is to add value labels for the categorical variables. We do this by first specifying a set of value labels:

```
label define sex 0 "Man" 1 "Woman"

label define srh 1 "Poor" 2 "Good" 3 "Excellent"
```

Note For simplicity reasons, they have the same name as the corresponding variables. But if we would have had, for example, a whole set of variables which all had the response options “No” and “Yes”, we could have created just one label and used it for all those variables.

Now it is time to apply our labels:

```
label values sex sex

label values srh srh
```

Do you suddenly realize that you need to adjust your labels? You can change them by using the following commands:

```
label define sex 0 "Man" 1 "Woman", replace

label define srh 1 "Poor" 2 "Good" 3 "Excellent", replace
```

You can also delete value labels by writing the following:

```
label drop sex

label drop srh
```

More information	help label

2.2.3 Manage variables

It is good to be able to write syntax – but sometimes it might be useful to also get a more overarching perspective on the dataset. This can be achieved by using Stata's Variables Manager. Use the following command to access it:

```
varmanage
```

2.2.4 Coding the questionnaires

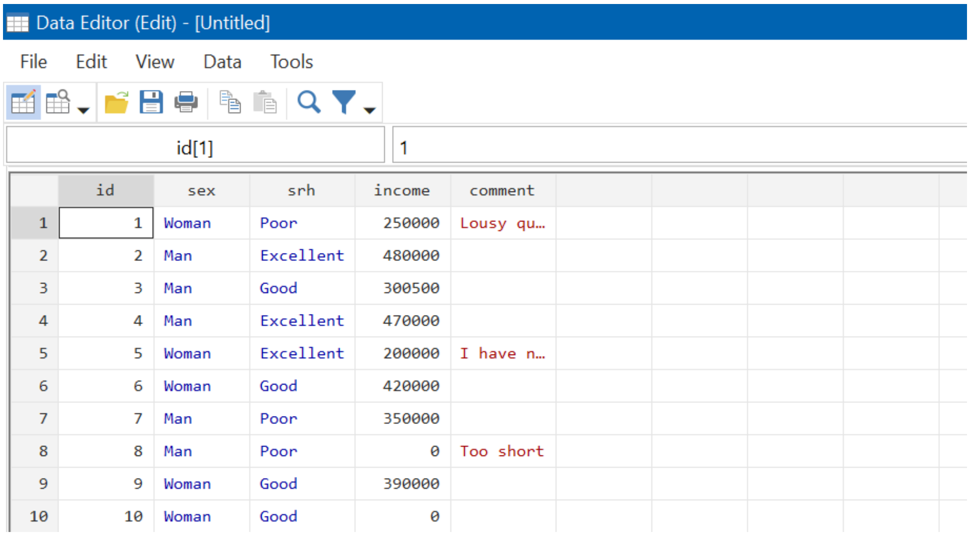
So far, so good! Now it is time to transfer the actual responses from the questionnaire to Stata.

Edit data

Add this command to open the Data Editor:

```
edit
```

Here you can simply add the responses you have in your questionnaires. Note that if you have specified value labels, these will appear instead of the actual value:



	id	sex	srh	income	comment
1	1	Woman	Poor	250000	Lousy qu…
2	2	Man	Excellent	480000	
3	3	Man	Good	300500	
4	4	Man	Excellent	470000	
5	5	Woman	Excellent	200000	I have n…
6	6	Woman	Good	420000	
7	7	Man	Poor	350000	
8	8	Man	Poor	0	Too short
9	9	Woman	Good	390000	
10	10	Woman	Good	0	

(Some entries for comments have been truncated. Id 1=Lousy questionnaire; Id 5=I have nothing to add; Id 8=Too short).

Note You can use the arrows on your keyboard to navigate the cells.

Close the Data Editor when you are done. Do not forget to save the dataset. We have chosen to call it TestData1.dta.

Note Every change that you perform in the Data Editor generates the corresponding command in the Results window. It might be good to log these commands in order to be able to re-create the data at a later point (see Section 2.1.3 for more information about log files).

Browse data

Do you want to have a look at the Data Editor without actually editing? Then you can use browse:

```
browse
```

If you just want to browse specific variables or portions of the variable list, this can be specified after browse:

```
browse varname
```

or

```
browse varname varname varname
```

or

```
browse varname-varname
```

For example:

```
browse sex income
```

Note You can also use browse together with if (see Section 2.7). An alternative to browse is list (see Section 2.3.1).

2.3 Adjusting an existing dataset

Often, we do not create an entire new dataset from scratch; we rather import another type of file (e.g. Excel or .csv) that contains data of some sort. In this section, we will walk you through some of the most useful commands in Stata.

2.3.1 Review dataset

Describe

When you want to have a quick overlook of your variables, the describe command might be very useful. It is basically like a summary of what you can see when using the Variables Manager (see Section 2.2.3).

```
describe
```

```
  obs:            10
 vars:             5
------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
id              int     %8.0g                 ID number
sex             int     %8.0g      sex        What is your biological sex?
srh             int     %9.0g      srh        How would you rate your health?
income          long    %8.0g                 What is your annual income?
comment         strL    %9s                   Do you have any comments on the
                                                survey?
------------------------------------------------------------------------------
Sorted by:
```

If you just want to describe specific variables or portions of the variable list, this can be specified after describe:

```
describe varname
```

or

```
describe varname varname varname
```

or

```
describe varname-varname
```

For example:

```
describe sex income
```

More information	help describe

Codebook

As a complement to describe, you can use codebook.

Note We suggest that you include the compact option, or you will get a lot of output.

```
codebook, compact
```

```
Variable   Obs Unique    Mean  Min     Max  Label
-----------------------------------------------------------------------------------------
id          10     10     5.5    1      10  ID number
sex         10      2      .5    0       1  What is your biological sex?
srh         10      3       2    1       3  How would you rate your health?
income      10      9  286050    0  480000  What is your annual income?
comment      3      3       .    .       .  Do you have any comments on the survey?
-----------------------------------------------------------------------------------------
```

You can also include the following option to see any potentially missing information in the dataset (e.g. missing labels or value labels).

```
codebook, problems
```

```
   Potential problems in dataset   C:\Users\yerik\Stata Guide\TestData.dta

               potential problem   variables
--------------------------------------------------
  string vars with embedded blanks   comment
--------------------------------------------------
```

Finally, if you want to explore a specific variable in a detailed way, use this option:

```
codebook srh, detail
```

```
-----------------------------------------------------------------------------------------
srh                                                        How would you rate your health?
-----------------------------------------------------------------------------------------

                  type:  numeric (int)
                 label:  srh

                 range:  [1,3]                        units:  1
         unique values:  3                        missing .:  0/10

            tabulation:  Freq.   Numeric  Label
                             3         1  Poor
                             4         2  Good
                             3         3  Excellent
```

More information	help codebook

List

As an alternative to browse (see Section 2.2.4), you can use list. However, it works best for datasets with a limited number of variables and observations, or if you only list a portion of your dataset – otherwise, the output will be extremely difficult to read.

```
list
```

```
      +---------------------------------------------------------+
      | id     sex         srh   income                 comment |
      |---------------------------------------------------------|
   1. |  1   Woman        Poor   250000     Lousy questionnaire |
   2. |  2     Man   Excellent   480000                         |
   3. |  3     Man        Good   300500                         |
   4. |  4     Man   Excellent   470000                         |
   5. |  5   Woman   Excellent   200000   I have nothing to add |
      |---------------------------------------------------------|
   6. |  6   Woman        Good   420000                         |
   7. |  7     Man        Poor   350000                         |
   8. |  8     Man        Poor        0               Too short |
   9. |  9   Woman        Good   390000                         |
  10. | 10   Woman        Good        0                         |
      +---------------------------------------------------------+
```

If you just want to describe specific variables or portions of the variable list, this can be specified after list:

```
list varname
```

or

```
list varname varname varname
```

or

```
list varname-varname
```

For example:

```
list sex income
```

You can also choose to only list a range of observations. Note that the output depends on how the observations are sorted.

```
list in x/x
```

For example:

```
list in 1/5
```

Note You can also use list together with if (see Section 2.7).

More information	help list

2.3.2 Convert variables

It might be the case that you have a lot of string variables in your dataset, although they are actually numeric. Why is this not desirable? Well, statistical analysis is all about numbers. Accordingly, we want to convert string variables to numeric variables as far as possible (of course, this cannot be as easily fixed for variables that actually contain – and should contain – strings of text).

There are a couple of different ways that we can convert string variables to numeric variable. We will use a dataset called TestData2.dta, which looks like this:

```
describe
```

```
  obs:            10
 vars:             5                          15 Jul 2020 20:12
--------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------------
---------------------
id              int     %8.0g                 ID number
sex2            str1    %9s                   What is your biological sex?
srh2            str9    %9s                   How would you rate your health?
income2         str7    %9s                   What is your annual income?
comment         strL    %9s                   Do you have any comments on the survey?
--------------------------------------------------------------------------------------
Sorted by:
```

```
list
```

```
     +-----------------------------------------------------------+
     | id   sex2         srh2   income2                  comment |
     |-----------------------------------------------------------|
  1. |  1      1         Poor   250,000      Lousy questionnaire |
  2. |  2      0    Excellent   480,000                          |
  3. |  3      0         Good   300,500                          |
  4. |  4      0    Excellent   470,000                          |
  5. |  5      1    Excellent   200,000    I have nothing to add |
     |-----------------------------------------------------------|
  6. |  6      1         Good   420,000                          |
  7. |  7      0         Poor   350,000                          |
  8. |  8      0         Poor         0                Too short |
  9. |  9      1         Good   390,000                          |
 10. | 10      1         Good         0                          |
     +-----------------------------------------------------------+
```

By reviewing the tables above, we can notice that three of the variables – sex2, srh2, and income2 are string variables although they actually could be numeric. This will make them rather impossible to use in statistical analysis. However, we need different approaches to actually convert them to numeric – described in detail below.

Real

The first alternative is to use real. This works for string variables that only contain numbers, such as sex2.

```
generate sex=real(sex2)
```

This will create a new variable called sex, which is a numeric version of sex2.

```
describe sex sex2
```

```
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------
sex             float   %9.0g
sex2            str1    %9s                   What is your biological sex?
```

More information	help real

Destring

For sex2, we could have achieved the almost same result by using destring.

```
destring sex2, gen(sex)
```

This too will create a new variable called sex, which is numeric version of sex2. An advantage is that we keep the variable label (and the automatically selected storage type is different).

```
describe sex sex2
```

```
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------
sex             byte    %10.0g                What is your biological sex?
sex2            str1    %9s                   What is your biological sex?
```

For the variable income2, it is not possible to use real at all, since this string variable contains non-numeric values (in this case, commas). If we use real, all cells that contain a non-numeric character will have missing values. Here our best option is to use destring, which allows us to ignore the non-numeric characters.

```
destring income2, gen(income) ignore(",")
```

```
describe income income2
```

```
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------------
income          long    %10.0g                What is your annual income?
income2         str7    %9s                   What is your annual income?
```

More information	help destring

Encode

The third type of string variable that we want to convert to a numeric variable, is srh2. This variable, however, have the actual categories coded in the cells (i.e. Poor, Good, and Excellent). We want these translated into numbers instead. We can use encode to achieve this.

```
encode srh2, gen(srh)
```

```
describe srh srh2
```

```
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------------
---------------------
srh             long    %9.0g      srh        How would you rate your health?
srh2            str9    %9s                   How would you rate your health?
```

Note that Stata automatically creates value labels for the new variable srh.

More information	help encode

Tostring and decode

Finally, sometimes we might want to convert numeric variables into string variables (e.g. to be able to use substring, see Section 2.4.6).

According to the same principles as we used destring and encode, we can apply tostring and decode.

More information	help tostring
	help decode

2.3.3 Rename variables

Renaming variables is very easy – just make sure to use a logical name (without spaces or special symbols) and not a name that is already taken by another variable:

```
rename oldvarname newvarname
```

For example:

```
rename var01 id
```

More information	help rename

2.3.4 Delete variables

Do you need to delete a variable? This is how you do it:

```
drop varname
```

For example:

```
drop var33
```

If you have a large number of variables that you want to delete, it might sometimes be easier to tell Stata which the variables you want to retain:

```
keep varname
```

For example:

```
keep var1-var10 var66
```

Note You cannot undo the deletion of a variable.

More information	help drop
	help keep

2.3.5 Sort dataset

It might be good to keep your dataset sorted, according to e.g. id number. Just keep in mind that Stata interprets missing numeric values as being larger than any other number, so they are placed last when you sort. When you sort on a string variable, however, null strings (empty strings) are placed first.

```
sort varname
```

For example:

```
sort id
```

More information	help sort

2.3.6 Create an id number variable

Not all datasets have an id number variable (for example, if you have performed an anonymous survey, there might not have been a need to mark the questionnaires with any id number). This is easily fixed in Stata – note, however that Stata will assign numbers in the order that the dataset is sorted, so make sure that you have it sorted in the way you like first):

```
gen varname = _n
```

For example:

```
gen id = _n
```

More information	help generate

2.3.7 Order variables

Is it driving you crazy that variables are ordered in an illogical way? This can be fixed by using order. When using order, you must specify which variables you want to move. First, we will show you how to move one variable (works also for more than one specified variable), then we show how to re-order the whole list of variables (works also for portions of the list)

Move the variable to the beginning of the dataset:

```
order varname, first
```

For example:

```
order id, first
```

Move the variable to the end of the dataset:

```
order varname, last
```

For example:

```
order mortality, last
```

Move the variable to before another variable:

```
order varname1, before(varname2)
```

For example:

```
order health_2007, before(health_2008)
```

Move the variable to after another variable:

```
order varname1, after(varname2)
```

For example:

```
order phystest2, after(phystest1)
```

Order all the variables in the dataset alphabetically:

```
order _all, alpha
```

or

```
order firstvarname-lastvarname, alpha
```

For example:

```
order id-mortality, alpha
```

Order a portion of the variable list alphabetically:

```
order phystest1-phystest10, alpha
```

More information	help order

2.4 Generate

We suspect that generate (or gen, for short) is perhaps one of the commands that you will use the most in Stata. We have already applied it in several parts of the guide, and here are some additional alternatives.

More information	help generate

2.4.1 Copy of an existing variable

Make a copy of an existing variable. This is highly useful e.g. when you want to experiment with a variable or to recode a variable without altering the original version.

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)

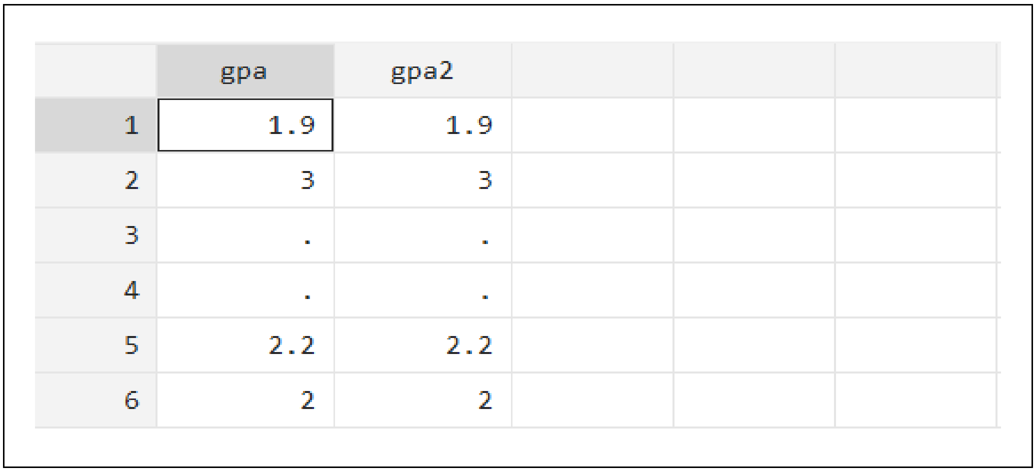
```
gen gpa2=gpa
```

As can be seen below, the new variable gpa2 will be an exact copy of the old variable gpa.

```
sum gpa gpa2
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      9,380    3.178614    .6996298          1          5
        gpa2 |      9,380    3.178614    .6996298          1          5
```

```
browse gpa gpa2
```



	gpa	gpa2			
1	1.9	1.9			
2	3	3			
3	.	.			
4	.	.			
5	2.2	2.2			
6	2	2			

2.4.2 New variable with a specific value

Here is one example of how we can create a new variable with a specific value.

Practical example

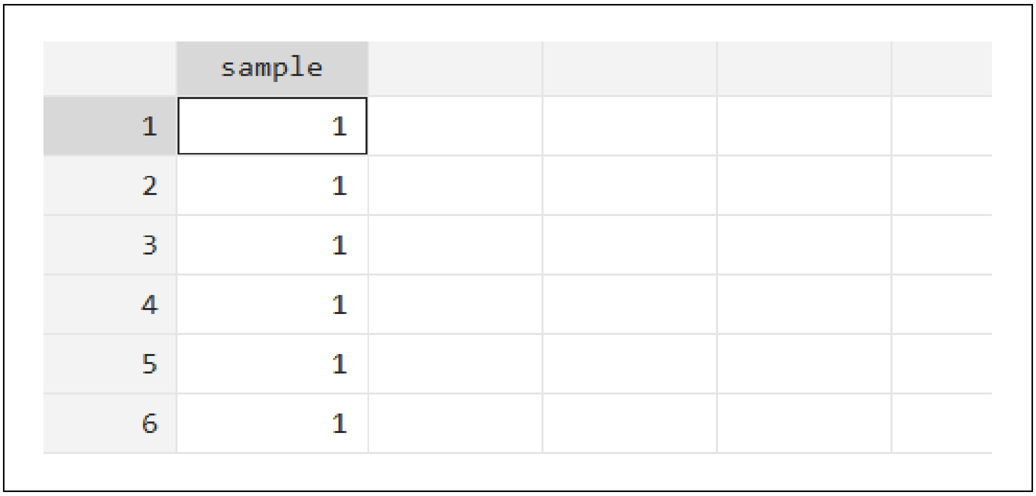
```
gen sample=1
```

This creates a new variable called sample, for which all observations will have the value 1.

```
sum sample
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      sample |     10,000           1           0          1          1
```

```
browse sample
```



	sample
1	1
2	1
3	1
4	1
5	1
6	1

2.4.3 New variable based on an expression

Create a new variable based on an expression that can include multiple variables and/or values. Below, two examples are presented.

Practical example 1

Dataset: StataData1.dta

Name	**Label**
unemp_42	Days in unemployment (Age 42, Year 2012)
unemp_43	Days in unemployment (Age 43, Year 2013)
unemp_44	Days in unemployment (Age 44, Year 2014)
unemp_45	Days in unemployment (Age 45, Year 2015)

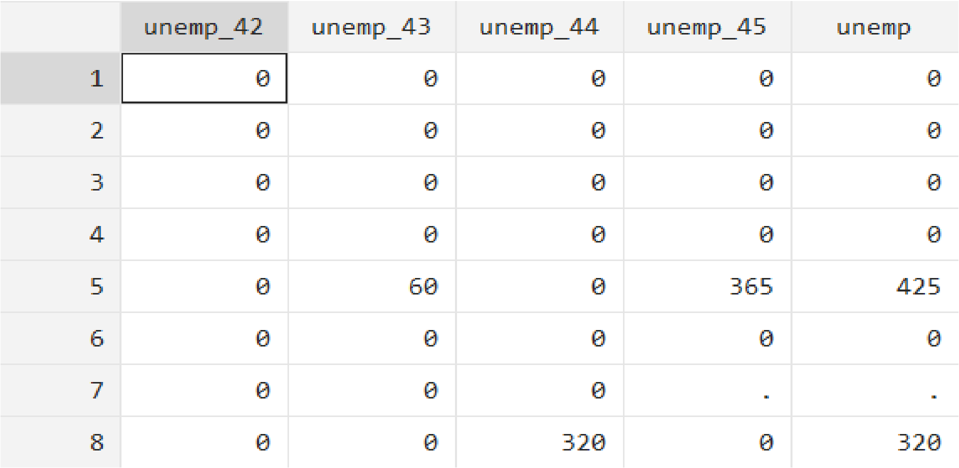
```
gen unemp=unemp_42+unemp_43+unemp_44+unemp_45
```

The new variable unemp contains the sum of the other four variables.

```
sum unemp_42 unemp_43 unemp_44 unemp_45 unemp
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    unemp_42 |      9,078    17.52787    58.68258          0        365
    unemp_43 |      8,994    7.593173    36.93804          0        365
    unemp_44 |      8,880     9.48018    44.70616          0        365
    unemp_45 |      8,773    5.531859    34.04111          0        365
       unemp |      8,672    39.39864    111.0481          0       1434
```

```
browse unemp_42 unemp_43 unemp_44 unemp_45 unemp
```



	unemp_42	unemp_43	unemp_44	unemp_45	unemp
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	60	0	365	425
6	0	0	0	0	0
7	0	0	0	.	.
8	0	0	320	0	320

Note If any of the variables that you include in the expression have missing values, the value for the new variable will be missing as well.

Practical example 2

Dataset: StataData1.dta

Name	**Label**
bweight	Birth weight (Age 0, Year 1970)

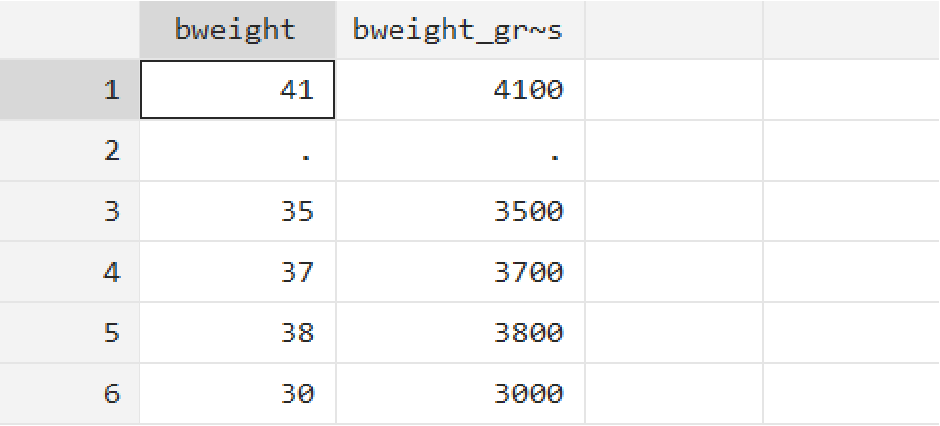
```
gen bweight_grams=bweight*100
```

The old variable bweight shows birth weight in hectograms, but now we have created the new variable bweight_grams which shows birth weight in grams instead.

```
sum bweight bweight_grams
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     bweight |      8,010    36.01074    5.406709          2         62
bweight_gr~s |      8,010    3601.074    540.6709        200       6200
```

```
browse bweight bweight_grams
```



	bweight	bweight_gr~s
1	41	4100
2	.	.
3	35	3500
4	37	3700
5	38	3800
6	30	3000

2.4.4 Rounding

Do you want to reduce the number of decimals by rounding a variable?

Practical example

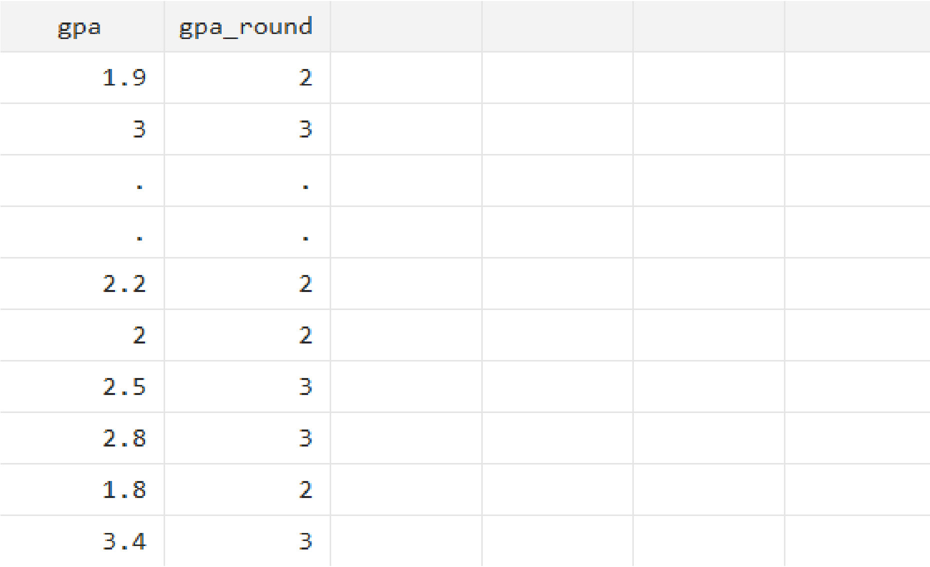
Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)

```
gen gpa_round = round(gpa,1)
```

Here, the values for gpa have been rounded to the nearest whole number and saved as gpa_round.

```
browse gpa gpa_round
```



gpa	gpa_round
1.9	2
3	3
.	.
.	.
2.2	2
2	2
2.5	3
2.8	3
1.8	2
3.4	3

2.4.5 Logarithmic transformation

Do you want to take the natural logarithm of a variable (log transformation) and create a new variable?

Practical example

Dataset: StataData1.dta

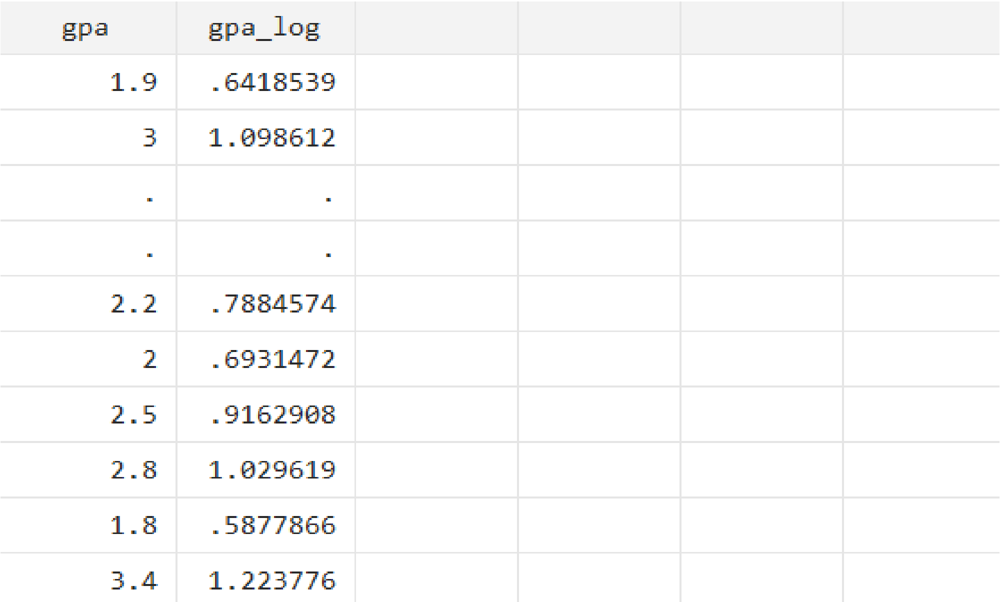
Name	**Label**
gpa	Grade point average (Age 15, Year 1985)

```
gen gpa_log=ln(gpa)
```

In Stata, it works exactly the same if you replace “ln” with “log”.

Note The command is “ln” (lower-case L, not upper-case i).

```
browse gpa gpa_log
```



gpa	gpa_log
1.9	.6418539
3	1.098612
.	.
.	.
2.2	.7884574
2	.6931472
2.5	.9162908
2.8	1.029619
1.8	.5877866
3.4	1.223776

2.4.6 Substring

If you need to subtract a portion (substring) from a string variable, you can use substr. The authors of the guide can happily reveal that they have applied this a lot when working with ICD codes (classification system for diagnoses).

Function

Basic command	gen newvarname= substr(oldvarname,start,length)	
Explanations	newvarname	Insert the name of the new variable (containing the substring).
	oldvarname	Insert the name of the new variable (the original string variable).
	substr	Extract a portion of the string variable.
	start	Specify which position that the starting character in the substring has.
	length	Specify the length of the substring.
More information	help substr	

Practical example

Dataset: StataData1.dta

Name	**Label**
cvd_date_str	Date of out-patient care due to CVD (Ages 41-50, Years 2011-2020)

We have a string variable called cvd_date_str that contains the date of out-patient care due to cardiovascular disease (CVD), coded like YYYYMMDD. Suppose that we want to extract the year (YYYY), month (MM), and day (DD) into separate variables.

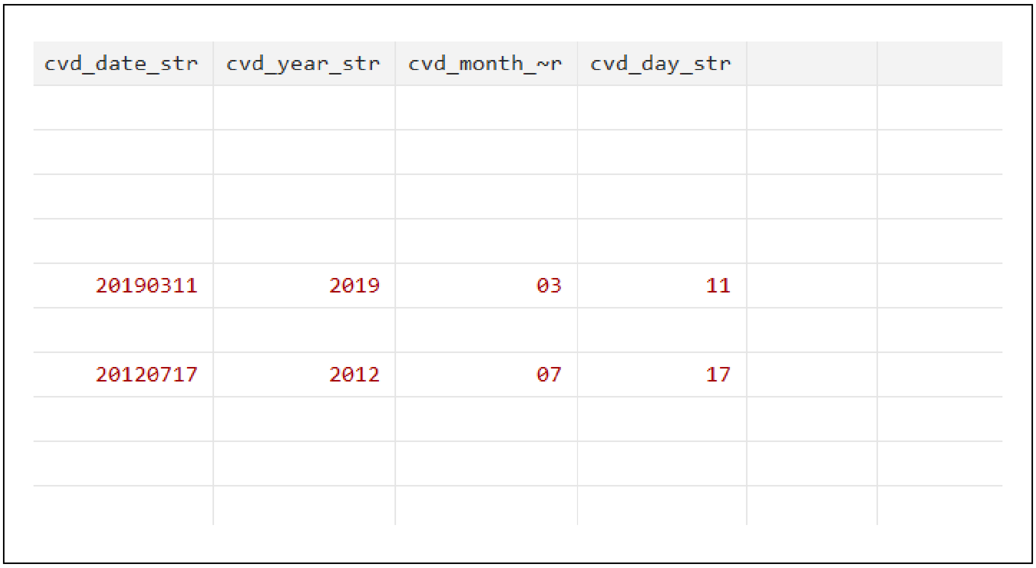
```
gen cvd_year_str= substr(cvd_date_str,1,4)
```

```
gen cvd_month_str= substr(cvd_date_str,5,2)
```

```
gen cvd_day_str= substr(cvd_date_str,7,2)
```

As can be noted in the command above, for year, we specify 1 as the position which the starting character in the substring has, and 4 as the length. For month, we specify 5 and 2. And, finally, for day, we specify 7 and 2.

```
browse cvd_date_str cvd_year_str cvd_month_str cvd_day_str
```



cvd_date_str	cvd_year_str	cvd_month_~r	cvd_day_str
20190311	2019	03	11
20120717	2012	07	17

Let us also add some variable labels for these new variables.

```
label variable cvd_year_str "Year of out-patient care due to CVD (Ages 41-50, Years 2011-2020)"
```

```
label variable cvd_month_str "Month of out-patient care due to CVD (Ages 41-50, Years 2011-2020)"
```

```
label variable cvd_day_str "Day of out-patient care due to CVD (Ages 41-50, Years 2011-2020)"
```

2.4.7 Date variables

Date variables – do not get us started. This is a science in itself! It might nonetheless be very useful later on if you want to perform time-to-event analysis (survival analysis) to be able to generate date variables.

In this example, we will use three variables that specify year, month, and day, respectively, and combine them into a nicely formatted date variable.

Note This requires that you have performed the practical example in Section 2.4.6 first.

Dataset: StataData1.dta

Name	**Label**
cvd_year_str	Year of out-patient care due to CVD (Ages 41-50, Years 2011-2020)
cvd_month_str	Month of out-patient care due to CVD (Ages 41-50, Years 2011-2020)
cvd_day_str	Day of out-patient care due to CVD (Ages 41-50, Years 2011-2020)

All three are string variables. To make things smoother, we will transform them into numeric variables, using real.

```
gen cvd_year=real(cvd_year_str)
```

```
gen cvd_month=real(cvd_month_str)
```

```
gen cvd_day=real(cvd_day_str)
```

Just to double-check that everything worked out:

```
sum cvd_year cvd_month cvd_day
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    cvd_year |        518    2015.506    2.817287       2011       2020
   cvd_month |        518    6.393822    3.435974          1         12
     cvd_day |        518    16.00579    8.859857          1         31
```

The next step it to generate the date variable.

```
gen cvd_date=mdy(cvd_month,cvd_day,cvd_year)
```

Note The term “mdy” means that the date is specified as month/day/year. This will create a special Stata date variable.

And finally, we format the date variable so it makes more sense for Stata:

```
format %d cvd_date
```

2.5 Egen

Is there something that you need to do, but the commands we have gone through so far do not seem to offer the solution? Then it is highly recommended that you explore egen, which is an extension generate.

More information	help egen

2.5.1 Standardization: z-scores

The standard score – or the z-score – is very useful when we have continuous (ratio/interval) variables with different normal distributions (see Section 3.4 for more information about distributions). For example, if we have one variable called income (measured as annual household income in Swedish crowns) and another variable called years of schooling (measured as the total number of years spent in the educational system), these variables obviously have very different distributions. Suppose we want to compare which one – income or years of schooling – has a larger statistical effect on our outcome. That is not possible using the variables we have. The solution is to standardize (i.e. calculate z-scores for) these two variables so that they are comparable.

Z-scores are expressed in terms of standard deviations from the mean. What we do is that we take a variable and "rescale" it so that it has a mean of 0 and a standard deviation of 1. Each individual's value on the standardized variable indicates its difference from the mean of the original (unstandardized) variable in number of standard deviations. A value of 1.5 would thus suggest that this individual has a value that is 1½ standard deviations above the mean, whereas a value of -2 would suggest that this individual has a value that is 2 standard deviations below the mean.

Function

Basic command	egen newvarname=std(oldvarname)	
Explanations	newvarname oldvarname	Insert the name of the new variable. Insert the name of the old variable.
Short names	Std	Standard deviation
More information	help egen	

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)

```
egen z_gpa=std(gpa)
```

```
egen z_cognitive=std(cognitive)
```

Now you have new versions – containing z-scores – of the two variables.

```
sum gpa z_gpa cognitive z_cognitive
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      9,380    3.178614    .6996298          1          5
       z_gpa |      9,380    2.87e-10           1  -3.113953   2.603357
   cognitive |      8,879    308.4708    72.18442        100        500
 z_cognitive |      8,879   -1.41e-09           1   -2.88803   2.653332
```

```
codebook gpa z_gpa cognitive z_cognitive, compact
```

```
Variable      Obs Unique      Mean        Min       Max  Label
-------------------------------------------------------------------------------------------------------
gpa          9380     41  3.178614          1         5  Grade point average (Age 15, Year 1985)
z_gpa        9380     41  2.87e-10  -3.113953  2.603357  Standardized values of (gpa)
cognitive    8879    101  308.4708        100       500  Cognitive test scores (Age 15, Year 1985)
z_cognitive  8879    101 -1.41e-09   -2.88803  2.653332  Standardized values of (cognitive)
-------------------------------------------------------------------------------------------------------
```

2.6 Recode

There are a lot of situations where recode is useful. For example: if you have continuous variable that you want to categorize, if you have a categorical variable for which you want to collapse categories, if you want to reverse the coding of a variable, or if you want to change any value(s) into missing.

For string variables, recode does not work. Instead, we can use replace.

More information	help recode
	help replace

2.6.1 Recode numeric variables

Function

Basic command	recode varname (rule)	
Useful options	recode varname (rule), gen(newvarname)	
Explanations	varname	Insert the name of the variable.
	(rule)	Specify which values you want to recode and how you want them to change.
	gen()	Add this if you want to generate a new variable with the recoding.
	newvarname	Name of the new variable.
More information	help recode	

Practical example

Dataset: StataData1.dta

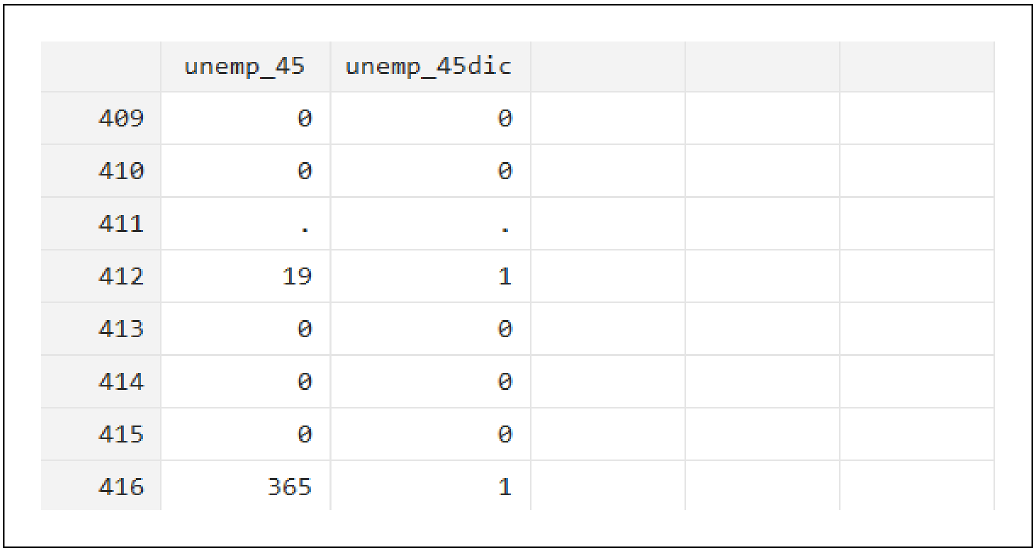
Name	**Label**
unemp_45	Days in unemployment (Age 45, Year 2015)

```
gen unemp_45dic=unemp_45
```

```
recode unemp_45dic (0=0) (1/365=1) (.=.)
```

The new variable unemp_45dic is a binary version of the original variable unemp_45, where all the individuals who had 0 days of unemployment at age 45 are given the value 0, and everyone who had 1-365 days of unemployment are given the value 1. Missing (".") is kept as missing.

```
browse unemp_45 unemp_45dic
```



	unemp_45	unemp_45dic			
409	0	0			
410	0	0			
411	.	.			
412	19	1			
413	0	0			
414	0	0			
415	0	0			
416	365	1			

2.6.2 Recode string variables

Recoding string variables builds on the same principle as for numeric variables. However, you need to use a command called replace instead of recode. Exactly as for numeric variables, it is preferable to generate a copy of the old variable before you start replacing values (or expressions, which is the term used below).

Note In this example, we are taking a sneak peek at if (which is described in more detail in Section 2.7).

Function

Basic command	replace varname="exp2" if varname=="exp1"	
Explanations	varname	Insert the name of the variable that you want to recode.
	exp1	Specify the value/expression that you want to change.
	exp2	Specify the value/expression that you want to change to.
More information	help replace	

Practical example

Dataset: StataData1.dta

Name	**Label**
marstat30	Marital status (Age 30, Year 2000)

First, let us have a look at this variable.

```
describe marstat30
```

```
              storage   display    value
variable name   type    format     label      variable label
---------------------------------------------------------------------------------------
marstat30       str20   %20s                  Marital status (Age 30, Year 2000)
```

```
tab marstat30
```

```
Marital status (Age |
      30, Year 2000) |      Freq.     Percent        Cum.
---------------------+-----------------------------------
                   D |        866        9.39        9.39
                   M |      5,120       55.54       64.94
                  UM |      3,206       34.78       99.72
                   W |         26        0.28      100.00
---------------------+-----------------------------------
               Total |      9,218      100.00
```

We can see that marstat30 is a string variable with four values specified (D, M, UM, and W). As it happens, we know that D=Divorced, M=Married, UM=Unmarried, and W=Widowed. This is what we want to change the values to.

```
replace marstat30="Divorced" if marstat30=="D"
```

```
replace marstat30="Married" if marstat30=="M"
```

```
replace marstat30="Unmarried" if marstat30=="UM"
```

```
replace marstat30="Widowed" if marstat30=="W"
```

```
tab marstat30
```

```
Marital status (Age |
      30, Year 2000) |      Freq.     Percent        Cum.
---------------------+-----------------------------------
            Divorced |        866        9.39        9.39
             Married |      5,120       55.54       64.94
           Unmarried |      3,206       34.78       99.72
             Widowed |         26        0.28      100.00
---------------------+-----------------------------------
               Total |      9,218      100.00
```

Now, it would be even easier to use encode to transform marstat30 into a numeric variable while retaining the values as value labels.

2.7 Condition the data with if

As a way of conditioning most other commands, we can use if. For example, you may want to get descriptive statistics only for those with a specific value on a variable (or several variables). You can also e.g. generate or recode a variable given certain properties of one or more other variables. There are simply so many things that if can be applied to, that it is impossible to do it justice with just a few examples.

More information	help if

Before we get into this, however, there are something that we should address first: logical operators.

Logical operators	
<	Less than
<=	Less than or equal to
==	Equal
>	Greater than
>=	Greater than or equal to
!=	Not equal to
&	And
|	Or
!	Not
()	Can be used for grouping to specify order or evaluation

2.7.1 Descriptive statistics with if

Practical example 1

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
sex	Sex

```
sum gpa if sex==0
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      4,752     3.07258    .6988425          1          5
```

```
sum gpa if sex==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      4,628    3.287489    .6836043        1.3          5
```

In the tables above, we see descriptive statistics for gpa, presented for men (sex==0) and women (sex==1) separately.

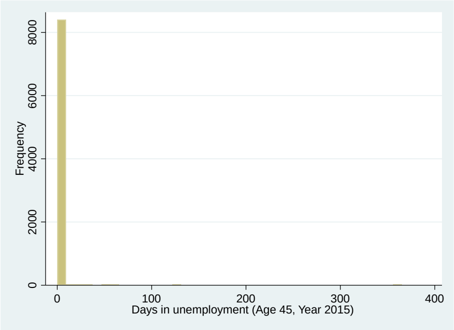
Practical example 2

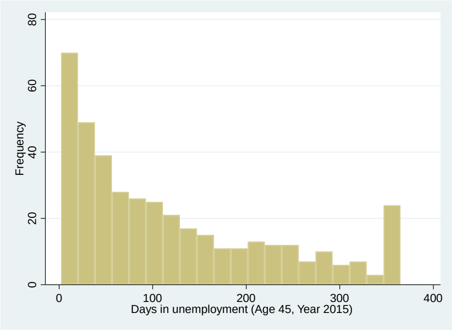
Dataset: StataData1.dta

Name	**Label**
unemp_45	Days in unemployment (Age 45, Year 2015)

```
histogram unemp_45, freq
```

```
histogram unemp_45 if unemp_45!=0, freq
```





The figure to the left shows a histogram for unemp_45 – not very useful since there are so many individuals with the value 0. In the figure to the right, the zeroes have been omitted so that it only shows individuals with at least one day in unemployment at age 45.

2.7.2 Recode with if

Practical example

Dataset: StataData1.dta

Name	**Label**
parmental	Parental mental illness (Ages 0-14, Years 1970-1984)
parcrim	Parental criminality (Ages 0-14, Years 1970-1984)

```
gen paradversity=.
```

```
recode paradversity (.=1) if parmental==0 & parcrim==0
```

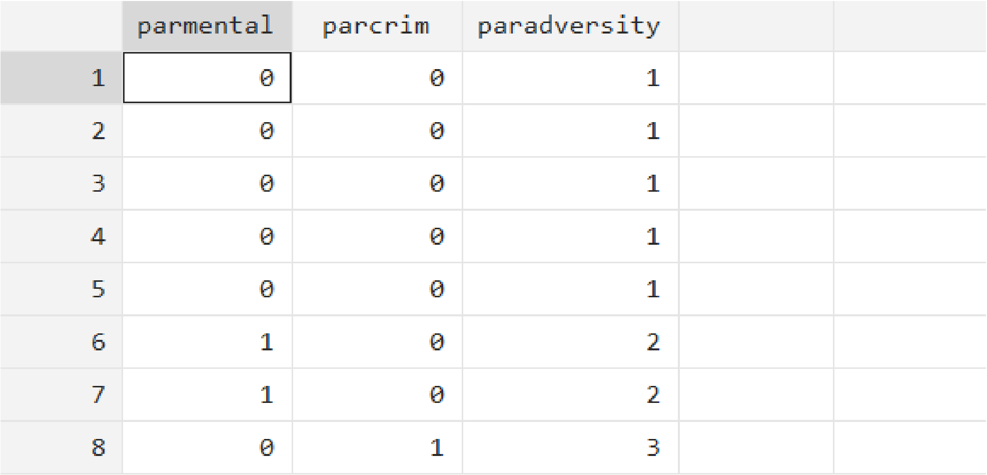
```
recode paradversity (.=2) if parmental==1 & parcrim==0
```

```
recode paradversity (.=3) if parmental==0 & parcrim==1
```

```
recode paradversity (.=4) if parmental==1 & parcrim==1
```

The new variable paradversity captures the four different combinations of parmental and parcrim.

```
browse parmental parcrim paradversity
```



	parmental	parcrim	paradversity
1	0	0	1
2	0	0	1
3	0	0	1
4	0	0	1
5	0	0	1
6	1	0	2
7	1	0	2
8	0	1	3

2.8 By

Similar to if, by can be used in combination with a lot of other commands. One particularly great application is to use it together with different types of graphs (e.g. bar charts, histograms, and scatterplots).

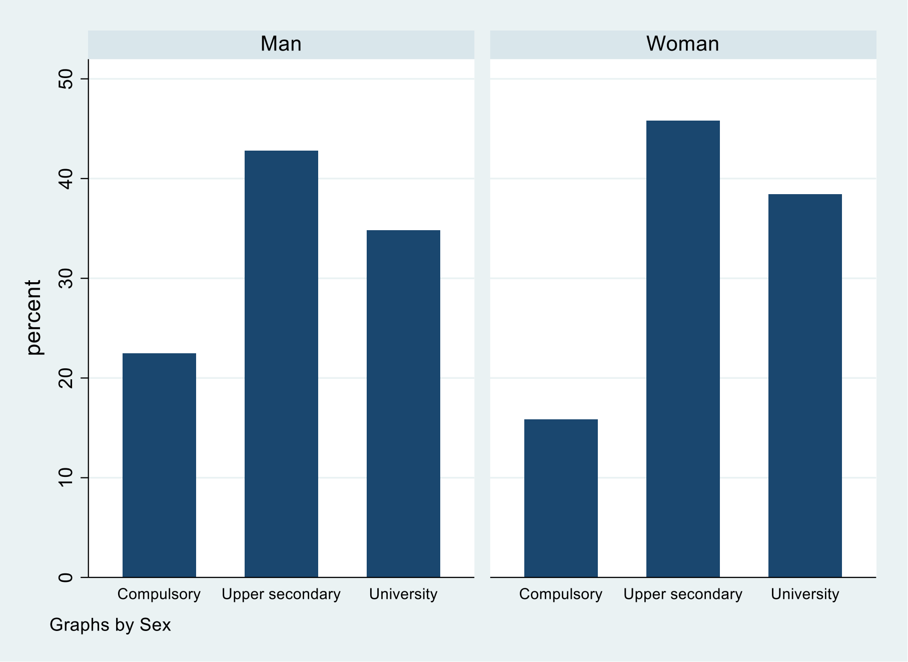
More information	help by

Practical example

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)
sex	Sex

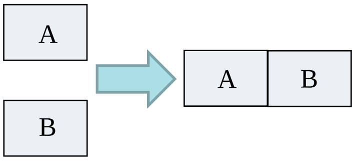
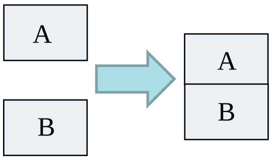
```
graph bar, over(educ) by(sex)
```



This application gives us two bar charts of educational level – separately for men and women.

2.9 Combining datasets

There are different ways of combining datasets in Stata, of which merge and append are the most useful ones. This is an illustration of the differences between these commands (A and B denote different datasets):

Merge	Append
	

2.9.1 Merge

Sometimes, it is necessary to combine two or more datasets. That is quite common for us working with register datasets, where different variables are kept in different files. For this purpose, it is possible to use merge.

More information	help merge

For merge to work, you need one or more variables to merge the datasets with. Most of the time, you have two datasets that contain the same number of individuals which are identified through an id variable.

Open the dataset that you want to merge something to, with the following command:

```
use "path\filename.dta"
```

Change path\filename to the full path (i.e. the folder on your computer that contains the file), and specify the file name, such as:

```
use "C:\Users\yerik\Stata Guide\TestDataMA.dta"
```

Then you can use the following command:

```
merge 1:1 varlist using "path\filename.dta"
```

1:1 means that you do a one-to-one merge on specified key variables. For varlist, you specify the variable(s) that you want to merge through.

Change path\filename to the full path to the dataset that you want to merge with (called the “using” dataset) the dataset that you have open (called the “master” dataset). For example:

```
merge 1:1 id using "C:\Users\yerik\Stata Guide\TestDataMB.dta"
```

This produces a variable in the first dataset that is called _merge. We also get a frequency table of this variable in the Results window. In our example, it looks like this:

```
Result                           # of obs.
-----------------------------------------
not matched                             0
matched                                10  (_merge==3)
-----------------------------------------
```

Thus, all 10 observations in the two datasets have been matched successfully.

Note Merging datasets can, of course, be a bit more complicated than this. If you have different amounts of individuals in the two datasets, you might need to use m:1, 1:m, or m:m, instead of 1:1 (m=many).

2.9.2 Append

If there is a situation when you would like to add individuals to a dataset, you can use append. This is, for example, useful if you want to combine different subsamples. Just consider first if you need to add a variable in your datasets that identifies which subsample the individuals belong to.

More information	help append

Open the dataset that you want to append something to, with the following command:

```
use "path\filename.dta"
```

Change path\filename to the full path (i.e. the folder on your computer that contains the file), and specify the file name, such as:

```
use "C:\Users\yerik\Stata Guide\TestDataAA.dta"
```

Then you can use the following command:

```
append using "path\filename.dta"
```

Change path\filename to the full path to the dataset that you want to append (called the “using” dataset) to the dataset that you have open (called the “master” dataset). For example:

```
append using "C:\Users\yerik\Stata Guide\TestDataAB.dta"
```

We suggest that you browse through your data next to make sure that everything worked out correctly.

Note Appending data can, of course, be a bit more complicated than this. Explore help, for more useful options.

3. BASIC STATISTICAL CONCEPTS

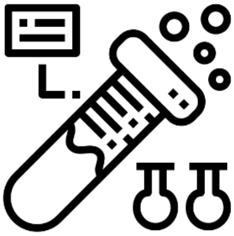
Outline

Content

The first part of this chapter is devoted to issues related to study designs, populations, and samples. These are things you need to be aware of in order to make correct judgements of your data material. Before it is possible to describe the variables in the dataset through the different commands in Stata, we need to know more about the specific variables. Here, we will address two major aspects: measurement scales and distributions.

3.1 Study design

There are two main types of study design: experimental design and observational design. With an experimental design, the researcher performs an intervention and measures the effect of this intervention on an outcome. In observational studies, the researcher observes the participants' outcome without performing any intervention.

Experimental design	Observational design
	

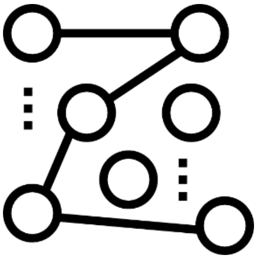
3.1.1 Experimental design

The literature distinguishes between many different types of experimental designs. Some terminology is focused on the location of the experiment: is it performed in a lab, out in the field, or is it a natural experiment? Other terms pay more attention to how individuals are assigned to different conditions/groups. For example, there is the pre-experimental design, where an intervention is performed for the entire sample of individuals (i.e. everyone is part of the so-called intervention group). Thus, there is no "control group" that one uses for comparison. The quasi-experimental design includes an intervention group *and* a control group, but the participants are not randomly assigned to the groups. Then we have the true experimental design: this includes a control group and an intervention group to which the participants are randomly assigned. Another term commonly used for the true experimental design, is randomised controlled trial (RCT).

Intervention group vs control group

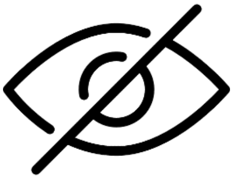
What is the purpose of dividing the participants into an intervention group and a control group? Well, the aim with an experimental study is to see whether the intervention (e.g. a treatment) has had any effect on the outcome. We therefore collect information before the intervention as well as after the intervention; has the outcome of interest changed across measurement points? What kind of participant conditions can predict this change? However, even if we can detect a change, how do we know that it is not by chance? It could be something else than the intervention that has caused this change. That is why we have a control group. If the same change is not observed in the control group, we can be more confident that the change in the intervention group is actually caused by the intervention.

Randomisation



So, why is it necessary to randomly assign individuals to the groups? This is because we want to be sure that a difference in the outcome between the intervention group and the control group is not simply because the groups are very different when it comes to the distribution of conditions (in epidemiology, one would be concerned with the equal distribution of risk factors in the two groups). Accordingly, randomisation has a great influence on the reliability of the results.

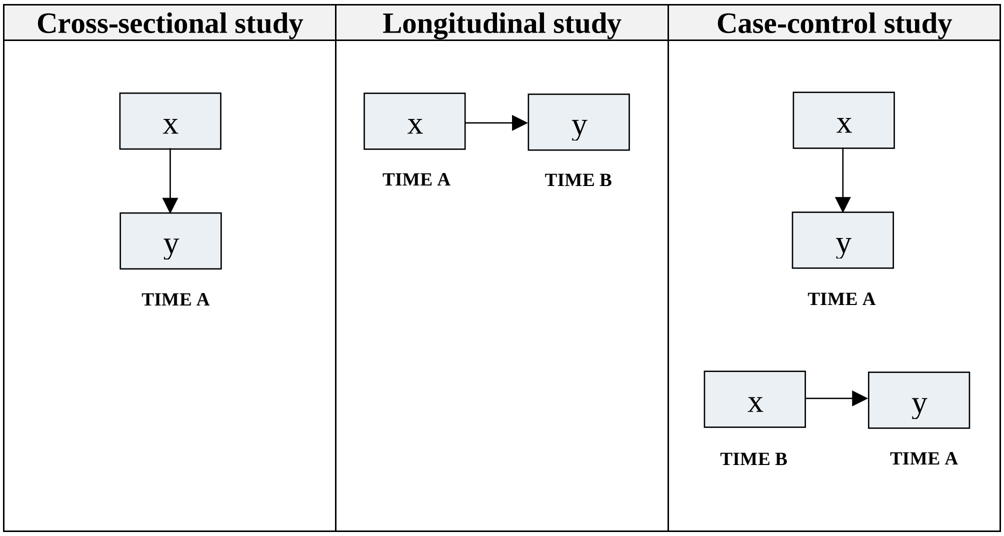
Blinding



Apart from randomisation, RCT also draws heavily on the concept of blinding. Some studies are blinded, meaning that the assignment to intervention group/control group is unknown for the participants, and sometimes also to those who provide the treatment and/or the researchers. This is to minimise the effect of expectations on the outcome. In the social sciences, blinding is usually not feasible.

3.1.2 Observational design

The three perhaps most common types of observational design are: cross-sectional studies, longitudinal studies, and case-control studies.



Cross-sectional studies

Cross-sectional studies are based on data that are collected at a single point in time. Thus, we measure the exposure and outcome simultaneously. This kind of study is perfect for estimating prevalence and can be used to explore patterns and associations in the data – but it does not allow us to draw any causal inferences (i.e. we cannot determine that the exposure is actually causing the outcome).

Longitudinal studies

In longitudinal studies, we have data from at least two measurement points: the exposure is measured at what is commonly called baseline, and the outcome at a later time, usually referred to as a follow-up.

There are at least three subtypes of longitudinal studies: trend studies, panel studies, and cohort studies.

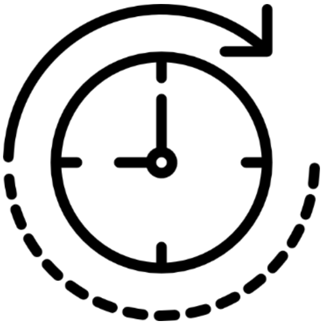
Trend study (repeated cross-section)	Following the same population over time. *Example*: Examining the prevalence of cannabis use among 15-year-olds in Sweden between 2009 and 2019.
Panel study	Following the same cross-section of individuals over time. *Example*: Exploring the association between exposure to bullying among those who were aged 10-18 in 2009 and the risk of cannabis use ten years later (2019).
Cohort study	Following the same cohort of individuals over time. *Example*: Studying the association between peer relationships in adolescence and drug dependence in adulthood among everyone born in Sweden in 1970.

Since time is such a fundamental concept in longitudinal studies, we also want to say something about longitudinal data and how these can be analysed to make full use of the detail. For example, when the outcome can occur at different points across the follow-up, we can use time-to-event analysis such as Cox regression (see Chapter 17). Often when our outcome is based on count data, we have collected information for a longer follow-up period; then we can use e.g. Poisson regression (see Chapter 16). But we might also model the longitudinal data in many other alternative ways. For example, individual developmental outcome patterns over time (in terms of the outcome’s frequency, duration, complexity, and sequencing) can be captured with methods such as latent growth models, group-based trajectory modelling, and sequence analysis. These methods are currently not covered by this guide.

Case-control studies

Then we have the case-control studies. Here, one focuses on a group that has a certain outcome (the “cases”), and matches them to a similar group without the outcome (the “controls”). These two groups are subsequently compared with regard to different exposures. Information about the exposure can reflect the same time point (i.e. cross-sectional data) or at an earlier time point (i.e. longitudinal data).

Retrospective vs prospective data



The terms retrospective and prospective are sometimes used a bit sloppy when it comes to observational studies. Basically, these terms refer to what one would assess first: the exposure or the outcome. In retrospective studies, the outcome is first established, after which one looks backwards in time to examine exposures. Thus, this is what one usually (but not always) do for case-control studies. In prospective studies, the exposure is first established, after which one looks forward in time for the outcome to occur. This is typically what we do in longitudinal studies (more specifically panel and cohort studies). It is nonetheless quite common that cohort studies are retrospective. This means that we define a cohort that has already experienced both the exposure(s) and the outcome(s) of interest, and we collect information on this through e.g. administrative records.

To make things more confusing, it is quite common that all sorts of observational studies (e.g. cross-sectional studies, cohort studies, and case-control studies) include retrospective questions, such as survey questions about past events and experiences. This is, however, not exactly what is meant by retrospective designs.

3.1.3 A comparison between study designs

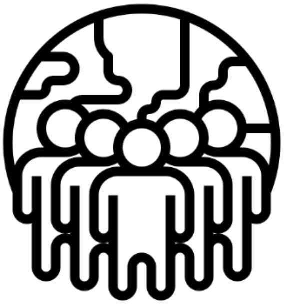
	Randomised controlled trial	**Cross-sectional study**	**Longitudinal study**	**Case-control study**
Population	Selected population, controlled environment	Mixed populations, varying contexts	Mixed populations, varying contexts	Mixed populations, varying contexts
Direction[a]	Exposure is introduced before the outcome is established	Exposure and outcome are measured simultaneously	Exposure is established before outcome is established	Outcome is established before exposure is established
Use[b]	Determine the effect of an intervention	Hypothesis testing, prevalence studies	Analyse associations over time	Analyse associations with rare outcomes
Analysis[c]	Simple, confounding taken care of by design	Sophisticated regression techniques, adjust for confounding	Sophisticated regression techniques, adjust for confounding	Sophisticated regression techniques, adjust for confounding
External validity /Generalisability	Low-medium	High	High	High
Internal validity /Causality[d]	High	Low	Medium	Low-medium

[a] See Chapter 9 for a discussion about exposures and outcomes.
[b] See Section 5.1 for information about hypothesis testing.
[c] See Section 9.3 for a discussion about confounding.
[d] See Section 9.4 for a discussion about causality.

3.2 Population and sampling

3.2.1 Population

When we design a study, we first need to establish what our population is, since the population is what we want to say something about. A population is often referred to by “N”.

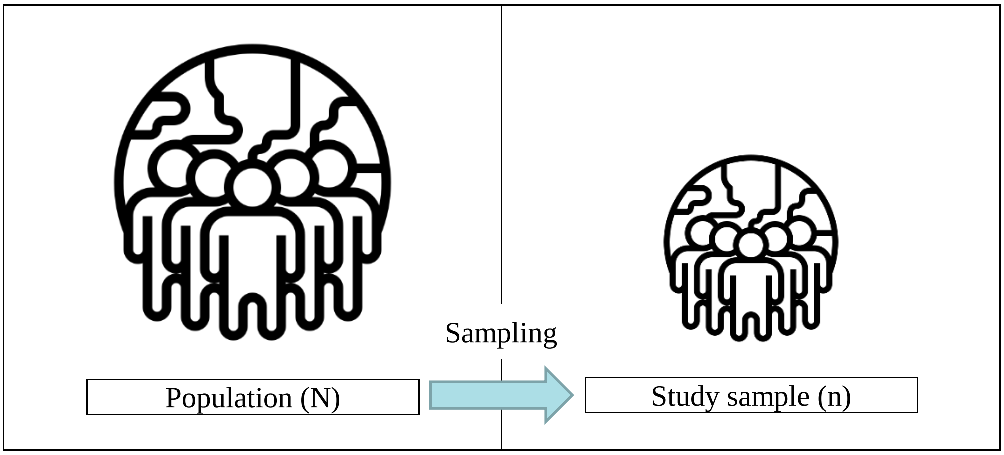


Population (N)

A population can be almost anything: We have populations which are geographically defined, such as the world, a country or a city; we have age-defined populations such as teenagers, infants and elderly, and also specific groups such as women, drug addicts, teachers, master students, and so on.

3.2.2 Sampling

It is seldom the case that we examine the whole population which we have chosen. Instead, we use sampling – that basically means that we take a smaller sample of the population: a study sample. A study sample is often denoted by “n”. The reasons behind sampling are primarily that it is very costly and time consuming to collect data for the entire population. However, sometimes you can include the whole population - like if you have small populations, such as one school or one hospital or one company (this is often referred to as a case study). Another example is when you use national registers (then you usually do not have to considered aspects such as time or cost since the data is already available).



There are many different sampling techniques available. Generally, they can be categorised into two types that include several sub types: non-probability sampling and probability sampling.

Non-probability sampling

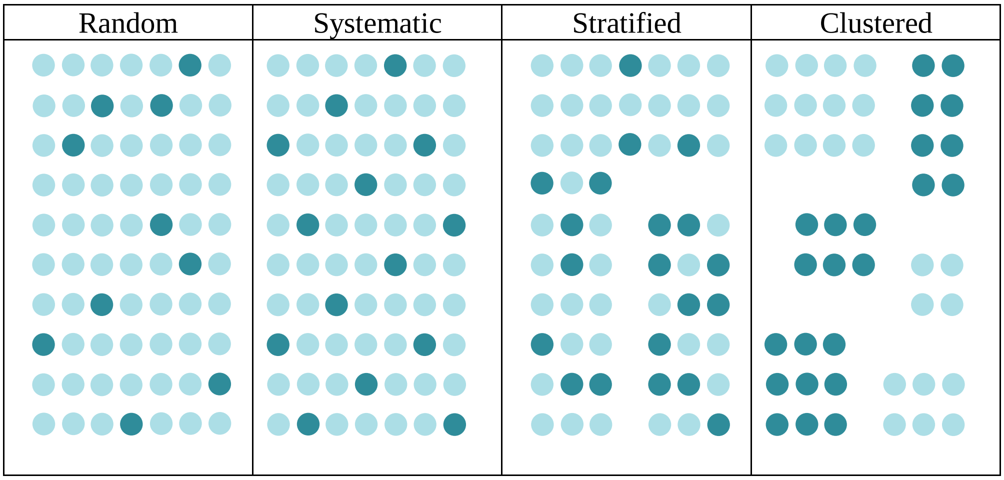
Types of non-probability sampling	
Snowball	Finding respondents through already selected respondents
Quota	Adding suitable individuals until a certain quota is achieved
Convenience	Easy access of respondents

Non-probability sampling is most common in small-scale studies, marketing research, interview studies, and studies like that. Snowball sampling means that you start out with some respondents and ask them to find other suitable respondents (like friends or other people they know). Quota sampling is often used in marketing research. For example, the researchers want to have 100 respondents who have tried a new coffee brand and stands outside the store until they have found 100 persons who have bought that specific brand. Then we have convenience sampling. This is when you pick respondents who are easy to get access to, like friends, family, or members of an organisation that you are a member of yourself, and so on.

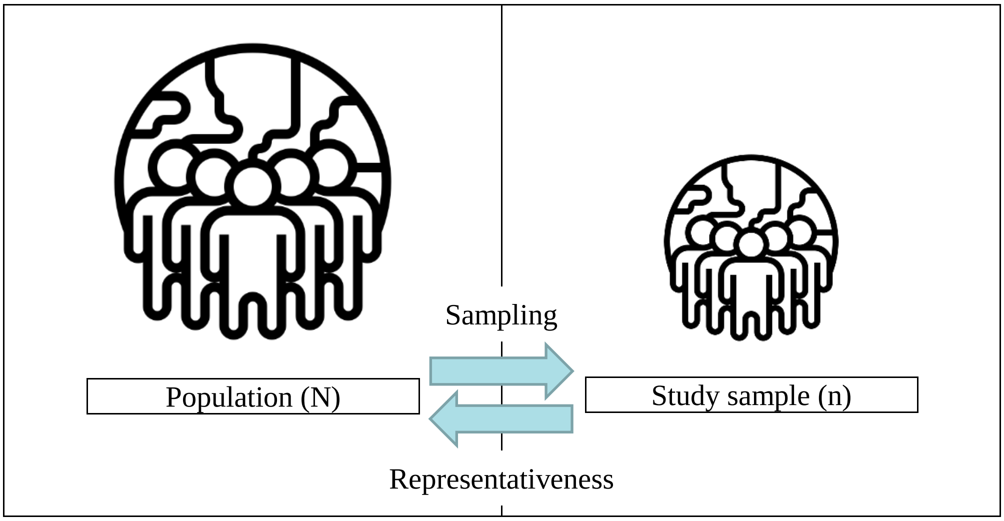
Probability sampling

Types of probability sampling	
Random	Every individual has the same chance of being selected
Systematic	Sampling with intervals, e.g. every fifth of a list
Stratified	Random sampling from different groups
Clustered	Random sampling of groups, choosing all individuals from these groups

When it comes to probability sampling, we first have the random probability sampling, which postulates that every individual in the population should have the equal chance of being selected. Another procedure is the systematic sampling, where you, for example, draw every fifth or seventh from a list of people. Stratified sampling is when you draw random samples from some specific groups. For example, if you want to compare labour market outcomes between native Swedes and immigrants, you may not get a large enough sample of immigrants if drawing a random sample from the entire population living in Sweden. Instead, you can draw a larger random sample from the smaller group. Finally, we have clustered sampling. Perhaps you start out by drawing a random sample of schools and then select all students attending ninth grade in these schools.



Probability sampling constitutes the foundation of quantitative data analysis. Why is it so important? Well, we want our study sample to be representative. This means that it should have the same characteristics as our population. This is a requirement to be able to draw conclusions about the population based on the study sample (also known as generalisability).

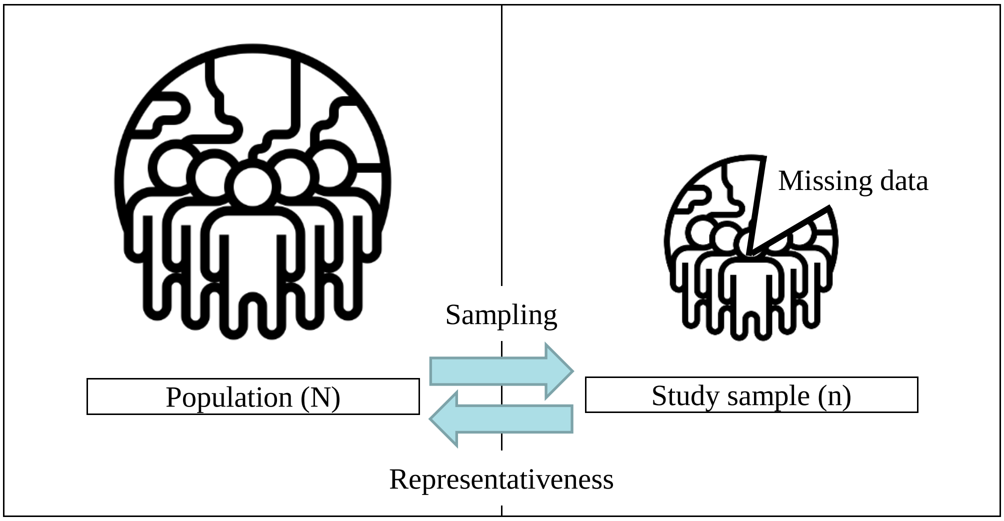


However, generalisability is not only about the ability to apply the results from the sample to the population:

It is possible to generalise the results from the sample…?	
To the population	Depends on the type of sampling. Can be assessed by comparing the sample characteristic with the population characteristics. In general: the bigger the sample, the better. *Example*: Our population consists of all children ages 2-5 in Sweden. We draw a random sample of 100 preschools and choose all children these preschools. Is the sample representative for the population?
Between populations	How unique is the population? Are there similar populations to which the results can be generalised? *Example*: Our population is defined as unaccompanied minors coming to Sweden from Iraq. Do our results also apply to unaccompanied minors from e.g. Afghanistan? Or to accompanied minors from Iraq?
Between interventions	Does the intervention (e.g. treatment) have to be exactly the same to generate the same results? What happens if we adjust the intervention? *Example*: Our population is defined as pregnant women who smokes. We include all pregnant women living in a specific Swedish city who reported that they were smoking at the time of enrolment in antenatal care. We randomise them into an intervention group and a control group. The intervention group participants in two hours of motivational interviews per week for two months. At the time of the child's birth, a higher proportion of women in the intervention group have stopped smoking compared to those in the control group. Would we see the same effect if we reduced the number of interviews?
Between contexts	How unique is the context? Is the study culture specific? *Example*: Our population is a total sample of all children (0-18) living in joint custody in Sweden. Can the results be applied to other countries, such as the United States?
Over time	Are the results specific for the historical time period for which we collected the data? Are the results and interpretations valid also for today? *Example*: Our population consists of all children who grew up in societal care in the 1960s. The results from our analysis suggests that these children experience much worse health in adulthood, compared to those who grow up with their biological parents. Would be find the same results if we were to follow-up children in societal care in the 2010s?

3.2.3 Missing data: attrition and non-response

An issue that almost all quantitative researchers deal with has to do with missing data. What is that? Well, when we have defined our population and conducted a probability sampling, we start collecting data for the individuals in our study sample – either through questionnaires or registers (or both). It is very seldom the case, however, that we get complete information for all individuals. We thus get missing data. When we use register data, missing data is commonly called attrition, and when we use survey data (i.e. questionnaire data), missing data is usually called non-response. If we have problems with missing data, we run into problems with representativeness, which may prevent us to draw conclusions about the population based on the study sample. This is discussed in further detail in Section 11.4.



3.3 Measurement scales

3.3.1 Types of scales

We use a scale to make the measurements of a variable, and the characteristics of the scale determine the characteristics of the data we collect and, in turn, how we describe our data. Generally speaking, there are four measurement scales: nominal, ordinal, ratio and interval. Nominal and ordinal variables are often called categorical (or qualitative), whereas ratio and interval variables are often referred to as continuous (or quantitative).

Name	Type
Nominal	Categorical/qualitative
Ordinal	
Ratio	Continuous/quantitative
Interval	

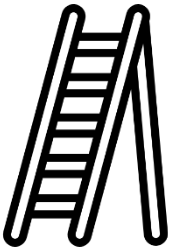
It should also be noted that a nominal variable with only two categories/values is called dichotomous (or binary, or dummy) whereas a nominal variable with more than two categories is called polytomous.

3.3.2 Differences between the scales

These scales differ in three important ways: hierarchy, distance, and zero point.

Checklist	
Is it possible to arrange/order the values hierarchically?	Yes/No
Is it the same distance between the values?	Yes/No
Does the scale have an absolute zero point?	Yes/No

Hierarchy

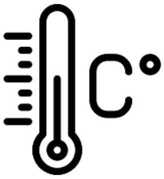


What does "arrange/order the values hierarchically" mean? If we take gender as an example, it is not reasonable to say that "Man" is less or more than "Woman". As another example, we can take nationality: it is not reasonable to see "Danish" as less or more than "Finnish". For variables such as self-rated health, on the other hand, it is possible to say that "Excellent health" is better than "Good health". Moreover, it is possible to say that the grade "A" is better than the grade "B".

Distance

What does "distance" mean? If we take income as an example, we know that 1000 dollars are twice as much as 500 dollars, and 2000 dollars are twice as much as 1000 dollars. The same logic applies to variables such as age: it is the same distance between 2 years and 4 years as between 6 years and 8 years. Thus, having the same distance between the values means that the differences between two values are the same regardless of which part of the scale you are looking at.

Zero point



What does “absolute zero point” mean? Basically, it means that the scale cannot have negative values. It is possible for the temperature to be minus 10 degrees Celsius, but is not possible to have less than zero years of schooling or having less than zero days of unemployment.

Examples

Below, we can see some examples of variables on the different measurement scales.

Scale	Values	Examples
Nominal	Order values: No Same distance: No Absolute zero point: Not applicable	Yes/no questions Gender Nationality
Ordinal	Order values: Yes Same distance: No Absolute zero point: Not applicable	Attitude questions Self-rated health Educational level
Ratio	Order values: Yes Same distance: Yes Absolute zero point: Yes	Age Income School marks
Interval	Order values: Yes Same distance: Yes Absolute zero point: No	Temperature (Celsius) Calendar time

A nominal variable is hence a variable for which the values cannot be ranked, and we do not have the same distance between the values, e.g. gender or questions that can be answered with yes or no. Ordinal variables are similar, but here the values can be ranked, such as for self-rated health: “Excellent is better than “Good”; “Good” is better than “Fair”; and “Fair” is better than “Poor”. However, for ordinal scales we do not have the same distance between the values: the “amount” of better health is not necessarily the same between “Poor” and “Fair” as between “Good” and “Excellent”. The ratio scale is similar to the ordinal scale, but here we do have the same distance between the values: for example, we know that 10 years of schooling is twice as much as 5 years of schooling. The interval scale is similar to the ratio scale, but here we do not have an absolute zero point.

3.3.3 Types of values

It is possible to distinguish between two types of values: discrete and continuous. Discrete values can only assume “whole” values, such as “Man”, “Women”, “Green”, “Car”, and “House”. Continuous values can assume any value along a scale, such as “3.5 years”, “58.3 seconds”, and “163.5 centimetres”. Note, however, that continuous variables (i.e. on a ratio or interval scale) do not necessarily have continuous values. For example, number of cars is a ratio variable but it has discrete values: while the average number of cars in a population may be 0.8, it is not correct (although many do) to say that any given individual in a population has 0.8 cars (since a car is a “whole” value).

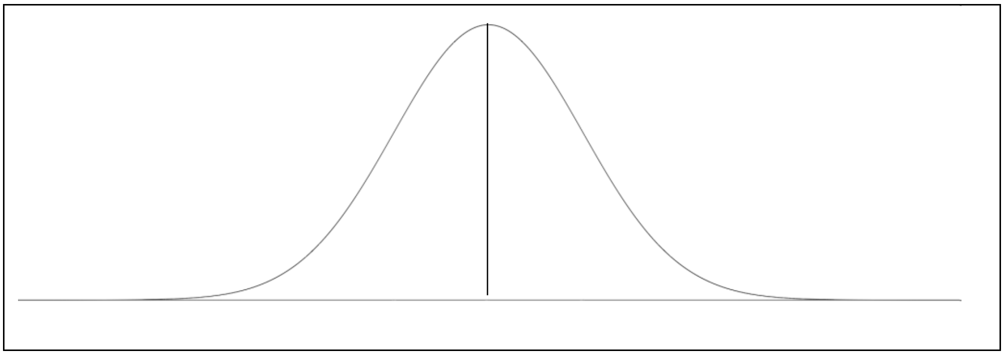
Name	Type
Discrete	“Whole” values
Continuous	Any value

3.4 Distributions

For continuous variables (i.e. on a ratio or interval scale) it is important to know what the distribution of values in the variable looks like.

3.4.1 Normal distribution

One common type of distribution is the “normal distribution”. Many statistical methods are based on normal distributions. Please note that “normal” in this setting should be interpreted as something that is typical (or regular), not as something that is natural.

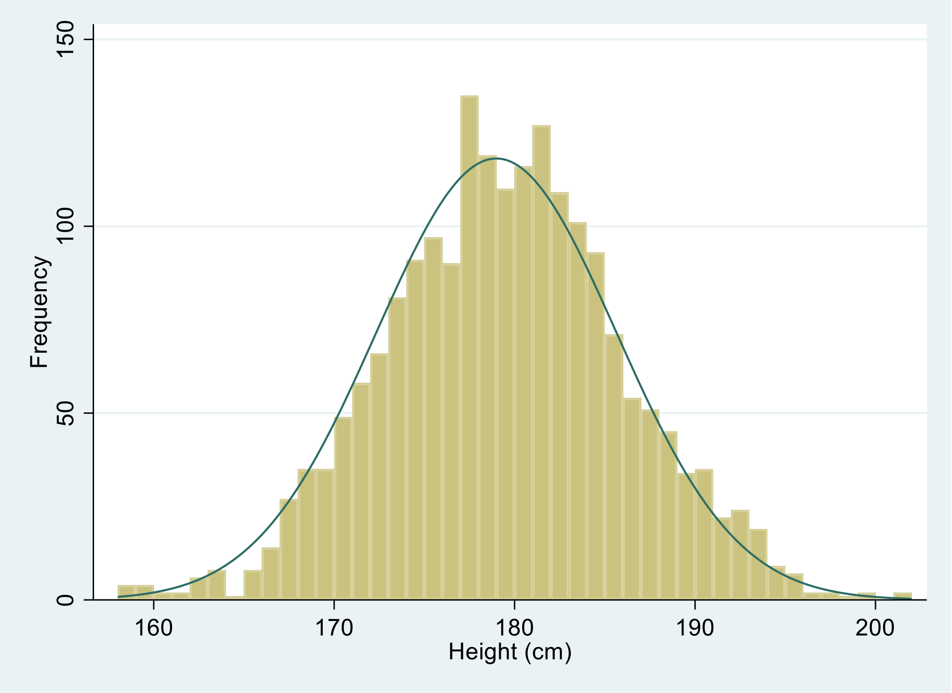


The above figure is a typical example of a normal distribution. Here are some basic facts about the normal distribution:

Basic facts about the normal distribution
Always bell-shaped.
The peak always indicates the mean value.
Always symmetrical, i.e. the tails on each side of the mean are equally large. This means that 50% of the values are on one side of the mean, and 50% of the values are on the other side of the mean.
The area under the curve is always 1 (100% of the values).

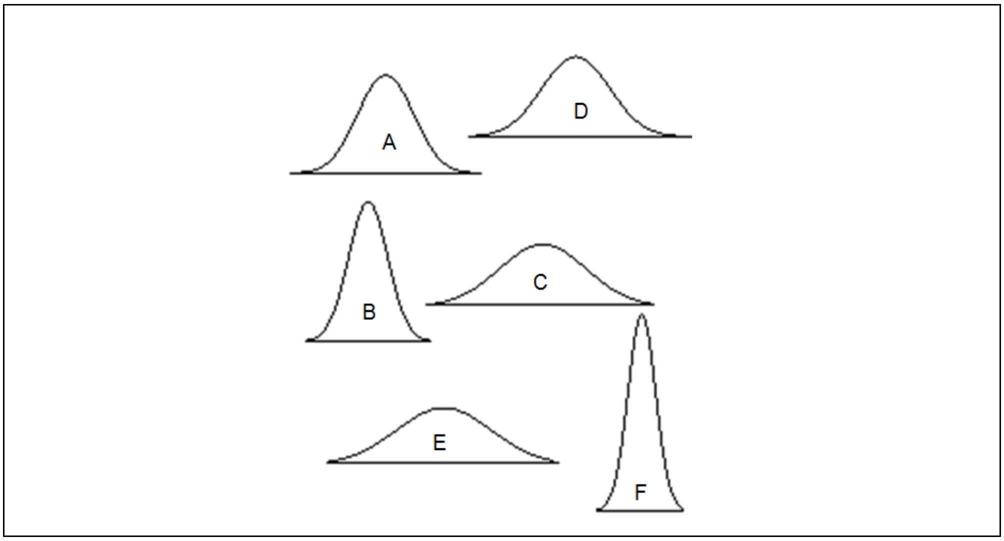
More about peaks

Below is an example of a (normal) distribution of height among Swedish men at the time of military service enlistment (in Swedish: “lumpen”). In this example, the mean height is about 179 centimetres. The less common a certain height gets, the smaller the area under the curve. Here, the tails are about equally large on both sides of the mean, suggesting that it is approximately as common for individuals in the sample to be shorter than the mean as it is for them to be taller than the mean.



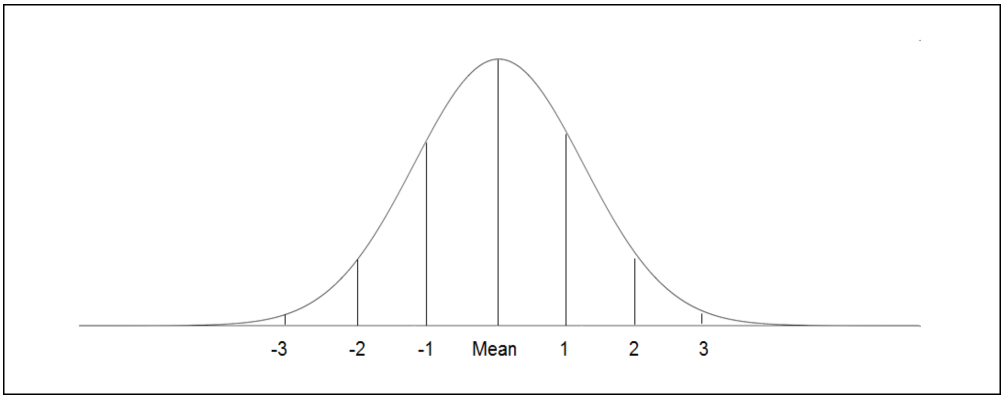
Why peaky or less peaky?

Normal distributions can look quite different. The figures below are all examples of normal distributions. The difference lies in the amount of spread of the values: because the shape of a normal distribution is not only defined by the mean value, but by the standard deviation!



What is standard deviation?

A simple definition of standard deviation is that it expresses how much variation exists from the mean for a given variable (see Section 4.6.2 for further discussion). If we have a small standard deviation, it suggests that the individuals in our data have values close to the mean, and if we have a large standard deviation, it indicates that the values are more spread out over a large range of values.



Empirical rule

The empirical rule of normal distributions tells us the following (see the figure above):

- 68% of the values fall within -1 and +1 standard deviations.
- 95% of all values fall within -2 and +2 standard deviations.
- Nearly 100% of all values fall within -3 and +3 standard deviations.

Example

We have collected information about weight for a sample of individuals. If the mean weight in this sample was 70 kilos and the standard deviation was 5 kilos, the empirical rule would give us the following information:

68% of the individuals have a weight of 65-75 kilos:
Lower limit: 70 kilos - (5 kilos*1); upper limit: 70 kilos + (5 kilos*1)

95% of the individuals have a weight of 60-80 kilos:
Lower limit: 70 kilos - (5 kilos*2); upper limit: 70 kilos + (5 kilos*2)

Nearly 100% of the have a weight of 55-85 kilos:
Lower limit: 70 kilos - (5 kilos*3); upper limit: 70 kilos + (5 kilos*3)

As long as we have information about the mean value and the standard deviation, it is possible to do the same calculation for all the normal distributions. Remember that a more pronounced peak indicates a low standard deviation, whereas a flat distribution indicates a high standard deviation.

3.4.2 Skewed distributions

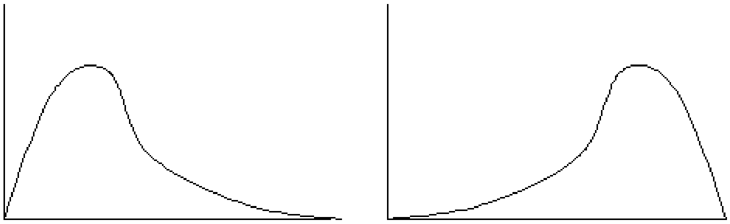
There are other types of distribution. One very common type of distribution is the skewed distribution. Here are some facts about skewed distributions:

Basic facts about skewed distributions
Always asymmetrical = Tails are different, i.e. the empirical rule does not apply.
Skew can be positive (right tail longer) or negative (left tail longer).

Positive or negative?

Examples of a positively skewed distribution (like the figure to the left) are: number of hospital visits, number of days in unemployment, number of telephone calls during a day. Most individuals will have the value zero or a low value, whereas a few will have increasingly high values.

Examples of a negatively skewed distribution (like the figure to the right) are: age of retirement, or a very easy test. Most individuals will have high values, and then a few will have very low values.



Positively skewed distribution Negatively skewed distribution

Measures of skewness

The skewness of the distribution can be indicated by two types of measure: skewness and kurtosis.

Facts about the skewness measure
Measure of the symmetry of a distribution.
Negative skewness value = Longer tail to the left.
Positive skewness value = Longer tail to the right.
A perfect normal distribution has a skewness of 0.
Skewness value between -2 and +2 is usually considered acceptable.

Facts about the kurtosis measure
Measure of the shape (or the “peakedness”) of a distribution.
A perfect normal distribution has a kurtosis of 0 (mesokurtic distribution).
Kurtosis value above 0 = Leptokurtic distribution (sharper peak and longer/fatter tails).
Kurtosis value below 0 = Platykurtic distribution (rounder peak and shorter/thinner tails).
Kurtosis value between -2 and +2 is usually considered acceptable.

4. DESCRIPTIVE ANALYSIS

Outline

Content

When we know about measurement scales and distributions, we can decide on how to best describe our variables. In this chapter, we go through a set of tables and graphs as well as measures of central tendency and variation. We will first cover frequency tables. With regard to graphs, we will discuss bar charts, pie charts, and histograms. For measures of central tendency, the mean, mode, and median are addressed. Moreover, some examples of measures of variation will be included here, namely minimum, maximum, variance, and standard deviation.

The chapter also includes a quite extensive section on epidemiological concepts, such as prevalence and incidence.

Finally, we end with a section on how to design informative tables and graphs for descriptive statistics.

4.1 Introduction

Going back to what we learnt about measurement scales and the distributions, this is generally how you could match the different types of variables with the different types of description:

Type of variable	
Categorical (nominal/ordinal)	Frequency table Bar chart Pie chart Mode
Continuous (ratio/interval)	Histogram Mean (if normal distribution) Median (if skewed distribution) Minimum Maximum Variance Standard deviation

4.2 Frequency table

Quick facts	
Number of variables	One
Scale of variable(s)	Categorical (nominal/ordinal)

A frequency table is a simple but very useful description of one variable and gives us both the frequency and various types of percentages of individuals with the different values.

This function is used primarily for categorical variables (i.e. nominal/ordinal) but can be used for any type of variable; the main concern is that the table becomes too lengthy if there are many categories/values in the variable.

The following information is included in the frequency table:

Types of statistic		
Freq.	Frequency	The number of individuals in the different categories of the variable.
Percent	Percent	The percentage distribution of individuals in the different categories of the variable.
Cum.	Cumulative percent	Adds the percentages from top to bottom.

Note Frequency tables do not automatically include information about missingness (but it is available as an option).

Function

Basic command	tab varname	
Useful options	tab varname, m tab varname, nol tab varname, sort	
Explanations	varname m nol sort	Insert the name of the variable you want to use. Treat missing values like other values. Display numeric codes rather than value labels. Display the table in descending order of frequency.
Short names	tab m nol	tabulate missing nolabel
Notes	Options can be used simultaneously, e.g: tab varname, m nol sort	
More information	help tabulate oneway	

Note You can tab multiple variables at the same time by using tab1 (for example: tab1 varname1 varname2 varname3)

Practical example

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)

```
tab educ
```

```
   Educational |
  level (Age 40, |
     Year 2010) |      Freq.     Percent        Cum.
----------------+-----------------------------------
     Compulsory |      1,763       19.20       19.20
Upper secondary |      4,062       44.23       63.43
     University |      3,358       36.57      100.00
----------------+-----------------------------------
          Total |      9,183      100.00
```

This is the simplest form of a frequency table. It shows the frequencies, the percentage distribution, and the cumulative percentages. In this particular example, we see the distribution of educational level. Here, we are mostly interested in the column called Percent. It shows that 19.2% of the sample have compulsory education, 44.2% have upper secondary education, and 36.6% have university education. This then actually tells us something about the mode/type value; it is the most common value – which in this case is upper secondary education.

```
tab educ, m
```

```
    Educational |
 level (Age 40, |
     Year 2010) |      Freq.     Percent        Cum.
----------------+-----------------------------------
     Compulsory |      1,763       17.63       17.63
Upper secondary |      4,062       40.62       58.25
     University |      3,358       33.58       91.83
              . |        817        8.17      100.00
----------------+-----------------------------------
          Total |     10,000      100.00
```

The table above includes missing values. We can see that 8.2% of the original sample has missing values for this variable.

```
tab educ, nol
```

```
Educational |
 level (Age |
   40, Year |
      2010) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,763       19.20       19.20
          2 |      4,062       44.23       63.43
          3 |      3,358       36.57      100.00
------------+-----------------------------------
      Total |      9,183      100.00
```

This table omits the value labels, and instead shows the actual values.

```
tab educ, sort
```

```
    Educational |
 level (Age 40, |
     Year 2010) |      Freq.     Percent        Cum.
----------------+-----------------------------------
Upper secondary |      4,062       44.23       44.23
     University |      3,358       36.57       80.80
     Compulsory |      1,763       19.20      100.00
----------------+-----------------------------------
          Total |      9,183      100.00
```

And here we can see the categories sorted from the most common one (upper secondary) to the least common one (compulsory).

4.3 Bar chart

Quick facts	
Number of variables	One
Scale of variable(s)	Categorical (ordinal)

A bar chart is like an illustration of a frequency table. On the x-axis (horizontal axis) you see the different values (or categories) of the variable and on the y-axis (vertical axis) you can choose to see either the percentage of individuals in each category (like in the graph below) or the number of individuals in each category.

As mentioned above, the bar chart is useful primarily for categorical variables (preferably ordinal, since the bars suggest that values are ranked) but can be used for any type of variable as long as it does not have too many values.

Function

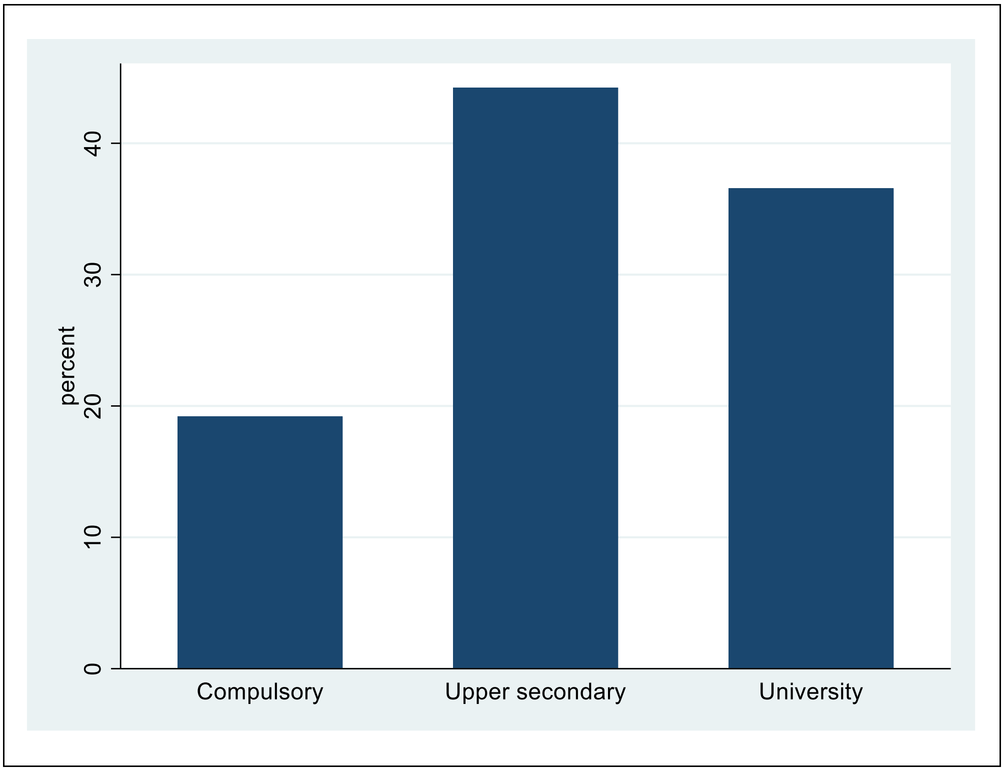
Basic command	graph bar, over(varname)	
Explanations	varname	Insert the name of the variable you want to use.
More information	help graph bar	

Practical example

Dataset: StataData1.dta

Name	Label
educ	Educational level (Age 40, Year 2010)

```
graph bar, over(educ)
```



The figure above is a bar chart for the variable educ. On the y-axis (vertical axis) we have percentages, and on the x-axis (horizontal axis), we have the different categories of the variable. It is rather easy to see that the category “Upper secondary” is the most common category, followed by “University” and then “Compulsory”.

Note You can use the Graph Editor (see Section 2.1.4) to edit the bar chart.

4.4 Pie chart

Quick facts	
Number of variables	One
Scale of variable(s)	Categorical (nominal)

Similar to a bar chart, a pie chart can also be seen as a simple illustration of a frequency table. The slices represent the different values (or categories) of the variable and they can be specified in terms of the percentage of individuals in each category or the number of individuals in each category.

This function is used only for categorical variables (preferably nominal, since it makes more sense to illustrate non-ranked values with slices than with bars). It is also recommended that the variable has relatively few categories – otherwise the pie chart will get too complex.

Function

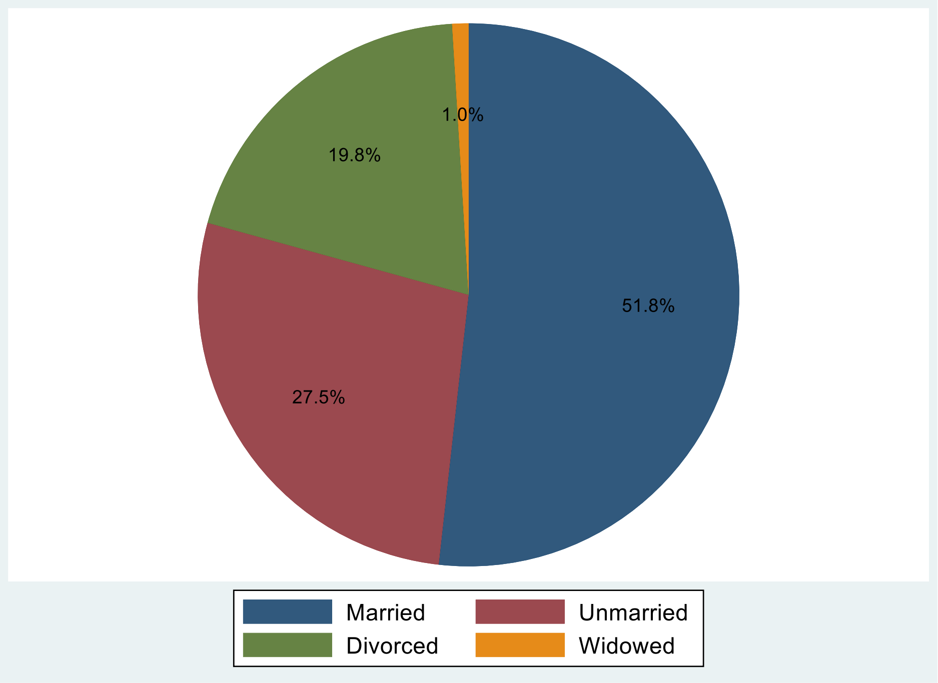
Basic command	graph pie, over(varname)	
Useful options	graph pie, over(varname) plabel(_all percent)	
Explanations	varname	Insert the name of the variable you want to use.
	plabel(_all percent, format(%12.1f))	Show the percentage distribution on the slices, with one decimal.
More information	help graph pie	

Practical example

Dataset: StataData1.dta

Name	**Label**
marstat40	Marital status (Age 40, Year 2010)

```
graph pie, over(marstat40) plabel(_all percent, format(%12.1f))
```



The figure above is a pie chart for the variable marstat40. It is rather easy to see that the category "Married" is the most common category (51.8%), followed by "Unmarried" (27.5%), "Divorced" (19.8%), and "Widowed" (1%).

Note You can use the Graph Editor (see Section 2.1.4) to edit the bar chart.

4.5 Histogram

Quick facts	
Number of variables	One
Scale of variable(s)	Continuous (ratio/interval)

A histogram is similar to a bar chart but, unlike the bar chart, it is suitable for continuous variables. The histogram will give us an idea about whether the distribution (of the continuous variable) is normal or skewed. It is also possible to include a normal curve in the chart in order to see how the data adheres to a normal distribution.

Function

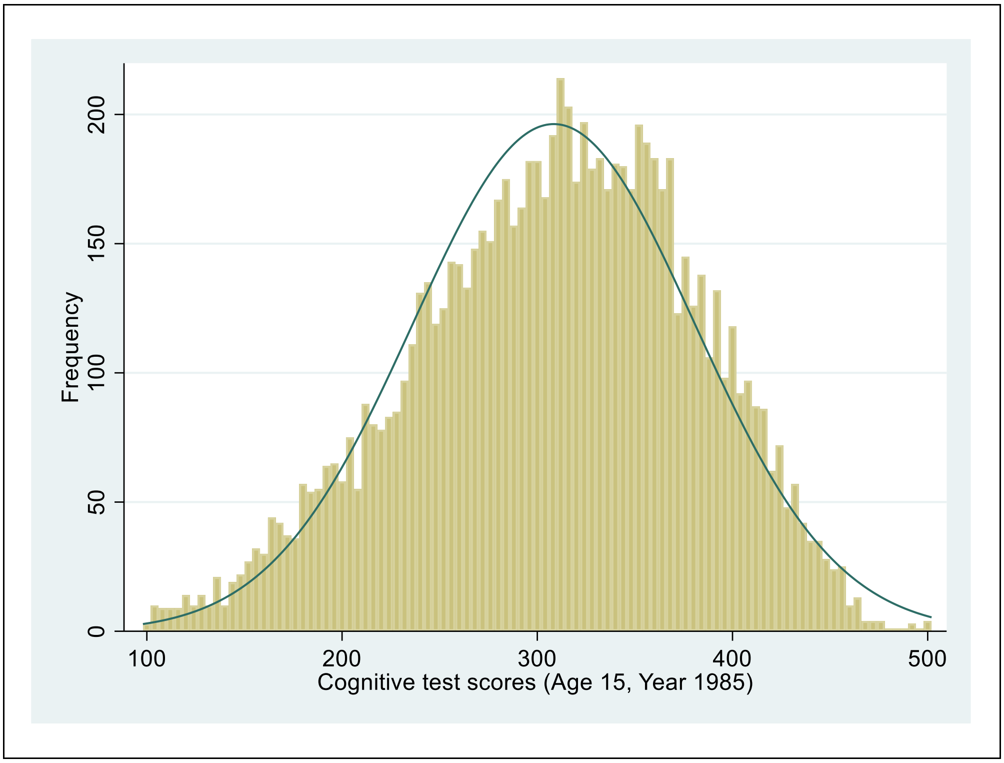
Basic command	histogram varname, freq	
Useful options	histogram varname, freq norm histogram varname, freq norm bin(x) histogram varname, freq norm d	
Explanations	varname	Insert the name of the variable you want to use.
	freq	Show frequencies on the y-axis.
	norm	Include a normal curve in the histogram.
	bin(x)	Here you can specify how many bins you want to histogram to show; might require some experimenting.
	d	Specify that data are discrete.
Short names	freq	frequencies
	norm	normal
	d	discrete
More information	help histogram	

Practical example

Dataset: StataData1.dta

Name	**Label**
cognitive	Cognitive test scores (Age 15, Year 1985)

```
histogram cognitive, freq norm d
```



This is a histogram of cognitive. The x-axis (horizontal axis) represents the values of the variable. The y-axis (vertical axis) represents the number of individuals. The line displays the normal curve.

Note You can use the Graph Editor (see Section 2.1.4) to edit the bar chart.

4.6 Measures of central tendency and variation

4.6.1 Central tendency

Central tendency can be defined as measures of the location of the middle in a distribution. The most common types of central tendency are:

Measure	Definition
Mean	The average value
Median	The value in the absolute middle
Mode	The most frequently occurring value

Mean

The mean is perhaps the most commonly used type of central tendency and we get it by dividing the sum of all values by the number of observations.

Example

We have four fishes that weigh:

1.1 kilo	0.8 kilo	1.1 kilo	1.0 kilo

What is the mean?
First, we add the values together: 1.1+0.8+1.1+1.0=4.0
Then we divide the sum of the values by the number of fishes: 4.0/4=1.
The mean is thus 1 kilo.

Median

The median – i.e. the value in the absolute middle of the distribution – is obtained by sorting all the observations' values from low to high and then identifying the value in the middle of the list.

Example

We have nine individuals who are of the following heights:

☺	☺	☺	☺	☺	☺	☺	☺	☺
158 cm	159 cm	164 cm	165 cm	173 cm	174 cm	175 cm	179 cm	181 cm

The median is thus 173 cm.

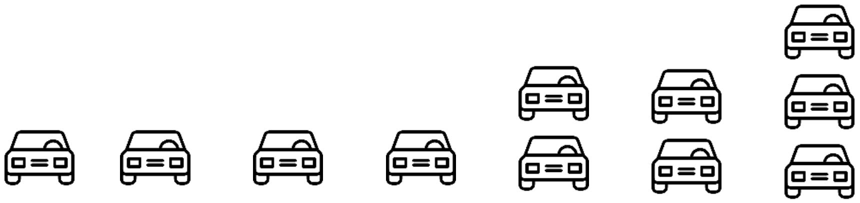
Note When we have an odd number of values, it is easy to identify the value in the absolute middle of the distribution. When we have an even number of values, we get the median by adding the two values in the middle together and dividing the sum by 2.

Mode

The mode – or type – is defined as the most frequently occurring value in a distribution. Here as well, one starts by sorting observations from the lowest to the highest value and then identifies the most common value.

Example

We have information about the number of cars in each of seven households:



Household 1 1 car	Household 2 1 car	Household 3 1 car	Household 4 1 car	Household 5 2 cars	Household 6 2 cars	Household 7 3 cars

The mode is thus 1 car (since this is the most common value).

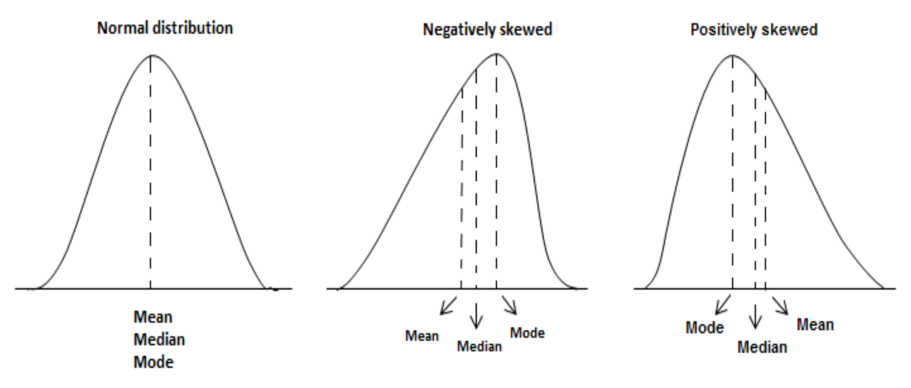
How to choose between mean, median, and mode?

The choice of type of central tendency is based on a) the measurement scale of the variable and b) the distribution of the variable. Generally, if the variable is categorical (nominal or ordinal), the mode is preferred. If the variable is continuous (ratio or interval), the mean or the median is preferred. In the latter case, the mean is chosen if the variable is normally distributed and the median is chosen if the variable has a skewed distribution.

Scale	Type	Central tendency
Nominal	Categorical	Mode
Ordinal		
Ratio	Continuous	Normal distribution: Mean Skewed distribution: Median
Interval		

Why should one not use the median or the mean for categorical variables? For nominal variables, it is easy to give an answer. Let us take country of birth as an example. In this example, the variable is coded into four categories: 1) Sweden, 2) China, 3) Canada, and 4) Norway. This is clearly a nominal variable. Since the order of the categories is random (i.e. the order of the categories does not really matter), the location of the absolute middle in the distribution would not tell us anything information about the variable: the "content" of the middle would change completely if we changed the order of the categories. Let us take gender (which is also on a nominal scale) as another example: it would not make any sense to give the mean or median of gender. For some ordinal variables, however, the median is *sometimes* used. For example, if we have five categories of occupational class, which can be ranked from lower class to upper class, it may be interesting to give the value of the median (for example, in this case, the median could be lower non-manuals which would tell us something about the distribution of values).

Why is it important to consider the distribution of the variable for continuous variables before we decide on the type of central tendency? If we take a look at the figures below, we can draw the following conclusions: if we have a perfectly normally distributed variable, the mean, median and mode would all be the same. However, if the distribution is skewed, the median would be a better description of the location of the middle in the distribution.



4.6.2 Variation

Besides the mean, the median and the mode, we may use some measures of variation to describe our variables further. Here are some of the most common measures of variation:

Measure	Definition
Minimum	The lowest value
Maximum	The highest value
Variance	The average of squared deviations from the mean value
Standard deviation	The squared root of the variance

These measures are most suitable for continuous variables (i.e. ratio or interval) but sometimes minimum and maximum are used for ordinal variables as well. However, they cannot be used for nominal variables (for the same reason as why we do not use mean or median to describe nominal variables).

The minimum and maximum are rather self-explanatory, but what about variance and standard deviation? Below, these measures are discussed in more detail.

Variance and standard deviation

Both variance and standard deviation are measured used to describe the dispersion (spread) of data around the mean value of a variable. To calculate the variance, we do the following:

Step	Calculation
1	Calculate the mean of the variable (the sum of all values, divided by the number of observations).
2	Subtract the mean from each value. These differences are often called deviations. Values below the mean will have negative deviations whereas values above the mean will be positive deviations.
3	Square each deviation to make it positive.
4	Add the squared deviations together. Divide by the number of observations.

However, the variance is quite difficult to interpret. That is why most would prefer to express dispersion in terms of standard deviation instead. To do this, we just add one more step:

Step	Calculation
5	Take the square root of the variance.

Population vs sample

The above calculations are based on the idea that the data we use encompass the entire population that we want to study. This is perhaps seldom the case; often we have drawn a sample from our population. Under such circumstances, we need to make a small adjustment (in *italics*):

Step	Calculation
1	Calculate the mean of the variable (the sum of all values, divided by the number of observations).
2	Subtract the mean from each value. These differences are often called deviations. Values below the mean will have negative deviations whereas values above the mean will be positive deviations.
3	Square each deviation to make it positive.
4	Add the squared deviations together. Divide by the number of observations *minus 1*.
5	Take the square root of the variance.

We will not go into detail regarding why we adjust Step 4, as described above. But basically, it has to do with the distinction between parameters and statistics: for populations, we can calculate parameters (fixed, “true” value), whereas for samples, we can calculate statistics (dependent on the selected sample, estimated value). We use “observations minus 1” (usually expressed as “n-1”) to produce as a less biased estimate. Want to know more? Read up on Bessel’s correction.

4.6.3 Summarize

Quick facts	
Number of variables	At least one
Scale of variable(s)	Continuous (ratio/interval) or ordinal

To generate descriptive statistics for your variables, you can use the summary command (or sum, for short). It is used primarily for continuous variables (i.e. ratio/interval) but could also be used for some ordinal variables that are approximately continuous (e.g. rating measures). The Stata function gives you the following statistics:

Types of statistic	
Obs	Number of observations
Mean	Mean value
Std. Dev.	Standard deviation
Minimum	Minimum (smallest) observed value
Maximum	Maximum (largest) observed value

If you combine sum with the option detail, you will additionally get the following statistics:

Additional types of statistic	
Median	Median
Variance	Variance
Kurtosis	Kurtosis and standard error of kurtosis
Skewness	Skewness and standard error of skewness

Function

Basic command	sum varname	
Useful options	sum varname, detail	
Explanations	varname detail	Insert the name of the variable you want to use. Display additional statistics
Short names	sum	summarize
Notes	Stata's calculations are statistics, not parameters (see Section 4.6.2).	
More information	help summarize	

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)

```
sum gpa
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      9,380    3.178614    .6996298          1          5
```

```
sum gpa, detail
```

```
           Grade point average (Age 15, Year 1985)
-------------------------------------------------------------
      Percentiles      Smallest
 1%          1.7              1
 5%            2              1
10%          2.2            1.1       Obs               9,380
25%          2.7            1.1       Sum of Wgt.       9,380

50%          3.1                      Mean           3.178614
                        Largest       Std. Dev.      .6996298
75%          3.7              5
90%          4.1              5       Variance       .4894818
95%          4.3              5       Skewness       .0244443
99%          4.7              5       Kurtosis       2.559851
```

Both tables show that this variable has 9,380 observations (Obs). The mean is 3.18 and the standard deviation (Std. Dev.) is 0.70. The minimum value is 1 (shown by both Min in the upper table and 1% Smallest in the lower table) and the maximum value is 5 (shown by both Max in the upper table and 99% Largest in the lower table). The lower table additionally shows that the median is 3.1 (as indicated by the 50% Percentiles). The Variance is 0.49, Skewness 0.02, and Kurtosis 2.56.

4.6.4 Tabstat

Quick facts	
Number of variables	At least one
Scale of variable(s)	Continuous (ratio/interval) or ordinal

A nice alternative to summarize is tabstat. Per default, you will only get the mean value, but you can additionally order the following descriptive statistics (just to give some examples):

Types of statistic	
Mean	Mean value
Count	Count of nonmissing observations
Sum	Sum
Min	Minimum (smallest) observed value
Max	Maximum (largest) observed value
Range	Max-min
Sd	Standard deviation
Variance	Variance
Skewness	Skewness
Kurtosis	Kurtosis
Median	Median

Function

Basic command	tabstat varname	
Useful options	tabstat varname, stat(x) tabstat varname, stat(x) by(groupvar)	
Explanations	varname	Insert the name of the variable you want to use.
	stat(x)	Replace "x" by specifying the statistics you want to show.
	by(groupvar)	Specify a group variable.
More information	help tabstat	

Practical example 1

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)

You can use tabstat for a single variable and specify as many statistics as you like:

```
tabstat gpa, stat(count mean median sd variance min max)
```

```
    variable |         N      mean       p50        sd  variance       min       max
-------------+--------------------------------------------------------------------
         gpa |      9380  3.178614       3.1  .6996298  .4894818         1         5
----------------------------------------------------------------------------------
```

And here is an example where sex is included as a group variable:

```
tabstat gpa, stat(count mean median sd variance min max) by(sex)
```

```
Summary for variables: gpa
     by categories of: sex (Sex)

   sex |         N      mean       p50        sd  variance       min       max
-------+----------------------------------------------------------------------
   Man |      4752   3.07258         3  .6988425  .4883808         1         5
 Woman |      4628  3.287489       3.2  .6836043  .4673148       1.3         5
-------+----------------------------------------------------------------------
 Total |      9380  3.178614       3.1  .6996298  .4894818         1         5
```

Practical example 2

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test scores (Age 15, Year 1985)

You can use also use tabstat for multiple variables:

```
tabstat gpa cognitive, stat(count mean median sd variance min max)
```

```
   stats |       gpa  cognit~e
---------+--------------------
       N |      9380      8879
    mean |  3.178614  308.4708
     p50 |       3.1       312
      sd |  .6996298  72.18442
variance |  .4894818   5210.59
     min |         1       100
     max |         5       500
------------------------------
```

4.7 Epidemiological measures

Since this guide is focused between the social sciences and medical sciences, it is also important to understand common terminology and measures used in epidemiology. For the language enthusiasts out there, epidemiology is derived from the Greek words *epi* (upon), *demos* (people), and *logos* (study). Together, quite literally, epidemiology is the study of what falls upon the people, or the science of epidemics. As such, epidemiology's origin was to study epidemics of (mostly) infectious diseases. Nowadays, epidemiology is focused on much more than infectious disease. Modern epidemiology is the study of the determinants and distribution of health-related states or events in specific populations. In other words, it is important to understand the time, place, and person(s) when describing the health of a population.

References (this section)

Ahrens. W., & Pigeot, I. (2014). *Handbook of Epidemiology*. 2nd Edition. New York: Springer Science+Business Media.

Bonita. R., Beaglehole, R., & Kjellström, T. (2006). *Basic Epidemiology*. 2nd Edition. World Health Organization.

Center for Disease Control and Prevention (2006). *Principles of epidemiology in public health practice: an introduction to applied epidemiology and biostatistics*. 3rd Edition.

4.7.1 Proportions, rates, and ratios

Ratios, proportions, and rates are all used to measure frequencies of health and disease. All three frequency measures compare one part of the population to either another part of the population, or to the entire population.

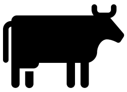
Measure	Definition
Ratio	A comparison of health event numbers or rates between groups
Proportion	A comparison of a part to the whole, or a type of ratio where the value of the numerator is included in the denominator
Rate	A measure of change in one quantity for each unit of another quantity

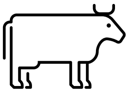
Ratio

A ratio is calculated by dividing, e.g., the number or rate of health events in one group by the number or rate of health events in a second group. Ratio results are often written as the result “to one” or result:1.

Example

We are reviewing results from a study of bovine spongiform encephalopathy (BSE, or Mad Cow Disease). What was the ratio of non-infected versus infected cows?





56	345

The ratio of non-infected to infected cows = 345/56 × 1 = 6.2:1.
In other words, for every infected cow, there are ~6.2 non-infected cows.

Proportion

Calculating a proportion is simply dividing, e.g., the number of persons or health events by the total of persons or health events. In this way, the numerator is a subset of the denominator.

Example

Let us return to our cows and calculate the proportion of infected bovines in the BSE study. There are still 56 infected cows, but now the denominator is the total number of cows in the study (56 infected + 345 non-infected = 401 cows).

56	401

Of the 401 cows in the study, 56 were infected with BSE.
The proportion of infected cows: 56/401 = 0.14 = 14%

Note Proportions are often used as descriptive measures in epidemiology. Specific proportions (e.g., the incidence proportion) will be discussed later in this section.

Rate

In epidemiology, a rate is the frequency with which a health event occurs within a specific population at or during a specific time or time interval. In other words, a rate is a measure of the risk of the health event. Even more specifically, it is the instantaneous risk that the health event will occur at the given time point.

In epidemiology, the change in one quantity for each unit of another quantity is often reported in terms of changes in the quantity of health events for units of time. For an example of a rate calculation, see “incidence rate” in Section 4.7.2.

4.7.2 Morbidity (disease)

Morbidity, which refers to disease, injury, and disability, broadly indicates a departure from the state of wellbeing, whether physiological or psychological. Incidence and prevalence are the key measures of morbidity frequency that aim to quantify population health in terms of health events and health status, respectively.

Measure	Definition
Incidence	The occurrence of a health event within an at-risk population during a specific time interval.
Prevalence	The proportion of health events (new *and* preexisting events) in an at-risk population at a specific timepoint or within a specific time interval.

A population at risk refers to the individuals who are susceptible to the health event. For example, only those with cervixes can logically be at risk for cervical cancer. Therefore, the at-risk population for cervical cancer might consist of women ages 20+ (the age restriction is imposed here since it is extremely rare for women to be diagnosed with cervical cancer before the age of 20).

Incidence refers to the proportion of at-risk individuals who *develop* the health event in question, whereas prevalence refers to the proportion of individuals who, at that point in time, *have* the health event in question. The two concepts are related, but that relationship may look different depending on which health event you are studying. For example, for type II diabetes, there may be low incidence (few people develop type II diabetes) but a high prevalence (many people already have type II diabetes) during the observation period. Conversely, for a disease like seasonal influenza, the incidence may be very high (many people develop the flu), but the prevalence may be low (people do not usually have the flu for very long).

Note Depending on the epidemiologist, incidence may refer to either the number of new events in a specific population or the number of new events per unit of a specific population.

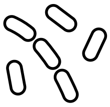
We will now look at measures of incidence and prevalence in a bit more detail. Incidence is often measured using the incidence proportion or the incidence rate.

Incidence proportion

The incidence proportion refers to the risk of occurrence of a health event (e.g. developing a disease, becoming injured or disabled) during a specific time interval among those who are at risk at the beginning of the time interval. Also sometimes referred to as cumulative incidence, the incidence proportion is calculated by dividing the number of new events within the at-risk population (numerator) by the entire at-risk population (denominator).

Example

Let us jump into a time machine and return to the 14th century to examine the incidence proportion of bubonic plague (the black death, caused by *Yersinia Pestis*) in Europe. If the at-risk population in the year 1347 was 70,000,000 and, of those, 860,000 individuals developed bubonic plague during the month under observation, what is the incidence proportion?



860,000	70,000,000

The incidence proportion is the number of at-risk individuals who developed bubonic plague divided by the total population at risk during that specific month: 860,000/70,000,000 = 0.01 = 1.2%

Incidence rate

The incidence rate has the same numerator as the incidence proportion: the number of new events within the at-risk population. However, the denominator is different. For the incidence rate, the denominator is the *time* each at-risk individual was observed, summed for all at-risk individuals. In other words, the denominator is the total time the population was at risk for the health event during the time interval under study. Incidence rates are often used to measure the speed at which a health event is distributed within a population.

Example

In a magical world where all individuals in the study are followed for the same amount of time and none are lost to follow-up, we have a study population of 100 people, each of whom is followed for 10 years. Within this population, 15 are newly diagnosed with heart disease.

15	100

We must first calculate the total time at risk. In this example, the unit of time at risk is measured in person-years. We have 100 people, each followed for 10 years.
100 people followed for 10 years = 1,000 person-years.

Now that we have our denominator, we can calculate the incidence rate.
Number of health events/time at risk = incidence rate.
15/1,000 person-years = 0.015 heart disease events per person-year.

Since health events are often reported per 100,000 person-years, we could also report the incidence rate as 1,500 heart disease events per 100,000 person-years.

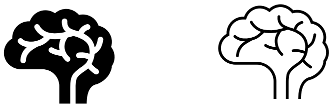
Note Measures of person-time will be discussed in much more detail in Chapter 17.

Prevalence

Prevalence can be calculated at a single point in time (point prevalence) or during a specific time interval (period prevalence). Calculating prevalence can be quite straightforward: we divide the number of people who have the health event (numerator) by the number of people in the at-risk population (denominator) at a specific time or during a specific time interval.

Example

We have an at-risk population of 2,000,000 adults between ages 50 and older who live in Europe. Of those, 18,800 have been diagnosed with dementia.



18,800	2,000,000

The prevalence of dementia within this population at this exact point in time is the number of at-risk people with dementia divided by the total population of persons at risk: 18,800/2,000,000 = 0.009 = 0.9%

Note Though prevalence is a relatively straightforward calculation, it is important to consider the factors that influence prevalence when interpreting your results, or results you read elsewhere. Prevalence may be influenced by, for example, how long the health event lasts (e.g., the duration individuals may live with a non-communicable versus a communicable disease), improved reporting, prolongation of life with the health event, or an increase or decrease in incidence of the health event.

Note Morbidity may also be reported in terms of risk ratios and odds ratios, which are discussed in more detail in Section 4.7.5.

4.7.3 Mortality (death)

Mortality is a valuable measure of population health. Mortality statistics are often used to assess the burden of disease within a certain population, as well as trends and changes in the disease burden over time.

Mortality is commonly reported as a rate, though the standard rates differ in terms of the population under study. Here are some frequently used measures of mortality:

Measure	Definition
Crude mortality rate	The number of deaths during a specific time interval divided by the total number of individuals at risk of dying during that time interval.
Cause-specific mortality rate	The number of deaths attributed to a given cause during a specific time interval divided by the total number of individuals at risk of dying during that time interval.
Age-specific mortality rate	The number of deaths among individuals in a certain age group during a specific time interval divided by the total number of individuals at risk of dying in the age group during that time interval.
Infant mortality rate	The number of deaths among children < 1 year old divided by the total number of live births during that time interval.
Maternal mortality rate	The number of deaths attributed to pregnancy- or childbirth-related causes during a specific time interval divided by the total number of live births during that time interval.

Note Since the denominator in, e.g., the crude mortality rate does not include the time at risk, some epidemiologists would argue that it is not a true rate. It should also be noted that the denominator in mortality rates based on vital statistics, such as the number of death certificates, often reflects the size of the population as of the *middle* of the time interval, rather than at the beginning of the time interval (i.e. the incidence proportion).

4.7.4 Natality (birth)

Measures of natality refer to population-based measures of birth. Commonly used natality measures are also reported as rates:

Measure	Definition
Crude birth rate	The number of live births during a specific time interval divided by the total population at the midpoint of that time interval.
General fertility rate	The number of live births during a specific time interval divided by the total number of women ages 15-44 at the midpoint of that time interval.

4.7.5 Risks and odds

Comparisons between what is *observed* and what is *expected* are very important when describing and analyzing epidemiologic data. Risks and odds are two measures of association that quantify the risk of occurrence of a health event. Measures of association in epidemiology are often reported as ratios so that we can compare the risk or odds between two groups. Two such ratios are discussed in more detail below.

Risk ratio

A risk ratio, or the relative risk, compares the risk of occurrence of a health event during a specific time interval for one group with the risk of occurrence of a health even during that same time interval for a second group. This should sound familiar because a risk ratio is calculated by dividing the *incidence proportion* for one group by the *incidence proportion* for the second group!

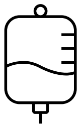
If we can calculate an incidence proportion, why do we need the risk ratio? For example, maybe we think the incidence of bubonic plague in neighborhood A seems quite high. Is this observed incidence proportion higher than what we expect, or higher than it is in other neighborhoods? We could use a risk ratio to compare the observed proportion in neighborhood A with neighborhood B, which represents the expected level of plague.

The two groups in a risk ratio are commonly differentiated in epidemiology according to a specific risk factor (i.e., whether they were exposed or unexposed); however, the two groups may also differ according to a demographic factor (e.g., born male versus female).

To calculate a risk ratio to compare the risk of occurrence of a health event between two groups, we need a two-by-two table.

Example

Let us revisit the European past. This time, we will go to 19th century London to apply a classic example to calculate a risk ratio. In a cholera outbreak in London in 1853, 28,200 of the 168,000 individuals served by the Southwark water supply company developed cholera, compared to 600 of the 19,200 individuals served by the Lambeth water supply company who developed cholera. To compare the risk between these two groups, we will summarize them in a two-by-two table, with two rows for the exposure (water supply company) and two columns for the outcome (cholera incidence).



	Cholera	No Cholera	Total
Southwark	28,200	139,800	168,000
Lambeth	600	18,600	19,200
Total	28,800	158,400	187,200

First, we will calculate the proportion of cholera incidence for each of the water companies:
Southwark: 28,200/139,800 = 0.202 = 20.2%
Lambeth: 600/18,600 = 0.032 = 3.2%

Then, we calculate the ratio of these two proportions:
Risk ratio = 20.2/3.2 = 6.3

From these data, we could conclude that individuals served by the Southwark water supply company were 6.3 times more likely to develop cholera than those served by the Lambeth water supply company.

Note A risk ratio > 1.0 indicates that the exposed group (numerator) has an increased risk for the health outcome, as in the above example. A risk ratio < 1.0 indicates that the exposed group has a decreased risk for the health outcome. In other words, the exposure might be protective against the occurrence of the health outcome. A risk ratio = 1.0 indicates that the risk for the health outcome is the same for both groups.

Odds ratio

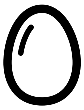
An odds ratio also quantifies the risk of occurrence of a health event, but it compares two categories: those who were exposed who have or do not have the health outcome, and those who were unexposed who have or do not have the health outcome.

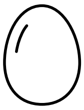
An odds ratio is also calculated using a two-by-two table, but instead of comparing across rows, we compare the numerator in one column to the denominator in the other. That is why the odds ratio is sometimes called the cross-product ratio: the calculation creates an X in the two-by-two table. Let's try an example.

Example

The Swedish public health agency is investigating an outbreak of food-borne infections potentially caused by *Salmonella* bacteria. They trace the outbreak to eggs served in a popular café in Södermalm. The public health agency obtains data for everyone who ate in the café during the weeks immediately before and after the oubreak began, including whether they consumed food made with eggs (exposure) or presented with a *Salmonella* infection (outcome).

Our two-by-two table is the same as the one we used for the risk ratio. Here, we have added letters to help you visualize the cross-product calculations.





	Salmonella	**No *Salmonella***	**Total**
Exposed	a = 132	b = 1,900	2,032
Unexposed	c = 10	d = 2,100	2,110
Total	142	4,000	4,142

To calculate the odds ratio, we first multiply a x d and b x c to get the cross products, and then divide the product of a x d by the product of b x c:
(132 x 2,100)/(1,900 x 10) = 14.6

Those who ate food containing eggs at this particular café during this two-week period were 14.6 times more likely to have contracted a *Salmonella* infection compared to those who did not eat eggs.

Note If the health outcome in question is rare, the odds ratio will be quite similar to the risk ratio. This is especially convenient for case-control studies, where the odds of exposure among cases are compared to the odds of exposure among controls. For case-control studies, we often do not know the size of the population from which the cases were drawn. This means that risks and rates (and therefore risk and rate ratios) cannot reliably be calculated, and so we instead calculate odds ratios.

4.7.6 Attributable proportion

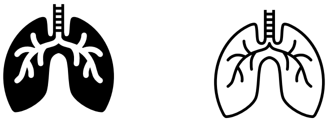
As a conclusion to this section on epidemiological measures of association and frequency, we want to briefly introduce one measure of public health impact, which can provide important context about the burden of different health outcomes on population health. Attributable proportion measures the quantity of the health outcome in the exposed group that can be attributed to the exposure. In other words, the attributable proportion represents the proportion of health events that hypothetically would be reduced if the exposure did not exist or could be removed from the equation.

Attributable proportion has a key assumption: if the number of health events in the unexposed group is the expected risk (baseline) for that event, then the difference in risk between the exposed and unexposed groups can be attributed, or caused by, the exposure. This also means that the attributable proportion should only be calculated for one exposure, or risk factor, that causes the health outcome.

The attributable proportion is calculated by subtracting the risk for the unexposed group from the risk for the exposed group, dividing the difference by the risk for the unexposed group, and then multiplying the quotient by 100.

Example

A classic example of attributable proportion is the relationship between smoking and mortality attributable to lung cancer. Our study population of male, British doctors in the 20th century contains daily smokers and non-smokers. If the mortality rate for lung cancer among daily smokers is 0.56 deaths per 1,000 persons per year, and 0.06 deaths per 1,000 persons per year, what is the attributable proportion?



0.56	0.06

The attributable proportion is: (risk for the exposed group – risk for the unexposed group)/risk for the exposed group, x 100.
Attributable proportion: (0.56 – 0.06)/0.56 x 100 = 89.3%

If the assumptions for calculating the attributable proportion hold, and assuming the two groups of doctors are comparable, these results would indicate that about 89% of deaths due to lung cancer within this group could be attributed to daily smoking. This would mean that if smoking did not exist, or we could remove smoking as an exposure, the mortality rate attributable to lung cancer would hypothetically decrease by 89.3%. However, the remaining 10.7% of deaths from lung cancer in this group would still occur.

4.8 Designing descriptive tables and graphs

What kind of descriptive statistics should one include in a study? This is a question with an unlimited number of answers. Below, we have some recommendations that you can draw inspiration from.

Note Make sure that you have defined your analytical sample before summarising the descriptive statistics in a table or inserting a graph in the document (see Section 11.5).

Note Of course, we also produce tables and graphs to present the results from our statistical analysis – but this specific section focuses on descriptive statistics.

4.8.1 Tables

Usually, when one writes up a manuscript for a study, there is at least one table with descriptive statistics. This table normally includes all the study variables. It is very common that the variables have different measurement scales but they can still be included in the same table.

Presentation of descriptive statistics in a table	
Continuous variables	Mean, median, standard deviation, min, max
Categorical variables	Percentage distribution

Checklist for tables
The tables are numbered sequentially throughout the document.
There is a descriptive heading placed *above* the table.
The number of observations (e.g. individuals) is included in the heading.
The table does not include any vertical lines/borders
The table includes as few horizontal lines/borders as possible.

Some examples

Below is a simple example of a descriptive table with only categorical variables.

Table 1. Descriptive statistics for the study variables (n=5,000).

	n	%
Sex		
Males	2,543	50.9
Females	2,457	49.1
Quality of life		
Low	570	11.4
Medium	1198	24.0
High	3232	64.6
Smoking		
Never smoker	2961	59.2
Former smoker	1433	28.7
Current smoker	606	12.1

Below is a simple example of a descriptive table with categorical and continuous variables.

Table 1. Descriptive statistics for the study variables (n=5,000).

		n		%
Drinks alcohol				
No		1631		32.6
Yes		3369		67.4
Uses illicit drugs				
No		4099		82.0
Yes		901		8.0
	Mean	Median	Std. dev.	Min, Max
Age	16.8	17.2	1.2	12, 19
Average school marks	14.9	14.3	2.5	10.1, 20.0

4.8.2 Figures

In some instances, there might be relevant to also produce a figure for one or more of the variables. This is perhaps particularly the case if one wants to illustrate the distribution of a variable in a more detailed way, or if one wishes to make a simple comparison between groups.

Checklist for figures
The figures are numbered sequentially throughout the document.
There is a descriptive heading placed *below* the figure.
The number of observations (e.g. individuals) is included in the heading.
The figure can be printed in black and white without becoming less informative.

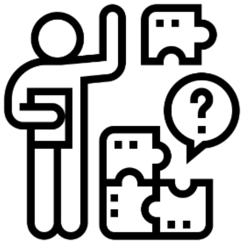
5. STATISTICAL SIGNIFICANCE

Outline

Content

This chapter focuses on theoretical issues concerning statistical significance, including a discussion on p-values and confidence intervals. We also present some practical alternatives for calculating confidence intervals for descriptive statistics.

5.1 Hypothesis testing



A lot of quantitative research is about examining relationships between variables (see Chapter 9 for a more detailed discussion about those issues). Assuming that all is done correctly, data analysis will give us information about the association (i.e. the change in the outcome per unit increase in the exposure) and the direction of the relationship (i.e. whether the relationship is negative or positive). These are the two most important outcomes of data analysis, but it is not uncommon that research inquiry instead focuses on a third point: statistical significance. Statistical significance can be seen as an indicator of the reliability of the results – although that is important, it is not what exclusively should guide which findings we focus on and which we discard. A fourth issue that needs to be considered is whether the findings have any practical or clinical importance – in order words; do they matter? We therefore suggest the following priority list when it comes to how results from data analysis should be interpreted and valued:

Priority list	
1. Effect	How much does the outcome change per unit increase in the exposure?
2. Direction	Is the relationship positive or negative?
3. Statistical significance	Is the relationship reliable?
4. Practical importance	Is the relationship relevant?

5.1.1 Hypotheses

Let us return to the matter of statistical significance: what is it really? Well, for example, if we find that cats are smarter than dogs, we want to know whether this difference is “real”. Hypothesis testing is how we may answer that question. We start by converting the question into two hypotheses:

Hypotheses		
Null hypothesis	(H_0)	There is no difference
Alternative hypothesis	(H_1)	There is a difference

There is no law saying that the null hypothesis is always “no difference” and the alternative hypothesis is always “difference”. However, for the null hypothesis, precedence is commonly given to the “simpler” (or more “conservative” or “normative”) hypothesis. Here, it is generally simpler to claim that there is no difference in intelligence between cats and dogs than to say that there is a difference.

5.1.2 Outcomes

There are two possible outcomes of hypothesis testing:

Outcomes of hypothesis testing	
Reject H_0 in favour of H_1	Suggests that the alternative hypothesis *may* be true (but it does not prove it)
Do not reject H_0	Suggests that there is not sufficient evidence against H_0 in favour of H_1 (but it does not prove that the null hypothesis is true)

Note We are never able to decide from hypothesis testing that we should reject or accept H_1. However, rejecting H_0 may lead us to suggest that H_1 might be accepted.

5.1.3 Errors

There are two types of error that may occur in hypothesis testing: a type I error occurs when the null hypothesis is rejected despite being true, whereas a type II error occurs when the null hypothesis is not rejected despite being false. In the example of cats and dogs, a type I error would thus occur if we concluded that there is a difference in the intelligence between cats and dogs although that is not true. A type II error, on the other hand, would occur if we concluded that there is no difference in intelligence when in fact there is.

Type I and type II errors			
		Conclusion	
		Reject H_0 in favour of H_1	**Do not reject H_0**
"Truth"	**H_0**	*Type 1 error*	*Right decision*
	H_1	*Right decision*	*Type II error*

Type I errors are generally considered to be more serious that type II errors. Type II errors are often due to poor statistical power (often because of small sample size).

5.1.4 Statistical hypothesis testing

Conducting a statistical hypothesis test is easy to do in statistical software such as Stata. These tests give us a probability value (p-value) that can help us decide whether or not the null hypothesis should be rejected. See Section 5.2 for a further discussion about the p-value.

5.2 P-values

The probability value – or p-value – helps us decide whether or not the null hypothesis should be rejected. There are some common misunderstandings about p-values:

The p-value is *not*…
… the probability that the null hypothesis is true
… the probability that the alternative hypothesis is false
… the probability of the occurrence of a type I error (falsely rejecting H_0)
... the probability that replicating the experiment would yield the same conclusion
… the probability that the finding is a "fluke"
… an indicator of the size of the effect or importance of the findings
… determining the significance level

Using the p-value to make this decision, it must first be decided what probability value we find acceptable. This is often referred to "the significance level". If the p-value is below this level, it means that we can reject the null hypothesis in favour of the alternative hypothesis, and if the p-value is above this level, it means that we cannot reject the null hypothesis. The smaller the p-value, the more convincing is the rejection of the null hypothesis.

5.2.1 Significance levels and confidence levels

Significance levels and confidence levels are just two ways of looking at the same thing. The level is set by the individual researcher – it that sense, it is quite arbitrary – but there are some levels that are widely used (asterisks are often used to illustrate these levels):

P-value	**Significance level**	**Confidence level**	
p<0.05	5%	95%	*
p<0.01	1%	99%	**
p<0.001	0.1%	99.9%	***

Note In some fields of research, p<0.10 – statistical significance at the 10% level – is also a commonly used significance level.

Let us return to the example of differences in intelligence between cats and dogs. For instance, if we find a difference in intelligence between these types of animal, and the p-value is below 0.05, we may thus state that the null hypothesis (i.e. no difference) is rejected at the 95% confidence level. The p-value does not, however, state whether the difference is small or big, or whether cats or dogs represent the smarter type of animal (in order to state such things, one would have to look at the direction and the effect size).

It should be noted that the p-value is affected by the sample size, which means that a smaller sample size often translates to a larger p-value. For example, if you have a data material of 100 individuals, the effect size has to be quite large (e.g. large income differences income between men and women) in order to get small p-values. Conversely, larger sample size makes it easier to find small p-values. For example, if you analyse a data material containing the entire population of a country, even tiny differences are likely to have small p-values. In other words, the size of the sample influences the chances of rejecting the null hypothesis (see Section 5.6 about power analysis).

5.2.2 Practical importance

As stated earlier in this section, statistical significance – determined by the p-value – is *not* the same as effect size or practical/clinical importance (i.e. whether it "matters"). We can use couple of examples to illustrate the differences:

Example 1
A pharmaceutical company has developed a drug to cure obesity. During tests of this drug, it appears as migraine could be one of the side effects of taking this drug. The null hypothesis would here be that there are no differences in the risk of migraine between people who had taken the drug and those who have not. The alternative hypothesis would then be that there are differences. When we run the analysis on this data material, we see that those who have taken the drug have ten times the risk of migraine, but the p-value is above the 95% confidence level (i.e. $p>0.05$). Thus, we cannot reject the null hypothesis. The difference is however large and is likely to have significant impact on people's lives. It could moreover be the case that a type II error has occurred here due to a small sample size.

Example 2
In the second example, researchers have gathered data on coffee consumption and happiness among 100,000 company employees. The null hypothesis would here be that there are no differences in happiness between people who drink coffee and those who do not. The alternative hypothesis would be that there are differences. The analysis suggests that there is a tiny difference in happiness between those who drink coffee and those who do not, to the advantage of the coffee drinkers. The p-value is below 0.05 which suggests that the null hypothesis can be rejected at the 95% confidence level. However, the difference is very small and the results may not be very useful.

5.3 Confidence intervals

Confidence intervals (CI) are closely related to the concept of statistical hypothesis testing, but they are more informative than p-values since they do not only suggest whether we should reject H_0 or not, they also provide the range of plausible values.

5.3.1 The "unknown population parameter"

Before we get into the discussion about confidence intervals, we need to address the concept of "unknown population parameter". A parameter tells us something about a population (while a "statistic" tells us something about a sample). The population parameter is thus basically a measure of any given population. Examples of population parameters are: the mean height of Swedish men, the average intelligence score in 12-year olds, or the mean number of children among highly educated people. The parameter is a fixed value, i.e. it does not vary. We seldom have information about the entire population, generally only for a part of it (a sample). In that case, the population parameter is unknown. Simply put, a confidence interval is a range that includes the "unknown population parameter".

5.3.2 Limits and levels

The interval has an upper and a lower bound (i.e. confidence limits). Similar to p-values, confidence intervals have "confidence levels" that indicate how certain we can be that the interval includes the true population parameter. Confidence intervals are typically stated at the 95% level. A 95% confidence level would thus mean that if we replicated a certain analysis in 100 samples of the population, we would expect that 95% of the intervals would include the true population parameter. Thus, strictly speaking, it is *not* correct to say that "with 95% probability, the true population parameter lies within this interval" (because the parameter either is or is not within the interval).

5.3.3 Confidence and precision

When discussing confidence intervals, it is important to be aware of the tension between precision and certainty: better precision means being less confident, whereas more confidence means less precision. As previously stated, confidence is reflected by the confidence level we choose; logically, a higher confidence level means more confidence. The higher the confidence level we choose, the wider the interval gets – and the wider the interval is, the less the precision we get.

Confidence versus precision
Higher confidence level = wider confidence interval = less precision
Lower confidence level = slimmer confidence interval = more precision

However, it is important to know that the width of the confidence interval is also affected by the sample size: the larger the sample size, the slimmer the interval (which means better precision).

Let us take an example to sum up what has been said about confidence intervals so far: We have gathered data on all sociology students at Stockholm University and find that their mean age is 26 years. Instead of highlighting this relatively non-informative fact, we can calculate the confidence interval (at the 95% level). In this case, it is 22-30. Therefore, we could make the more informative statement that: “with 95% confidence, we conclude that the mean age of sociology students is 22 to 30 years”.

The most common application for confidence intervals as a way of significance testing is when we are interested in the difference between two samples. For example: the difference in the mean income between men and women, or the difference in the percentage of daily smokers among individuals with a lower level of education versus those with a higher level of education. In this case, we may look at the overlap between the confidence intervals estimated for each sample. Suppose that we have an upcoming election and just got the results from the latest poll. There are two parties in the race: the green party and the yellow party. The results from the poll show that the green party got 42% of the votes and the confidence interval is 40-44 (at the 95 % level). The yellow party got 58% of the votes and the confidence interval is 54-62 (at the 95% level). What does this tell us? First of all, we can conclude that the yellow party has a greater share of votes. Looking at the two confidence intervals, we see that the intervals do not overlap. Why is that important? Well, remember that *all* values in a confidence interval are plausible. Hence, if the confidence intervals do not overlap, it means that the estimates (in this case: the share of votes) are indeed different given the chosen confidence level (in this case: at the 95% level). However, it should be emphasized that while non-overlap always mirrors a significant difference, overlap is not always the same as a non-significant difference.

5.4 Choice between p-values and confidence intervals

Now you are maybe wondering; should you use p-values or confidence intervals? Almost all disciplines would recommend using both because they capture several different dimensions. In the following, the advantages and disadvantages of p-values and confidence intervals will be described and discussed.

P-value is an important part of research, most likely the heart of it. The p-value is based on “yes-or no”-questions in which it shows how much evidence we have against the null hypothesis. P-values are much clearer than confidence intervals and it helps the researcher to make quick judgments about his research. Another advantage with the p-value is that it can give the difference from a previous specified statistical level. Unfortunately, there are misconceptions about the p-value among researchers and many disciplines rely on them to draw conclusions rather than understanding the background. One of the common mistakes among researchers is that they do not further analyse their data in order to ensure that the p-value is not affected by other factors. Moreover, p-values cannot alone permit any direct statements about the direction or size of difference. In order to make those decisions, one must always look at the confidence intervals.

A confidence interval informs the researcher about the power of the study and whether the data is compatible, it also shows the likelihood of the null hypothesis being true and that in turn tells us how much confidence we have in our findings. The width of the confidence interval indicates the precision of the point estimates, in which a small interval indicates a more precise estimate, while a wide interval indicates a less precise estimate. The precision is related to the sample size and power in which it tells us that the larger sample size we have, the greater, the more precise estimates we have. The intervals are useful when having small sample sizes. Normally, small studies fail to find statistically significant treatments, when including point estimates with wide intervals that include the null value may be consistent and significant. The intervals provide the researcher an understanding of the sample size. This can also be a disadvantage when having large data because it produces statistically significant results even if the difference between the groups is small. Another advantage with the confidence interval is that it can provide means of analysis for studies that seek to describe and explain, rather than make decisions about treatments effects. A disadvantage with the confidence interval is that it captures several elements at the time, in which it may not give precise information like the p-values.

As mentioned, a majority of disciplines recommend including both p-values and confidence intervals because they capture information in different dimensions. Neither p-values nor confidence intervals can prevent biases or other problems but the combination of them provides a more flexible approach and highlights new perspectives on the data. Confidence intervals permit us to draw several conclusions at the same time and they are more informative about sample sizes and point estimates.

They are also useful in studies when we have small sample sizes. But they are not as precise as p-values when it comes to accepting and rejecting the null hypothesis. Thus, when we combine them together we can be more certain.

The figure below shows the advantages and disadvantages when interpreting and drawing conclusions with the help of p-values and confidence intervals.

P-values versus confidence intervals		
	P-values	Confidence intervals
Accept/reject	👍	👍
Degree of support	👍	👍
Estimate and uncertainty	👎	👍

5.5 Calculate confidence intervals for descriptive statistics

Now that we have discussed what confidence intervals are (and what they are not), we thought it would be good time to show how to calculate them for descriptive statistics. For this purpose, we can use the commands ci, centile, and proportion.

5.5.1 Confidence intervals for means

Can be used for continuous variables with a normal distribution.

Function

Basic command	ci means varlist	
Useful options	ci means varlist, level(#)	
Explanations	Varlist	Insert the name(s) of the variable(s) that you want to use.
	level(#)	Specify the confidence level. Default is 95.
More information	help ci	

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)

```
ci means gpa
```

```
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         gpa |      9,380    3.178614    .0072238        3.164454    3.192774
```

In this example, we can see that the mean value for gpa is 3.18. The 95% confidence interval is 3.16-3.19.

5.5.2 Confidence intervals for median

Can be used for continuous variables (with a normal or skewed distribution).

Function

Basic command	centile varlist	
Useful options	centile varlist, level(#)	
Explanations	varlist	Insert the name(s) of the variable(s) that you want to use.
	level(#)	Specify the confidence level. Default is 95.
More information	help centile	

Practical example

Dataset: StataData1.dta

Name	**Label**
cognitive	Cognitive test score (Age 15, 1985).

```
centile cognitive
```

```
                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
   cognitive |     8,879         50         312             312         316
```

In this example, we can see that the median cognitive test score is 312, and the 95% confidence interval is 312-316.

5.5.3 Confidence intervals for variances and standard deviations

Can be used for variables that are continuous.

Function

Basic command	ci variances varlist	
Useful options	ci variances varlist, level(#) ci variances varlist, sd level(#)	
Explanations	varlist	Insert the name(s) of the variable(s) that you want to use.
	sd	Option to display confidence interval for standard deviation.
	level(#)	Specify the confidence level. Default is 95.
More information	help ci	

Practical example

Dataset: StataData1.dta

Name	**Label**
cognitive	Cognitive test score (Age 15, 1985).

```
ci variances cognitive
```

```
    Variable |        Obs       Variance        [95% Conf. Interval]
-------------+------------------------------------------------------
   cognitive |      8,879        5210.59        5060.642     5367.335
```

Here, the variance (5210) and its confidence interval (5061-5367) is shown.

```
ci variances cognitive, sd
```

```
    Variable |        Obs       Std. Dev.       [95% Conf. Interval]
-------------+------------------------------------------------------
   cognitive |      8,879       72.18442        71.13819      73.2621
```

This shows the standard deviation (72.18) and its 95% confidence interval (71.14-73.26).

5.5.4 Confidence intervals for counts

Can be used for continuous variables that are counts.

Function

Basic command	ci means varlist, poisson	
Useful options	ci means varlist, poisson level(#)	
Explanations	varlist	Insert the name(s) of the variable(s) that you want to use.
	level(#)	Specify the confidence level. Default is 95.
More information	help ci	

Practical example

Dataset: StataData1.dta

Name	**Label**
unemp_42	Days in unemployment (Age 42, Year 2012)

```
ci means unemp_42, poisson
```

```
                                                         -- Poisson  Exact --
    Variable |   Exposure        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    unemp_42 |       9078    17.52787     .043941        17.44185    17.61421
```

In this example, the mean is 17.53 days in unemployment. The 95% confidence interval is 17.45-17.61 days.

5.5.5 Confidence intervals for proportions

Can be used for categorical variables.

Function

Basic command	proportion varlist	
Useful options	proportion varlist, level(#)	
Explanations	varlist	Insert the name(s) of the variable(s) that you want to use.
	level(#)	Specify the confidence level. Default is 95.
More information	help proportion	

Practical example

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)

```
proportion educ
```

```
Proportion estimation                 Number of obs   =      9,183

------------------------------------------------------------------
                 |                                   Logit
                 | Proportion   Std. Err.     [95% Conf. Interval]
-----------------+------------------------------------------------
            educ |
      Compulsory |   .1919852   .0041101      .1840571     .200171
 Upper secondary |   .4423391   .0051829       .432205    .4525214
      University |   .3656757   .0050259      .3558812    .3755826
------------------------------------------------------------------
```

Here, we get the proportions (which can be translated into percentages) and its confidence interval for the three categories of educ. In this example, 19.2% have compulsory education (95% CI: 18.4-20.0), 44.2% have upper secondary education (95% CI: 43.2-45.3%), and 36.6% have university education (95% CI: 35.6-37.6%).

5.6 Power analysis

Although one should perform a power analysis when planning a study, many want to use it as a sort of post-hoc test. Study design is not covered in this guide, and performing post-hoc power analysis is not really something that we encourage. Accordingly, this guide will not elaborate on power analysis in more detail. If you nevertheless want to try it out on your own, there is a whole module for power calculations in Stata.

More information	help power

6. COMPARE GROUPS

Outline

Content

In this chapter, we focus on different ways of comparing groups (and measurement points/samples). It starts with some descriptive statistics (box plots and crosstables), and continues with how to perform t-tests and one-way ANOVA (including their non-parametric alternatives), as well as chi-square tests.

Before you start comparing groups (or measurement points/samples), it is important to know about the variables' measurement scale and distribution (as discussed in Chapter 3). For an overview, see the next page.

X (Independent variable/exposure/group variable)		**Y** (Dependent variable/ outcome/test variable)		**CHOICE OF TEST**	**DESCRIPTION**
Categorical	+	**Categorical**	⇨	**Chi2**	**Crosstable**
Categorical 2 categories (groups)	+	**Continuous** (normal distribution)	⇨	**T-test: Independent samples**	**E.g. Box plot**
Categorical 2 categories (groups)	+	**Continuous** (skewed distribution)	⇨	**Mann-Whitney**	**E.g. Box plot**
Categorical More than 2 categories (groups)	+	**Continuous** (normal distribution)	⇨	**Oneway ANOVA**	**E.g. Box plot**
Categorical More than 2 categories (groups)	+	**Continuous** (skewed distribution)	⇨	**Kruskal-Wallis**	**E.g. Box plot**

Measurement point 1		**Measurement point 2**		**CHOICE OF TEST**	**DESCRIPTION**
Continuous (normal distribution)	+	**Continuous** (normal distribution)	⇨	**T-test:Paired samples**	**E.g. scatterplot**
Continuous (skewed distribution)	+	**Kontinuerlig** (skewed distribution)	⇨	**Wilcoxon**	**E.g. scatterplot**

6.1 Descriptives

There are many ways that we can use descriptive statistics to compare two or more groups (or samples). Here, we will focus on box plots and crosstables.

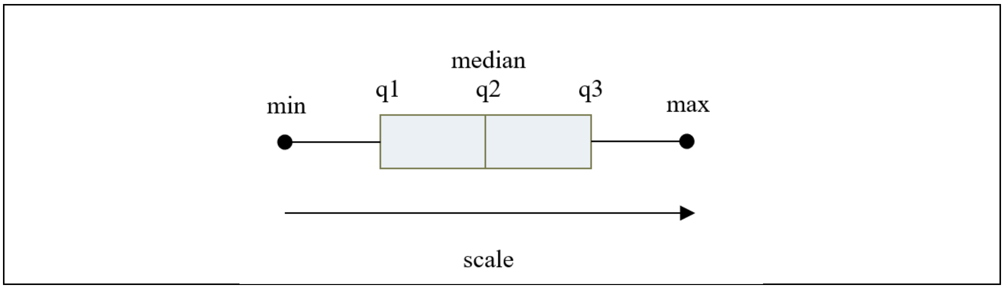
To use other types of graphs, try them out in combination with by (see Section 2.8).

6.1.1 Box plot

Quick facts	
Number of variables	One group variable (optional) One test variable
Scale of variable(s)	Group variable: categorical (nominal/ordinal) Test variable: continuous (ratio/interval)

A box plot – or box and whisker plot – is a four-part summary of a variable. The four parts are made up by five components: minimum, first quartile, median, third quartile, and maximum. Below is a simple illustration: we draw a box from the first quartile (q1) to the third quartile (q3). The line in the middle of the box represents the median (q2). The whiskers represent the minimum (min) and maximum (max) values. This means that each of the four parts contain approximately 25% of the values.

It is not necessary to include a group variable in a box plot, but we chose to place box plots in this chapter instead of Chapter 4, since we think that it is a nice alternative for comparing groups in a descriptive way.



Box plots are sensitive to outliers, so if you discover that your variable has any extreme values, you might need to reconsider your box plot (e.g. by excluding the outliers).

Function

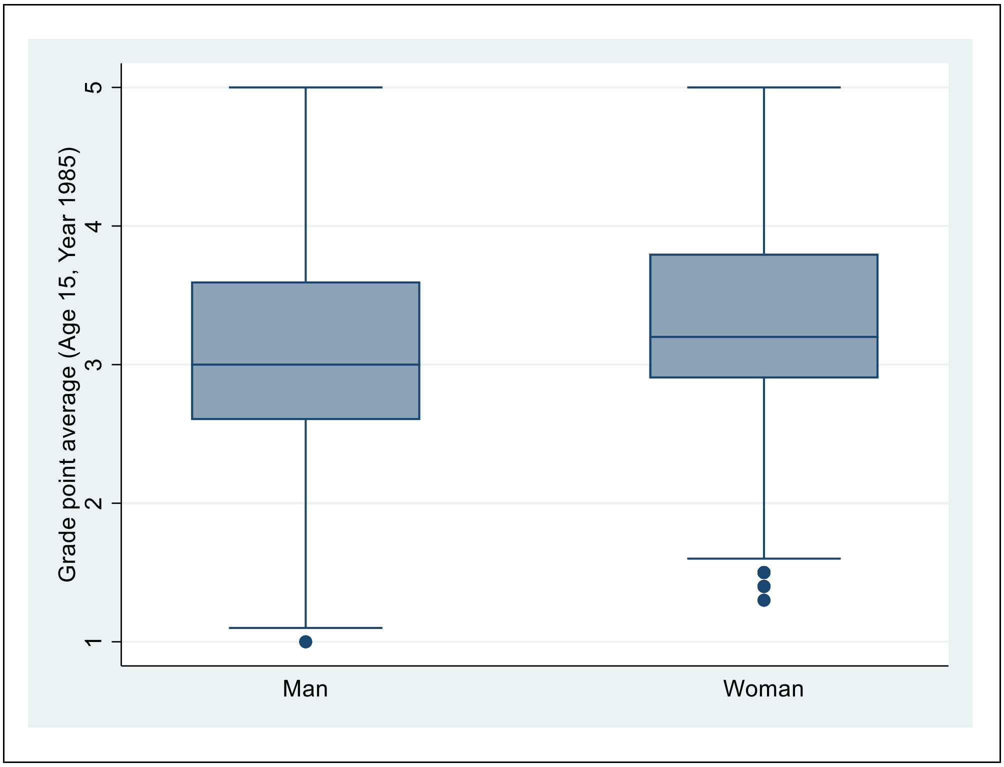
Basic command	graph box yvar, over(groupvar)	
Explanations	yvar	Insert the name of the variable that you want to use as your y-variable.
	groupvar	Insert the variable defining the groups.
More information	help graph box	

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
sex	Sex

```
graph box gpa, over(sex)
```



The box plot above shows the distribution of gpa according to sex. We can see that the distribution is slightly shifted upwards among women compared to men: their median grade point average is higher. There are some outliers, but this does not seem to be a big problem (the dots are few).

6.1.2 Crosstable

Quick facts	
Number of variables	Two
Scale of variable(s)	Categorical (nominal/ordinal)

A crosstable is a description of how individuals are distributed according to two variables.

This function is used primarily for categorical variables (i.e. nominal/ordinal) but can be used for any type of variable; the main concern is that the table becomes too complex and difficult to interpret if there are many categories/values in the variables used. Moreover, it is possible to add a chi-square to the crosstable (for more information about chi-square, see Section 6.5).

Unless otherwise specified, a crosstable will only show the frequency distribution. This is usually not what we are after; we rather would like to see the percentage distribution. There are two options to choose from: column and row percentages. The frequencies (i.e. the number of individuals) in the cells are the same, but the percentages are different since the focus shifts between the tables. If you find this difficult to separate in your mind, one good advice is perhaps to see where the percentages add up to 100% in Total - in the rows or in the columns.

Note If we would have individuals with missing information with regard to any of the two variables, these would be excluded from the crosstable unless otherwise specified.

Function

Basic command	tab varname1 varname2	
Useful options	tab varname1 varname2, row tab varname1 varname2, col	
Explanations	varname1	Insert the name of the first variable you want to use (is included as the row variable).
	varname2	Insert the name of the first variable you want to use (is included as the column variable).
	row	Show row percentages.
	col	Show column percentages.
	m	Include missing
Short names	tab	tabulate
	col	column
	m	missing
Notes	Options can be used simultaneously, e.g: tab varname1 varname2, row col m	
More information	help tabulate twoway	

Practical example

Dataset: StataData1.dta

Name	**Label**
sex	Sex
bullied	Exposure to bullying (Age 15, Year 1985)

```
tab sex bullied
```

```
           |  Exposed to bullying
           |  (Age 15, Year 1985)
       Sex |        No        Yes |     Total
-----------+----------------------+----------
       Man |     3,799        346 |     4,145
     Woman |     3,981        593 |     4,574
-----------+----------------------+----------
     Total |     7,780        939 |     8,719
```

In the table above, we specified sex as the row variable, and bullied as the column variable. Frequencies are shown in the different cells. We can observe that:

- There are 4,145 men and 4,574 women. There are 7,780 individuals who have not been exposed to bullying and 939 who have.
- Focusing on the frequency distribution of bullying across gender:
  Among the men, there are 3,799 who have not been exposed to bullying and 346 who have. The corresponding numbers among women are 3,981 and 593, respectively.
- Focusing on the frequency distribution of gender across bullying:
  Among those who have not been exposed to bullying, there are 3,799 men and 3,981 women. Among those who have been exposed to bullying, there are 346 men and 593 women.

Note that the two last bullet points use the same numbers but refer to them in different ways. This is important to keep in mind for the interpretation of the percentage distributions presented below.

Comparing frequencies are, however, rather tricky since the sample size differs across the categories of the variables. That is why it is often more practical to include percentages as well.

```
tab sex bullied, row
```

```
           |  Exposed to bullying
           |  (Age 15, Year 1985)
       Sex |        No        Yes |     Total
-----------+----------------------+----------
       Man |     3,799        346 |     4,145
           |     91.65       8.35 |    100.00
-----------+----------------------+----------
     Woman |     3,981        593 |     4,574
           |     87.04      12.96 |    100.00
-----------+----------------------+----------
     Total |     7,780        939 |     8,719
           |     89.23      10.77 |    100.00
```

In the table above, we have added row percentages. Since sex is our row variable, we will here see the percentage distribution of bullying across sex.

- In total, 89% have not been exposed to bullying whereas 11% have.
- Among the men, 92% have not been exposed to bullying whereas 8% have. The corresponding figures among women are 87% and 13%, respectively.

```
tab sex bullied, col
```

```
           |  Exposed to bullying
           |  (Age 15, Year 1985)
       Sex |        No        Yes |     Total
-----------+----------------------+----------
       Man |     3,799        346 |     4,145
           |     48.83      36.85 |     47.54
-----------+----------------------+----------
     Woman |     3,981        593 |     4,574
           |     51.17      63.15 |     52.46
-----------+----------------------+----------
     Total |     7,780        939 |     8,719
           |    100.00     100.00 |    100.00
```

In the table above, we have added column percentages. Since bullied is our column variable, we will here see the percentage distribution of sex across exposure to bullying.

- In total, 48% are men and 53% are women.
- Among those who have not been exposed to bulling, 49% are men and 51% are women. The corresponding figures among those who have been exposed to bullying are 37% and 63%, respectively.

Note We have rounded the percentages (it is seldom necessary to report decimals for percentages).

6.2 T-test: Independent samples

Quick facts	
Number of variables	One group variable One test variable (y)
Scale of variable(s)	Group variable: categorical with two values (binary) Test variable: continuous (ratio/interval)

The independent samples t-test is a parametric method for comparing the mean of one variable between two (unrelated) groups. For example, you may want to see if the income salary of teachers differs between men and women, or if the score of a cognitive test differs between children who have parents with low versus high education.

Mean income salary among men	Mean income salary among women

Assumptions

First, you have to check your data to see that the assumptions behind the independent samples t-test hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Continuous test variable	Your test variable should be continuous (i.e. interval/ratio). For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: "How satisfied with your income are you?" (on a scale 1-10) or "To what extent do you agree with the previous statement?" (on a scale 1-5).
Two unrelated categories in the group variable	Your group variable should be categorical and consist of only two groups. Unrelated means that the two groups should be mutually excluded: no individual can be in both groups. For example: men vs. women, employed vs. unemployed, low-income earner vs. high-income earner, and so on.
No outliers	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.

Function

Basic command	ttest testvar, by(groupvar)	
Explanations	testvar	Insert the name of the variable that you want to test.
	groupvar	Insert the variable defining the two groups.
More information	help ttest	

Practical example

Dataset: StataData1.dta

Name	**Label**
cognitive	Cognitive test score (Age 15, Year 1985)
sex	Sex

```
ttest cognitive, by(sex)
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Man |   4,495     311.943    1.091507    73.17985    309.8032    314.0829
   Woman |   4,384    304.9106    1.072037    70.98148    302.8088    307.0123
---------+--------------------------------------------------------------------
combined |   8,879    308.4708    .7660576    72.18442    306.9691    309.9724
---------+--------------------------------------------------------------------
    diff |            7.032464    1.530502                4.032326     10.0326
------------------------------------------------------------------------------
    diff = mean(Man) - mean(Woman)                                t =   4.5949
Ho: diff = 0                                     degrees of freedom =     8877

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

We can start by noting that the overall mean is 308.4708. As can be seen, men have a slightly higher mean value compared to women (311.943 vs. 304.9106).

The t-test statistic in this example is 4.5949, with 8877 degrees of freedom. The corresponding p-value is 0.0000 (look below “Ha: diff != 0”). This is below 0.05, which allows us to reject the null hypothesis (which states that there is no mean difference between the two groups).

In other words, there is a significant difference in mean cognitive test scores between men and women in this example, to the advantage of men.

6.2.1 Non-parametric alternative: Mann-Whitney u-test

It is not uncommon that at least one of the assumptions behind the independent samples t-test is violated. While you most commonly will be able to conduct the test anyway, it is important to be aware of the possible problems. Alternatively, you can use the Mann-Whitney u-test instead, which is a nonparametric independent t-test that relaxes some of the assumptions that were presented earlier.

Function

Basic command	ranksum testvar, by(groupvar)	
Explanations	testvar	Insert the name of the variable that you want to test.
	groupvar	Insert the variable defining the two groups.
More information	help ranksum	

Practical example

Dataset: StataData1.dta

Name	Label
cognitive	Cognitive test score (Age 15, Year 1985)
sex	Sex

```
ranksum cognitive, by(sex)
```

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

         sex |      obs    rank sum    expected
-------------+---------------------------------
         Man |     4495    20573556    19957800
       Woman |     4384    18849204    19464960
-------------+---------------------------------
    combined |     8879    39422760    39422760

unadjusted variance   1.458e+10
adjustment for ties  -4007390.8
                     ----------
adjusted variance     1.458e+10

Ho: cognit~e(sex==Man) = cognit~e(sex==Woman)
             z =   5.100
    Prob > |z| =   0.0000
```

The z-statistic in this example is 5.100, with a p-value of 0.0000. Since the p-value is below 0.05, this allows us to reject the null hypothesis (which states that there is no mean difference between the two groups).

In other words, there is a significant difference in mean cognitive test scores between men and women in this example, to the advantage of men (just like the previous t-test also showed).

6.3 T-test: Paired samples

Quick facts	
Number of variables	Two (reflecting repeated measurement points)
Scale of variable(s)	Continuous (ratio/interval)

A dependent or paired samples t-test is used to see the difference or change between two measurement points. This is a parametric type of test. For example, you could apply this test to see if the staff’s job satisfaction has improved after their boss has taken a course in “socio-emotional skills” compared to before, or if the rate of cigarette smoking in certain schools has declined since the introduction of a new intervention programme.

For the independent samples t-test, you were supposed to have two groups for which you compared the mean. For the paired samples t-test, you instead have two measurements of the same variable, and you look at whether there is a change from one measurement point to the other.

Happiness score before summer vacation	Happiness score after summer vacation

Assumptions

First, you have to check your data to see that the assumptions behind the paired samples t-test hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Continuous variables	Your two variables should be continuous (i.e. interval/ratio). For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: "How satisfied with your income are you?" (on a scale 1-10) or "To what extent do you agree with the previous statement?" (on a scale 1-5).
Two measurement points	Your two variables should reflect one single phenomenon, but this phenomenon is measured at two different time points for each individual.
Normal distribution	Both variables need to be approximately normally distributed. Use a histogram to check (see Section 4.5).
No outliers in the comparison between the two measurement points	For example, if one individual has an extremely low value at the first measurement point and an extremely high value at the second measurement point (or vice versa), this will distort the test. Use a scatterplot to check (see Section 7.1.1).

Function

Basic command	ttest testvar1==testvar2	
Explanations	testvar1	Insert the name of the variable for the first measurement point.
	testvar2	Insert the name of the variable for the first measurement point.
More information	help ttest	

Practical example

Dataset: StataData1.dta

Name	**Label**
unemp_42	Days in unemployment (Age 42, Year 2012)
unemp_43	Days in unemployment (Age 43, Year 2013)

Note Since the variables are extremely skewed (a lot of zeros), we are restricting the analysis to those who did not have the value 0 at age 42.

```
ttest unemp_42==unemp_43 if unemp_42!=0
```

```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
unemp_42 |   1,058    147.9981    3.043827    99.00628    142.0255    153.9707
unemp_43 |   1,058    38.62571    2.389486    77.72261    33.93703    43.31438
---------+--------------------------------------------------------------------
    diff |   1,058    109.3724    3.518878    114.4582    102.4676    116.2772
------------------------------------------------------------------------------
     mean(diff) = mean(unemp_42 - unemp_43)                       t =  31.0816
 Ho: mean(diff) = 0                              degrees of freedom =     1057

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

As can be seen, the mean is much higher at age 42 (147.998) compared to age 43 (38.625), which is a difference of 109.372.

The t-test statistic in this example is 31.0816, with 1057 degrees of freedom. The corresponding p-value is 0.0000 (look below “Ha: mean(diff) != 0”). This is below 0.05, which allows us to reject the null hypothesis (which is that there is no mean difference between the two measurement points).

In other words, there is a significant difference between the two measurement points, suggesting that the mean number of days in unemployment is significantly lower at age 43 compared to age 42 in this example.

6.3.1 Non-parametric alternative: Wilcoxon signed rank test

It is not uncommon that at least one of the assumptions behind the paired samples t-test is violated. While you most commonly will be able to conduct the test anyway, it is important to be aware of the possible problems.

Alternatively, you can use the Wilcoxon signed rank test instead, which is a nonparametric paired samples t-test that relaxes some of the assumptions that were presented earlier. The null hypothesis is that the distribution at the two measurement points are the same.

You should nevertheless note that this test is primarily suitable for samples with <= 200 observations. By default, if you then will obtain an exact p-value based on the actual randomization distribution of the test statistic. If you have more than 200 observations, you need to use the option called exact. The exact computation is only available for samples where n <= 2,000. Regardless of sample size, you always get an approximate p-value which is based on a normal approximation to the randomization distribution.

Function

Basic command	signrank testvar1=testvar2	
Useful options	signrank testvar1= testvar2, exact	
Explanations	testvar1	Insert the name of the variable for the first measurement point.
	testvar2	Insert the name of the variable for the first measurement point.
	exact	Specifies that the exact p-value be computed in addition to the approximate p-value.
More information	help signrank	

Practical example

Dataset: StataData1.dta

Name	**Label**
unemp_42	Days in unemployment (Age 42, Year 2012)
unemp_43	Days in unemployment (Age 43, Year 2013)

Note Since the variables are extremely skewed (a lot of zeros), we are restricting the analysis to those who did not have the value 0 at age 42.

```
signrank unemp_42=unemp_43 if unemp_42!=0, exact
```

```
Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |      945      518604      280104
    negative |      111       41604      280104
        zero |        2           3           3
-------------+---------------------------------
         all |     1058      560211      560211

unadjusted variance    98830557
adjustment for ties   -1179.625
adjustment for zeros      -1.25
                     ----------
adjusted variance      98829376

Ho: unemp_42 = unemp_43
             z =  23.991
    Prob > |z| =   0.0000
    Exact Prob =   0.0000
```

The z-statistic in this example is 23.991, with an approximate p-value (Prob > |z|) of 0.0000, and an exact p-value (Exact Prob) of 0.0000. Since the latter p-value is below 0.05, this allows us to reject the null hypothesis (which states that there is no difference in the distribution between the two measurement points), thus confirming what we also saw for the paired samples t-test.

6.4 One-way ANOVA

Quick facts	
Number of variables	One group variable (x) One test variable (y)
Scale of variable(s)	Group variable: categorical (nominal/ordinal) Test variable: continuous (ratio/interval)

The one-way ANOVA is very similar to the independent samples t-test. The difference is that the one-way ANOVA allows you to have more than two categories in your group variable. For example, you can compare how many cups of coffee people drink per day depending on if they have a low-stress, medium-stress, or high-stress job. Or you can see if the number of days of paternity leave differs between fathers in Sweden, Denmark, Norway and Finland.

Note The one-way ANOVA does not automatically tell you exactly which groups are different from each other; it only tells you that at least two of the groups differ in terms of the outcome.

Mean number of ice cones per week during May in Swedish children ages 5-10	Mean number of ice cones per week during June in Swedish children ages 5-10	Mean number of ice cones per week during July in Swedish children ages 5-10

Assumptions

First, you have to check your data to see that the assumptions behind the one-way ANOVA hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Continuous and normally distributed test variable	Your test variable should be continuous (i.e. interval/ratio) and normally distributed. For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: "How satisfied with your income are you?" (on a scale 1-10) or "To what extent do you agree with the previous statement?" (on a scale 1-5).
Two or more unrelated categories in the group variable	Your group variable should be categorical (i.e. nominal or ordinal) and consist of two or more groups. Unrelated means that the groups should be mutually excluded: no individual can be in more than one of the groups. For example: low vs. medium vs. high educational level; liberal vs. conservative vs. socialist political views; or poor vs. fair, vs. good vs. excellent health; and so on.
Equal variance	The variance in the test variable should be equal across the groups of the group variable.
No outliers	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.

Function

Basic command	oneway testvar groupvar	
Useful options	oneway testvar groupvar, tab oneway testvar groupvar, bonferroni	
Explanations	testvar	Insert the name of the test variable.
	groupvar	Insert the name of the group variable.
	tab	Produce summary table.
	bonferroni	Reports the results from a Bonferroni multiple-comparison test.
Short names	tab	Tabulate
Notes	Options can be used simultaneously, e.g: oneway testvar groupvar, tab bonferroni	
More information	help oneway	

Practical example

Dataset: StataData1.dta

Name	**Label**
income	Annual salary income (Age 40, Year 2010)
educ	Educational level (Age 40, Year 2010)

`oneway income educ, tab bonferroni`

```
Educational |
 level (Age |   Summary of Annual salary income
   40, Year |         (Age 40, Year 2010)
      2010) |        Mean   Std. Dev.       Freq.
------------+------------------------------------
  Compulsor |   164316.86   86528.011       1,376
  Upper sec |   178904.49   97666.116       3,560
  Universit |   238989.77   131440.58       3,128
------------+------------------------------------
      Total |   199722.22    114851.4       8,064

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      8.0909e+12      2   4.0454e+12    331.85     0.0000
 Within groups      9.8267e+13   8061   1.2190e+10
------------------------------------------------------------------------
    Total           1.0636e+14   8063   1.3191e+10

Bartlett's test for equal variances:  chi2(2) = 452.7846  Prob>chi2 = 0.000

            Comparison of Annual salary income (Age 40, Year 2010)
                   by Educational level (Age 40, Year 2010)
                                (Bonferroni)
Row Mean-|
Col Mean |   Compulso   Upper se
---------+----------------------
Upper se |    14587.6
         |      0.000
         |
Universi |    74672.9    60085.3
         |      0.000      0.000
```

The first table provides some summary statistics. Here we can see that the mean income for those with compulsory education is 164316.86, versus 178904.49 for those with upper secondary education, and 238989.77 for those with university education.

Next table gives the F statistics, which in this example is 331.85. The p-value is 0.0000 (i.e. below 0.05), which tells us that the means between the groups are not equal. At the lower part of the table, we get the results from Barlett's test for equal variances. P-value (Prob>chi2) below 0.05 suggests that the assumption of equal variances is violated. However, this can often happen with large datasets like the one used in the example. Also, the test is rather sensitive to data which is not normally distributed (the income variable used here is slightly skewed).

The fact that the F statistics tell us that the means between the groups are not equal says very little about wherein the differences lie: which groups are different? To answer this, we can take a look at the third table, showing the results from the Bonferroni test. The first lines of entries for each combination represents the mean differences. The second lines of entries are Bonferroni-adjusted p-values. In this example, they are all 0.000 – which suggest that there are significant differences between all three groups.

Postestimation commands

There are many different postestimation commands that you can apply to ANOVA. These options are described here:

```
help anova postestimation
```

6.4.1 Non-parametric alternative: Kruskal-Wallis ANOVA

It is not uncommon that at least one of the assumptions behind the one-way ANOVA is violated. While you most common will be able to conduct the test anyway, it is important to be aware of the possible problems.

Alternatively, you can conduct a Kruskal-Wallis ANOVA, which is nonparametric type of ANOVA. This test is robust over moderate violations against the normality assumption. Note, however, that the group sizes should be approximately equal and that the distributions of the groups also are approximately equal (they cannot be skewed in different directions, i.e. one is positively skewed and another is negatively skewed).

Note The Kruskal-Wallis ANOVA will only tell you that there is a difference between the groups (or not), but not which of groups that are different from one another.

Function

Basic command	kwallis testvar, by(groupvar)	
Explanations	testvar	Insert the name of the variable that you want to test.
	groupvar	Insert the variable defining the groups.
More information	help kwallis	

Practical example

Dataset: StataData1.dta

Name	**Label**
income	Annual salary income (Age 40, Year 2010)
educ	Educational level (Age 40, Year 2010)

```
kwallis income, by(educ)
```

```
Kruskal-Wallis equality-of-populations rank test

  +-------------------------------------+
  |             educ |   Obs | Rank Sum |
  |------------------+-------+----------|
  |       Compulsory | 1,376 | 4.61e+06 |
  | Upper secondary  | 3,560 | 1.29e+07 |
  |       University | 3,128 | 1.50e+07 |
  +-------------------------------------+

chi-squared =   548.642 with 2 d.f.
probability =     0.0001

chi-squared with ties =   549.333 with 2 d.f.
probability =     0.0001
```

The table provides some summary statistics. Here we can see number of observations per group (Obs) and the rank of each group (Rank Sum). These ranks are u-values.

Below the table, two sets of chi2 values and probabilities are reported. If the rank variable (in this case income) do not uniquely define individuals (i.e. individuals can have the same income and thus the same rank), then we should focus on the latter set ("chi-squared with ties").

In this example, the chi2 value is 549.333 with 2 degrees of freedom. The probability is moreover 0.0001. Since this is below 0.05, it means that there is a significant difference in income according to educational level (confirming what we found for the one-way ANOVA).

6.5 Chi-square

Quick facts	
Number of variables	Two
Scale of variable(s)	Categorical (nominal/ordinal)

There are two different forms of the chi-square test: a) The multidimensional chi-square test, and b) The goodness of fit chi-square test. It is the first form that will be covered in this part of the guide. The second form is discussed in other sections.

The multidimensional chi-square test assesses whether there is a relationship between two categorical variables. For example, let us assume that you want to see if young women smoke more than young men. The variable gender has two categories (men and women) and, in this particular case, the variable smoking consists of the categories: no smoking, occasional smoking and frequent smoking. The multidimensional chi-square test can be thought of as a simple crosstable where the distribution of these two variables is displayed:

	No smoking	*Occasional smoking*	*Frequent smoking*
Men (age 15-24)	85%	10%	5%
Women (age 15-24)	70%	20%	10%

Assumptions

First, you have to check your data to see that the assumptions behind the chi-square test hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Two or more unrelated categories in both variables	Both variables should be categorical (i.e. nominal or ordinal) and consist of two or more groups. Unrelated means that the groups should be mutually excluded: no individual can be in more than one of the groups. For example: low vs. medium vs. high educational level; liberal vs. conservative vs. socialist political views; or poor vs. fair, vs. good vs. excellent health; and so on.

Function

Basic command	tab varname1 varname2, chi2	
Useful options	tab varname1 varname2, chi2 exact	
Explanations	varname1	Insert the name of the first variable you want to use (is included as the row variable).
	varname2	Insert the name of the second variable you want to use (is included as the column variable).
	chi2	Report Pearson's chi-squared.
	exact	Report Fisher's exact test (useful if you have empty cells in your crosstable).
Short names	tab	Tabulate
More information	help tabulate twoway	

Practical example

Dataset: StataData1.dta

Name	**Label**
marstat40	Marital status (Age 40, Year 2010)
earlyret	Early retirement (Age 50, Year 2020)

```
tab earlyret marstat40, chi2
```

```
     Early |
retirement |
  (Age 50, |     Marital status (Age 40, Year 2010)
Year 2020) |   Married  Unmarried   Divorced    Widowed |     Total
-----------+--------------------------------------------+----------
        No |     4,161      1,904      1,407         60 |     7,532
       Yes |       394        470        313         24 |     1,201
-----------+--------------------------------------------+----------
     Total |     4,555      2,374      1,720         84 |     8,733

          Pearson chi2(3) = 217.3417   Pr = 0.000
```

Here we can see the crosstable of our two variables. It is followed by the chi-square value (Pearson chi2) and a p-value (Pr). If the p-value is below 0.05 it means that the two variables are not independent from one another. In this example, since the p-value is 0.000, it means that there are significant differences in early retirement according to marital status (or, by principle, vice versa).

7. CORRELATION ANALYSIS

Outline

Content

This chapter deals with correlation. We begin with descriptive statistics, in terms of scatterplots, and continue with correlation analysis (including non-parametric alternatives).

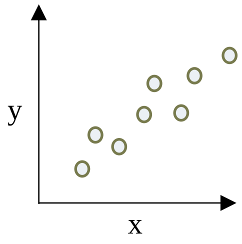
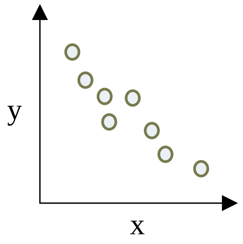
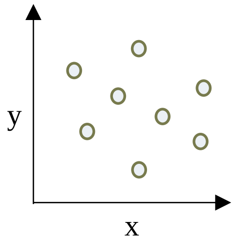
7.1 Descriptives

7.1.1 Scatterplot

Quick facts	
Number of variables	Two
Scale of variable(s)	Continuous (ratio/interval)

When we had two categorical variables, we could produce a crosstable to see how these two variables were related. If we have two continuous variables, we may use something called a scatterplot instead. Each dot in the scatterplot represents one individual in our data. We may also include a reference line here, to see if we have a pattern in our data (this will be discussed later).

The scatterplot can thus be used to illustrate how two continuous variables co-vary – or “correlate” – in their pattern of values. If increasing values of one variable correspond to increasing values of another variable, it is called a positive correlation. If increasing values of one variable correspond to decreasing values of another variable, we have a negative correlation. In the graph below, different types of correlation are presented. The letter “x” stands for x-axis (horizontal axis) and the letter “y” stands for y-axis (vertical axis).

Positive correlation	**Negative correlation**	**No correlation**
		

Note While not addressed here, patterns can of course also be non-linear (in contrast to the positive and negative correlations shown in the graphs above).

Function

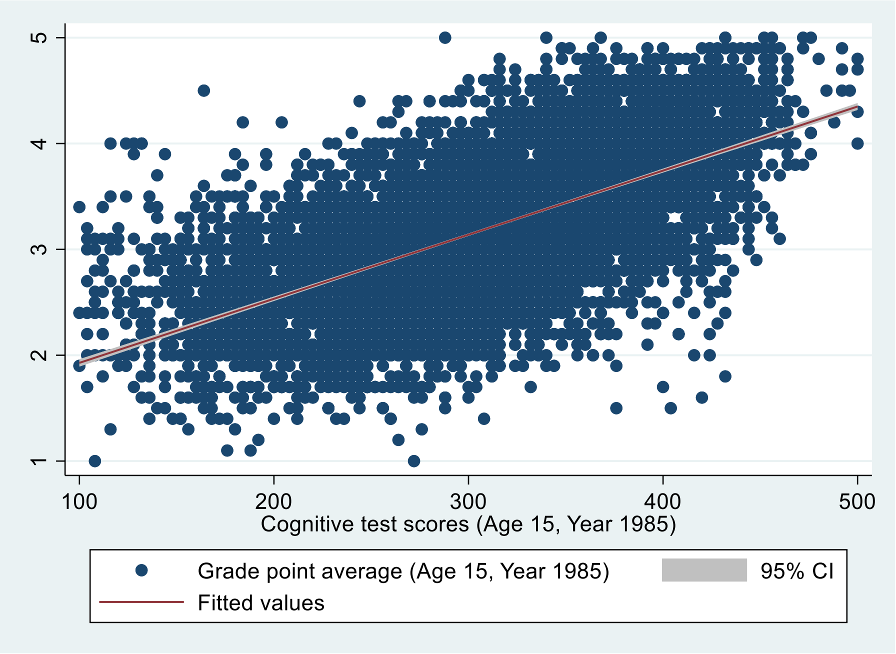
Basic command	graph twoway scatter yvar xvar	
Useful options	graph twoway (scatter yvar xvar) (lfit yvar xvar) graph twoway (scatter yvar xvar) (lfitci yvar xvar)	
Explanations	yvar	Insert the name of the first variable you want to use. This variable will be chosen for the y-axis (vertical axis).
	xvar	Insert the name of the first variable you want to use. This variable will be chosen for the x-axis (horizontal axis).
	lfit	Fit a regression line.
	lfitci	Fit a regression line and include confidence intervals.
More information	help scatter	

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)

```
graph twoway (scatter gpa cognitive) (lfitci gpa cognitive)
```



In the scatterplot above, we display gpa on the y-axis (vertical axis) and cognitive on the x-axis (horizontal axis). We can see a quite clear positive correlation here: the higher the cognitive test scores, the higher the grade point average. This is also illustrated by the fitted regression line.

Note You can use the Graph Editor (see Section 2.1.4) to further edit the scatterplot.

7.2 Correlation analysis

Quick facts	
Number of variables	Two or more
Scale of variable(s)	Continuous (ratio/interval)

A correlation analysis tests the relationship between two continuous variables in terms of: a) how strong the relationship is, and b) in what direction the relationship goes. The strength of the relationship is given as a coefficient (the Pearson product-moment correlation coefficient, or simply Pearson's r) which can be anything between -1 and 1. But how do we know if the relationship is strong or weak? This is not an exact science, but here is one rule of thumb:

Strength		
Negative	**Positive**	
-1	1	**Perfect**
-0.9 to -0.7	0.7 to 0.9	**Strong**
-0.6 to -0.4	0.4 to 0.6	**Moderate**
-0.3 to -0.1	0.1 to 0.3	**Weak**
0	0	**Zero**

Thus, the coefficient can be negative or positive. These terms, "negative" and "positive", are not the same as good and bad (e.g. excellent health or poor health; high income or low income). They merely reflect the direction of the relationship.

Direction	
Negative	As the values of Variable 1 increases, the values of Variable 2 *decreases*
Positive	As the values of Variable 1 increases, the values of Variable 2 *increases*

Note Correlation analysis does not imply anything about causality: Variable 1 does not *cause* Variable 2 (or vice versa). The correlation analysis only says something about the degree to which the two variables co-vary (in a linear fashion).

Assumptions

First, you have to check your data to see that the assumptions behind the correlation analysis hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Two continuous variables	Both variables should be continuous (i.e. interval/ratio). For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still rather common to use ratings as continuous variables, such as: "How satisfied with your income are you?" (on a scale 1-10) or "To what extent do you agree with the previous statement?" (on a scale 1-5).
Normal distribution	Both variables need to be approximately normally distributed. Use a histogram to check (see Section 4.5).
Linear relationship between the two variables	There needs to be a linear relationship between your two variables. You can check this by creating a scatterplot (described in Section 7.1.1).
No outliers	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.

Function alternative 1

Basic command	corr varname1 varname2	
Explanations	varname1	Insert the name of the first variable you want to use.
	varname2	Insert the name of the second variable you want to use.
Short names	corr	Correlate
Notes	You can include more than two variables at the same time in the analysis.	
More information	help correlate	

Function alternative 2

Basic command	pwcorr varname1 varname2	
Useful options	pwcorr varname1 varname2, sig pwcorr varname1 varname2, star(level)	
Explanations	varname1	Insert the name of the first variable you want to use.
	varname2	Insert the name of the second variable you want to use.
	sig	Print a p-value for each entry.
	star(level)	Denote statistically significant entries with an asterisk (*). Change “level” to the preferred significance level (e.g. 0.05, 0.01, 0.001).
Notes	Options can be used simultaneously, e.g.: pwcorr varname1 varname2, sig star(level) You can include more than two variables at the same time in the analysis.	
More information	help pwcorr	

There are two alternative commands if you want to do a correlation analysis in Stata: corr and pwcorr. The first difference between these commands has to do with how Stata handles missing values, and is only relevant if you include more than two variables in the analysis. In that case, corr will use listwise deletion (i.e. removing all observations that have missing information from any of the included variables), whereas pwcorr uses pairwise deletion (i.e. only removing observations with missing values for each specific pair of variables). The second difference is that they have different options (with the options for pwcorr being slightly more useful).

Since we highly recommend that you restrict your analysis to a sample with only valid information for all study variables anyway, it does not matter whether you would go for corr or pwcorr. But since we like the options to include p-values and asterisks, we will base our following example on pwcorr.

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)

```
pwcorr gpa cognitive, sig star(0.05)
```

```
             |      gpa cognit~e
-------------+------------------
         gpa |   1.0000 
             |
             |
   cognitive |   0.6276*  1.0000 
             |   0.0000
```

In the diagonal, we can see the perfect (and totally irrelevant) correlations between gpa and gpa, and between cognitive and cognitive. What is interesting here is the correlation coefficient between cognitive and gpa: 0.6276. According to our earlier specified rules of thumb, this would be a moderately strong correlation (close to strong). We get a p-value of 0.0000, which is lower than $p<0.05$ (as we can also note this by the asterisk). Thus, the correlation between cognitive test score and grade point average is statistically significant.

7.3 Non-parametric alternatives: Spearman's rank correlation and Kendall's rank correlation

It is not uncommon that the assumption of normality (i.e. normally distributed variables) is violated. As an alternative, you can conduct Spearman's rank correlation (also called Spearman's rho) or Kendall's rank correlation (also called Kendall's tau) instead.

Function alternative 1

Basic command	spearman varname1 varname2	
Useful options	spearman varname1 varname2, star(level)	
Explanations	varname1	Insert the name of the first variable you want to use.
	varname2	Insert the name of the second variable you want to use.
	star(level)	Denote statistically significant entries with an asterisk (*). Change "level" to the preferred significance level (e.g. 0.05, 0.01, 0.001).
Notes	You can include more than two variables at the same time in the analysis. Asterisks only appear if you specify more than two variables.	
More information	help spearman	

Function alternative 2

Basic command	ktau varname1 varname2	
Useful options	ktau varname1 varname2, star(level)	
Explanations	varname1	Insert the name of the first variable you want to use.
	varname2	Insert the name of the second variable you want to use.
	star(level)	Denote statistically significant entries with an asterisk (*). Change "level" to the preferred significance level (e.g. 0.05, 0.01, 0.001).
Notes	You can include more than two variables at the same time in the analysis. Asterisks only appear if you specify more than two variables.	
More information	help ktau	

The two commands are largely the same. However, it has been suggested that Kendall's rank correlation is slightly more robust since it is less sensitive to small samples (which usually has bigger problems with outliers). We have chosen to stick to Spearman's rank correlation since our sample is large. We have not included the option to show asterisks, since we only use two variables.

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)

```
spearman gpa cognitive
```

```
Number of obs =     8751
Spearman's rho =       0.6361

Test of Ho: gpa and cognitive are independent
    Prob > |t| =       0.0000
```

The correlation coefficient (Spearman's rho) is 0.6361. This is roughly the same as the coefficient we got with pwcorr. The p-value (Prob > |t| =) is 0.0000, which is lower than $p<0.05$. Thus, the correlation between cognitive test score and grade point average is statistically significant also with this test.

8. FACTOR ANALYSIS

Outline

Content

This chapter focuses entirely on factor analysis, and also includes a section on how to assess internal consistency with Cronbach's alpha. Factor analysis can be seen as a method of data reduction, which is rather different from other methods presented in this guide.

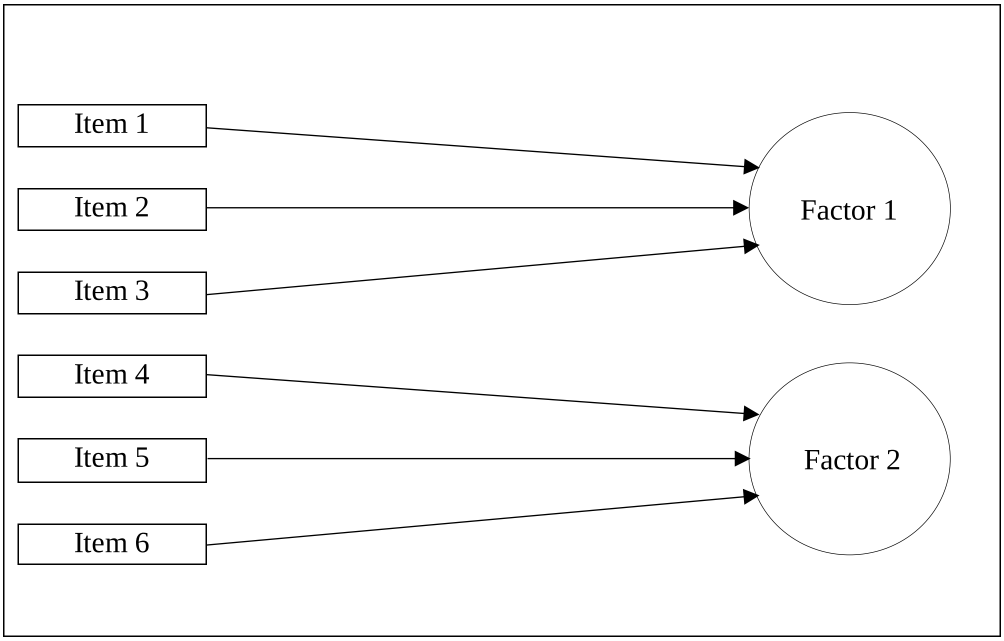
8.1 Introduction

Quick facts	
Number of variables	Two or more
Scale of variable(s)	Continuous (ratio/interval) or approximately continuous

There are two general types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The most important difference is that CFA has clear expectations on a specific factor structure, which is what we test, whereas EFA does not rely upon any expected structure. In this guide, we will focus on EFA (hereafter referred to simply as factor analysis). If you are interested in learning more about CFA, we suggest that you look up structural equation modelling (SEM), which is a very useful framework.

More information	help sem

The main feature of factor analysis is that is enables us to investigate the underlying structure in the pattern of correlations between a number of variables (often referred to as "items"). There are many different ways of using factor analysis, but one very practical application is cases where we have several items from a questionnaire that we want to create an index for. By conducting a factor analysis, we are able to see whether the items represent the same factor (or "dimension"). If so, we can create our index. Factor analysis can also tell us how to improve our index (e.g. by excluding one or more items), or if we actually have more than one factor and thus need to consider creating separate indices.



8.2 Assumptions

First, you have to check your data to see that the assumptions behind the factor analysis hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Ratio/interval/ordinal variables	Your variables should be continuous (i.e. interval/ratio) or ordinal (but still approximately continuous). For example: Income, height, weight, number of years of schooling, or ratings.
Linear associations	The variables in the factor analysis should be associated with each other in a linear fashion (use scatterplots to check, see Section 7.1.1).
Sample size	Factor analysis requires rather large samples. However, recommendations on this topic vary greatly. Some recommendations highlight the absolute sample size (here, lower limits range from n=100 to n=500) whereas others say that subject-to-variable ratio is important (and here, ratios from 2:1 to 20:1 are suggested).
No outliers	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.

Suppose that we have asked a bunch of individuals, six questions about their health. We conduct a factor analysis to see how many dimensions these questions reflect: do all questions reflect only one dimension (namely "health") or can they be categorised into two or more dimensions (i.e. different types of health)?

8.3 Number of factors

How do we ascertain how many factors/dimensions there are in our data? Well, there are several different ways to do this. It is nevertheless important to keep in mind that we want to aim for a balance between simplicity and accuracy: as few factors as possible, that explain as much of the variance as possible.

Determining the number of factors	
Eigenvalue > 1	Eigenvalues are indicators of the variance explained by a factor. Using the rule "eigenvalue is greater than one" is very common. The reasoning behind this rule is that a factor should account for at least as much variance as any single variable. Thus, the average of all eigenvalues is one, and the factor analysis should thus extract factors that have an eigenvalue greater than this average value.
Scree plot	In a scree plot, factors have their eigenvalues plotted alongside the y-axis (i.e. vertical axis) in the order or magnitude. Factors explaining large amounts of variable appear to the left, whereas factors explaining little variance are aligned to the right. The somewhat weird task is here to "locate the elbow". This means to identify the number of factors stated before the line starts becoming flat.
Communalities and uniqueness	Communality refers to how much variance of each variable that can be reproduced by the factor extraction. A general rule of thumb is that the extracted factors should explain at least 50% of the variables' variance (i.e. the communalities should be between 0.5 and 1). Stata, however, reports on the opposite of communalities: uniqueness (which is 1-communality). The similar threshold applies here, i.e. the uniqueness should be between 0 and 0.5.

8.4 Factor loadings

Once we have decided on the number of factors, we retrieve the "factor loadings". A factor loading is basically a correlation coefficient (see Chapter 7) and, thus, it varies between -1 and +1 (where a value closer to -1 or +1 indicates a stronger correlation). Factor loadings are given for each variable, for each factor separately. In other words, a factor loading shows how strongly a certain variable correlates with the given factor. There are no exact rules for deciding on when a loading is strong enough, but one suggested rule of thumb is below -0.5 or above 0.5. However, sometimes a variable has strong loadings for more than one factor (called "cross-loading"). This can for example happen if you have not extracted enough factors, or if the factors are correlated. Sometimes a variable has weak loadings for all factors; this may suggest that this variable is weakly related to all other variables or that you need to explore an additional factor (or maybe even exclude this specific variable).

8.5 Rotation

A factor analysis has the most interpretative value when: 1) Each factor loads strongly on only one factor; 2) Each factor shows at least three strong loadings; 3) Most loadings are either high or low; and 4) We get a "simple" factor structure. Rotation is a way of maximizing high loadings and minimizing low loadings so that we get the simplest factor structure possible. There are two main types of rotation:

Rotation	
Orthogonal	Assumes that the factors are uncorrelated Examples of sub types: varimax, quartimax, and equamax
Oblique	Assumes that the factors are correlated Examples of sub types: promax and oblimin

Thus, orthogonal rotation relies on the assumption that the factors are not correlated to each other, i.e. that the different factors represent different unrelated dimensions of what you are examining. This is not always the case. For example, if you have several variables measuring health, and find one factor that reflects physical health and another one reflecting psychological health, it may not be reasonable to assume that physical and psychological health two unrelated dimension. In that case, you need to change the type of rotation to oblique.

In Stata, orthogonal rotation with the varimax option is default.

8.6 Postestimation

There are some tests that you can use to decide whether your factor analysis offers a good fit for your data or not. For example, there is a test called Kaiser-Meyer-Olkin Measure of Sampling Adequacy (in short: the KMO test), which reflects the sum of partial correlations relative to the sum of correlations. It varies between 0 and 1, where a value closer to 1 is better. It has been suggested to use 0.5 as a minimum requirement. Thus, if the value is lower than 0.5, factor analysis may be inappropriate.

8.7 Factor analysis vs principal component analysis

Principal component analysis (PCA) is a term that is often used interchangeably with factor analysis. While both approaches aim to simplify the structure of a set of variables and the analyses are structured in similar ways, they are not exactly the same thing. PCA performs data reduction by using a linear combination of a set of variables, in order to create one or more index variables (components). Factor analysis is modelling the measurement of a latent (i.e. unobserved) variable.

To make it even more confusing, many statistical programs (e.g. SPSS) apply PCA as the default estimation method for factor analysis. In Stata, PCA is not default (but an option). Rather, Stata uses the principal-factor method (pf) to analyse the correlation matrix. When the principal-component factor method (pcf) is used, the communalities are assumed to be 1.

8.8 A practical example

Function Step 1

Basic command	factor varlist	
Useful options	factor varlist, mineigen(number) factor varlist, pcf *or* ipf *or* ml	
Explanations	varlist	List which variables that you want to include in the analysis.
	pcf *or* ipf *or* ml	Specify the estimation method. Default is pf.
Short names	pf	Principal factor method
	pcf	Principal-component factor method
	ipf	Iterated principal-factor method
	ml	Maximum-likelihood factor method
Notes	Options can be used simultaneously, e.g: factor varlist, mineigen(number) pcf	
More information	help factor	

Performing a factor analysis can be seen as an iterative process: you conduct the analysis, evaluate it, might tweak it a bit, and then conduct it again. We will start by performing a simple factor analysis with the principal-component factor method (pcf).

Practical example

Dataset: StataData2.dta

Name	**Label**
imp_ideas	Important to think up new ideas
imp_rich	Important to be rich
imp_secure	Important living in secure surroundings
imp_good	Important to have a good time
imp_help	Important to help the people
imp_success	Important being very successful
imp_risk	Important with adventure and taking risks
imp_behave	Important to always behave properly
imp_environ	Important looking after the environment
imp_trad	Important with tradition

```
factor imp_ideas-imp_trad, pcf
```

```
Factor analysis/correlation                        Number of obs    =     58,466
    Method: principal-component factors            Retained factors =          2
    Rotation: (unrotated)                          Number of params =         19

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.98870      1.36904            0.2989       0.2989
        Factor2  |      1.61967      0.69108            0.1620       0.4608
        Factor3  |      0.92859      0.08586            0.0929       0.5537
        Factor4  |      0.84274      0.08451            0.0843       0.6380
        Factor5  |      0.75823      0.11357            0.0758       0.7138
        Factor6  |      0.64466      0.04326            0.0645       0.7783
        Factor7  |      0.60139      0.04007            0.0601       0.8384
        Factor8  |      0.56133      0.02132            0.0561       0.8945
        Factor9  |      0.54000      0.02531            0.0540       0.9485
       Factor10  |      0.51469            .            0.0515       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(45) = 1.0e+05 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
       imp_ideas |   0.5303    0.3315 |      0.6089
        imp_rich |   0.4695    0.5222 |      0.5068
      imp_secure |   0.5888   -0.2645 |      0.5834
        imp_good |   0.4139    0.4282 |      0.6453
        imp_help |   0.6127   -0.2993 |      0.5350
     imp_success |   0.6658    0.2754 |      0.4809
        imp_risk |   0.4091    0.5807 |      0.4954
      imp_behave |   0.6034   -0.3716 |      0.4978
     imp_environ |   0.5787   -0.3684 |      0.5293
        imp_trad |   0.5330   -0.4552 |      0.5087
    -------------------------------------------------
```

In the first table, we first look at the column called Eigenvalue. We see that Factor1 and Factor2 produce eigenvalues above 1 (2.98870 and 1.61967, respectively). Next, focusing on the column called Proportion, we see that Factor1 accounts for 30% (0.2989) and Factor2 for (16% (0.1620) of the variance.

In the second table, we get the factor loadings for each item. When we use the option pcf, factor loadings are only shown for factors with eigenvalues above 1. For Factor1, loadings range between 0.4091 and 0.6658. For Factor2, they range between -0.3716 and 0.5807. The uniqueness values range between 0.4809 and 0.6089. Earlier, we suggested that factor loadings between 0.5 and 1 were acceptable, as well as uniqueness values between 0 and 0.5. Thus, our factor solution is quite poor. Moreover, it is not entirely clear which item belongs to which factor – we might need some rotation here.

Function step 2

Basic command	rotate	
Useful options	rotate, quartimax rotate, equamax rotate, promax(number) rotate, oblimin(number)	
Explanations	quartimax	Orthogonal rotation with the quartimax option.
	equamax	Orthogonal rotation with the equamax option.
	promax(number)	Oblique rotation with the promax option, replace "number" with preferred power (default is 3).
	oblimin(number)	Oblique rotation with the oblimin option, replace "number" with preferred gamma (default is 0).
Notes	Orthogonal rotation with the varimax option is default. To clear the results from rotation, use: rotate, clear	
More information	help rotate	

The next step is to rotate the results to minimize the complexity of the factor structure and facilitate interpretation. Since it is unlikely that our factors are uncorrelated (they seldom are, in the social sciences), we will go with an oblique rotation (more specifically, we try out promax).

Practical example

Dataset: StataData2.dta

Name	**Label**
imp_ideas	Important to think up new ideas
imp_rich	Important to be rich
imp_secure	Important living in secure surroundings
imp_good	Important to have a good time
imp_help	Important to help the people
imp_success	Important being very successful
imp_risk	Important with adventure and taking risks
imp_behave	Important to always behave properly
imp_environ	Important looking after the environment
imp_trad	Important with tradition

```
rotate, promax
```

```
Factor analysis/correlation                        Number of obs    =     58,466
    Method: principal-component factors            Retained factors =          2
    Rotation: oblique promax (Kaiser off)          Number of params =         19

    --------------------------------------------------------------------------
         Factor  |     Variance   Proportion    Rotated factors are correlated
    -------------+------------------------------------------------------------
        Factor1  |      2.62665       0.2627
        Factor2  |      2.35045       0.2350
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(45) = 1.0e+05 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
       imp_ideas |   0.1247    0.5794 |      0.6089
        imp_rich |  -0.0634    0.7171 |      0.5068
      imp_secure |   0.6192    0.0790 |      0.5834
        imp_good |  -0.0315    0.6034 |      0.6453
        imp_help |   0.6627    0.0607 |      0.5350
     imp_success |   0.2636    0.6018 |      0.4809
        imp_risk |  -0.1508    0.7370 |      0.4954
      imp_behave |   0.7110   -0.0087 |      0.4978
     imp_environ |   0.6911   -0.0191 |      0.5293
        imp_trad |   0.7245   -0.1210 |      0.5087
    -------------------------------------------------

Factor rotation matrix

    --------------------------------
                 | Factor1  Factor2
    -------------+------------------
         Factor1 |  0.8576   0.7306
         Factor2 | -0.5143   0.6828
    --------------------------------
```

The rotation made the factor loadings more clearly reflect the two factors.

If we identify for with factor each item has the higher loading, we can conclude that the two factors contain the following items:

Factor 1

- Important living in secure surroundings (security)
- Important to help the people (benevolence)
- Important to always behave properly (conformity)
- Important looking after the environment (universalism)
- Important with tradition (tradition)

Factor 2

- Important to think up new ideas (self-direction)
- Important to be rich (power)
- Important to have a good time (hedonism)
- Important being very successful (achievement)
- Important with adventure and taking risks (stimulation)

The ten variables used in this factor analysis actually stem from a theory of human values, developed by Schwartz. According to this theory, the variables should be categorised in the following way:

- Conservation: security, tradition, and conformity
- Openness to change: self-direction, stimulation, and hedonism
- Self-enhancement: power and achievement
- Self-transcendence: benevolence and universalism

If we compare the theoretical categories with the factors derived from factor analysis, we actually see that the Factor 1 includes all variables theoretically associated with conservation and self-transcendence, whereas Factor 2 includes all variables theoretically associated with openness to change and self-enhancement. What do we do with this information then? Well, we need to examine possible reasons as to why the factor analysis did not reveal the same factors as the theory proposes. If we find no apparent problems with the empirics (e.g. missing data, problems with the questionnaire itself, etc.) we may suggest that the theory needs to be modified. At least it is important to discuss the differences between the theory and the empirics.

Sometimes, we do not have a clear theory guiding the factor analysis and, thus, we have no a priori understanding about which factors that are reasonable to expect. In that case, it is common practice to focus on a factor solution with good properties (i.e. clear factor structure and high factor loadings). It is always a trade-off between theory and empirics: if theory has precedence over empirics, we may be more disposed to accept lower factor loadings.

In practice, all of this might mean that we go on to create two indices (e.g. sum score, or mean score), with each reflecting one factor, which we can then include in another analysis (such as regression analysis).

Function step 3

Basic command	estat kmo screeplot	
Explanations	kmo screeplot	Kaiser-Meyer-Olkin measure of sampling adequacy. Plot eigenvalues.
Notes	Orthogonal rotation with the varimax option is default. To clear the results from rotation, use: rotate, clear	
More information	help estat factor	

The third step is to do some postestimations, such as looking at the Kaiser-Meyer-Olkin measure of sampling adequacy and a screeplot, to see if our two-factor solution makes sense.

Note that if we here find any problems with our factor analysis or the chosen number of factors, we should go back and make some adjustments in order to find a better solution. For instance, we can try out different estimation methods, rotate the solution differently, or remove one or several of the items.

Practical example

Dataset: StataData2.dta

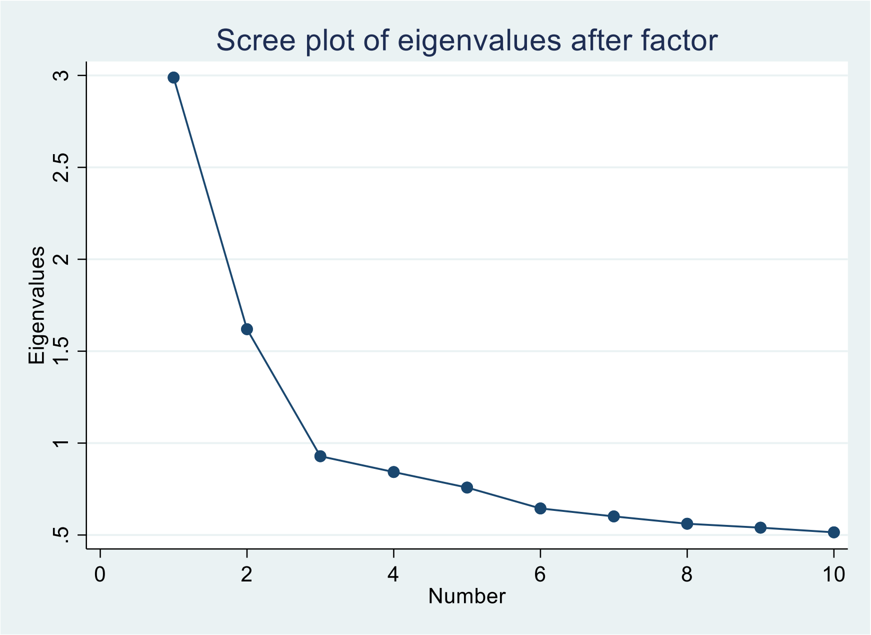
Name	**Label**
imp_ideas	Important to think up new ideas
imp_rich	Important to be rich
imp_secure	Important living in secure surroundings
imp_good	Important to have a good time
imp_help	Important to help the people
imp_success	Important being very successful
imp_risk	Important with adventure and taking risks
imp_behave	Important to always behave properly
imp_environ	Important looking after the environment
imp_trad	Important with tradition

```
estat kmo
```

```
Kaiser-Meyer-Olkin measure of sampling adequacy

    -----------------------
        Variable |      kmo
    -------------+---------
       imp_ideas |   0.8118
        imp_rich |   0.7261
      imp_secure |   0.8018
        imp_good |   0.8072
        imp_help |   0.8039
     imp_success |   0.8267
        imp_risk |   0.7250
      imp_behave |   0.8031
     imp_environ |   0.7916
        imp_trad |   0.7975
    -------------+---------
         Overall |   0.7918
    -----------------------
```

The KMO test produces an overall value of 0.7918, which shows that our factor analysis appears to be appropriate.

```
screeplot
```



In the screeplot, we can see that the “elbow” begins with the third factor, thus reflecting that a two-factor solution seems feasible.

8.9 Cronbach's alpha

Quick facts	
Number of variables	Two or more
Scale of variable(s)	Continuous (ratio/interval) or approximately continuous

When we have a composite measure (i.e. an index) – often derived from factor analysis – it is possible to evaluate it by means of the Cronbach's alpha. Formally speaking, the Cronbach's alpha is a measure of internal consistency; how closely related a number of items are as a group. The coefficient ranges between 0 and 1. A high alpha value indicates that items measure an underlying factor. However, it is not a statistical test but a test of reliability/consistency.

One important thing to note is that the Cronbach's alpha is affected by the number of variables: including a higher number of variables automatically increases the alpha value to some extent.

There are many rules of thumb with regard to what is considered a good or bad alpha value. Generally, an alpha value of at least 0.7 is considered acceptable.

Alpha values	
Between 0.7 and 1.0	Acceptable
Below 0.7	Not acceptable

Function

Basic command	alpha varlist	
Useful options	alpha varlist, item	
Explanations	varlist	List which variables that you want to include in the analysis.
	item	Display item-test and item-rest correlations. Useful to see what the effect would be if removing an item.
More information	help alpha	

Practical example

Dataset: StataData2.dta

Name	**Label**
imp_secure	Important living in secure surroundings
imp_help	Important to help the people
imp_behave	Important to always behave properly
imp_environ	Important looking after the environment
imp_trad	Important with tradition

Name	**Label**
imp_ideas	Important to think up new ideas
imp_rich	Important to be rich
imp_good	Important to have a good time
imp_success	Important being very successful
imp_risk	Important with adventure and taking risks

```
alpha imp_secure imp_help imp_behave imp_environ imp_trad, item
```

```
alpha imp_ideas imp_rich imp_good imp_success imp_risk, item
```

```
Test scale = mean(unstandardized items)

                                                            average
                                item-test     item-rest    interitem
Item          |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
imp_secure    | 61560   +       0.6676        0.4439        .5688143      0.6814
imp_help      | 61677   +       0.6627        0.4735        .5876615      0.6721
imp_behave    | 61479   +       0.7213        0.5092        .5121266      0.6547
imp_environ   | 61377   +       0.6766        0.4787        .5678347      0.6683
imp_trad      | 61621   +       0.7088        0.4802        .5221226      0.6680
-------------+-----------------------------------------------------------------
Test scale    |                                             .5517124      0.7165
-------------------------------------------------------------------------------


Test scale = mean(unstandardized items)

                                                            average
                                item-test     item-rest    interitem
Item          |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
imp_ideas     | 61206   +       0.6173        0.3970         .697211      0.6341
imp_rich      | 61547   +       0.6878        0.4579        .6041251      0.6059
imp_good      | 61405   +       0.6167        0.3579        .6943734      0.6523
imp_success   | 61218   +       0.6913        0.4772        .6046821      0.5982
imp_risk      | 61073   +       0.6863        0.4436        .6032935      0.6133
-------------+-----------------------------------------------------------------
Test scale    |                                               .64074      0.6723
-------------------------------------------------------------------------------
```

The scores for Test scale show the actual alpha values. For the first example, it is 0.7165 and for the second 0.6723. This is largely acceptable (at least for the first one). We can also see from the column called alpha that deleting any of the items would actually decrease the alpha score.

9. X, Y, AND Z

Outline

Content

This short chapter discusses the roles that variables (theoretically) can play when we conduct quantitative data analysis. It ends with a discussion about causality/causal inference.

9.1 Introduction

We talk a lot about variables in this guide, because variables are the cornerstones of quantitative data materials and quantitative data analysis. Other terms are sometimes used instead of “variables” – such as “indicators”, “measures” or “items”.

A variable is supposedly capturing the concept that we are interested in. The process of “translating” a concept to a variable is often called operationalisation. Some concepts are rather vague and not particularly easy to operationalise. One such example is “health”: should it be assessed by administrative health records or self-reported information? Is it simply the absence of disease or something more than that? Concepts such as “income” are more concrete since it refers to units (money) that can quite easily be measured. Still, it can be operationalised in many ways: monthly or annual income; income before or after taxes; individual income or household income; and so on.

Operationalisation should always be carefully reflected on and clearly motivated in research, since it might have important consequences for the analysis and therefore for the interpretation of the results.

Associations

In many types of analysis – such as regression analysis – we are interested in the association between two (or more) variables. The term association (or relationship) reflects the hypothesis that the variables are linked to one another in some way.

Effects

The way that regression analysis is constructed, however, assumes that one variable one variable has an “effect” on another variable. Here, we are talking about statistical effect, *not* causal effect. In other words, while we may find that one of the variables has a statistical effect on the other variable, it does not mean that we have proved that the first variable *causes* the second variable. A phrase commonly used in statistics to reflect this is: “correlation does not imply causation”. Just note that while it is more correct to talk about statistical effects, it is not all that uncommon that there are either implicit or explicit ideas about causal effects, guided by previous studies and theories. Sometimes such assumptions are quite reasonable, but the extent to which we can be certain about making causal inferences depends on the study design.

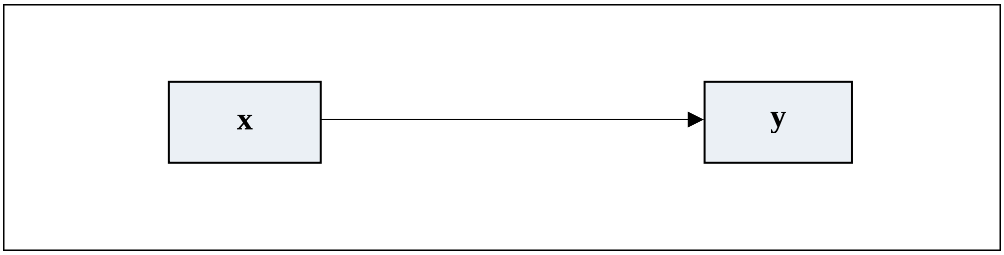
We will come back to the issue of causality later in this chapter (see Section 9.4).

X, y, and z

Variables play different roles in analysis. Researchers often use various terms to distinguish between these roles. Here, we will try to shed some light on the terms that are used.

Variables	
x	Independent variable; Exposure; Predictor
y	Dependent variable; Outcome
z	Covariate; Confounder; Mediator; Moderator; Effect modifier

9.2 X and y

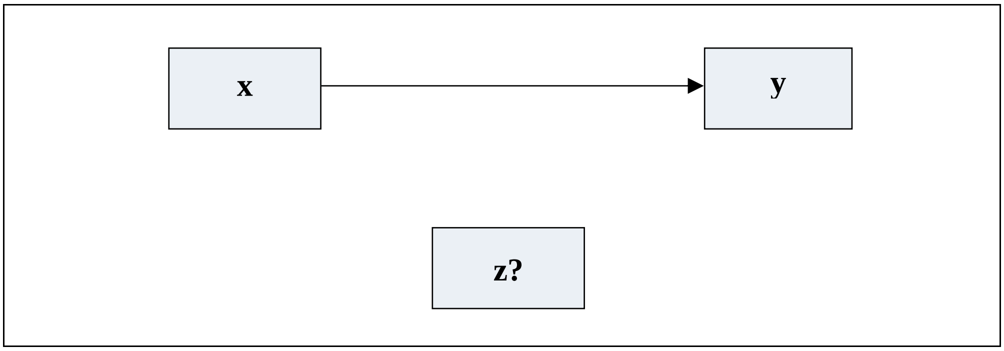


If you read about a variable being “independent”, an “exposure”, or a “predictor” – what does that mean? Basically, it means that someone thinks that this variable has an (statistical) effect on another variable. For the sake of simplicity, let us just call this type of variable “x”. The other variable – the one that x is assumed to affect – is called “dependent” variable or “outcome”. Again, to make it simpler, we can call it “y”.

Examples
Smoking (x) -> Lung cancer (y) Unemployment (x) -> Low income (y) Yoga lessons (x) -> Lower stress levels (y)

The examples presented above may suggest that it is easy to know which variable is x and which is y, but this is not always the case. Sometimes the situation is more complex. As an example, let us take the association between health and educational attainment: does a lower educational attainment (x) lead to worse health (y) or does poor health (x) result in lower educational attainment (y)? These kinds of issues are sometimes discussed in terms of “direction of causality” (again, see Section 9.4 for a more thorough discussion about causality). In cases like that you need to think about what is more reasonable: what does the previous literature/theory say about the association? Preferably, we would want to design the study in a way that solves the issue of directionality.

9.3 Z: confounding, mediating and moderating variables



The association – between x and y – that we are most interested in is often called “main association”. This is the focus of our analysis. However, sometimes there are other variables that we might find important for this main association. Strictly speaking, those variables are also called “x” (or “covariates”) but for clarity we will label them “z”. There are three important types of z-variables that are common in data analysis:

Types of “z”	
Confounder	Both x and y are affected by z
Mediator	A part of the association between x and y goes through z
Moderator	Z affects the association between x and y

Temporality

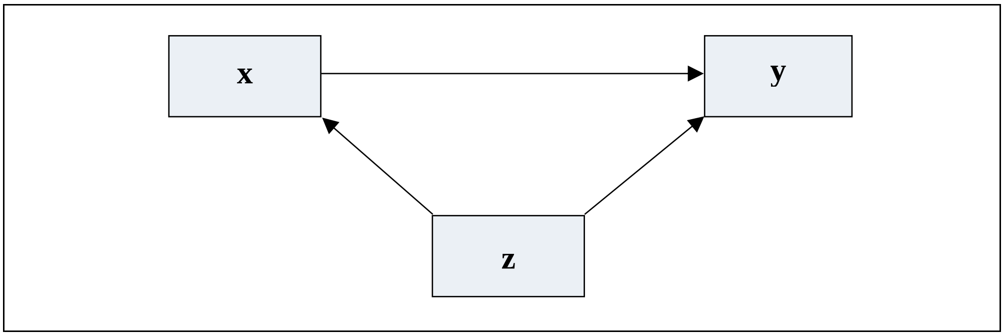
One thing that is good to keep in mind is the concept of temporality, i.e. timing. When did the phenomena that we are examining happen? This is often the same thing as the time at which the phenomenon was measured – but not always. For example, sometimes data are collected retrospectively (as with case-control studies and retrospective cohort studies), and sometimes survey questions are retrospective (e.g. asking adult respondents about childhood conditions).

Either way, it is important to ascertain the following:

- Nothing can predict an outcome if it happens after the outcome. In other words, whatever x-variable you have, it must measure something that happened *before* the phenomenon that your y-variable is thought to capture.
- A confounder is something that is believed to affect the x-variable and the y-variable. Therefore, it cannot measure something that happened *after* the x-variable and/or *after* the y-variable.
- A mediator is something that is believed to be affected by the x-variable as well as affect the y-variable. This means that is cannot come *before* the x-variable or *after* the y-variable.
- A mediator is something that is thought to affect the effect of the x-variable on the y-variable. It cannot come *after* the y-variable.

The details might still be a bit murky at this point, but we will return to all of them later in this chapter.

9.3.1 Confounding variables



A confounder is a variable that influences both the x-variable and the y-variable and, therefore, makes you think that there is an actual relationship between x and y (but it is due to z). Put differently, the confounder distorts the analysis. Suppose that we find that people who consume a lot of coffee (x) have an increased risk of lung cancer (y). A probable confounder could be cigarette smoking (z): smokers drink more coffee and have greater risk of lung cancer.

One should always worry about confounding in research, both when we conduct our own research and when we review others' research.

Address confounding by study design

If you are about to collect your own data, there are many ways to design a study to reduce potential confounding (see Section 3.1). The most obvious solution might be to do an experimental study (e.g. a randomised controlled trial; RCT). Experimental studies are, however, not always feasible, and most of the time, we do observational studies (e.g. cohort studies or case-control studies). Here, it is necessary to review the scientific literature, and then make sure to collect data on all potential confounders.

Address confounding in statistical analysis

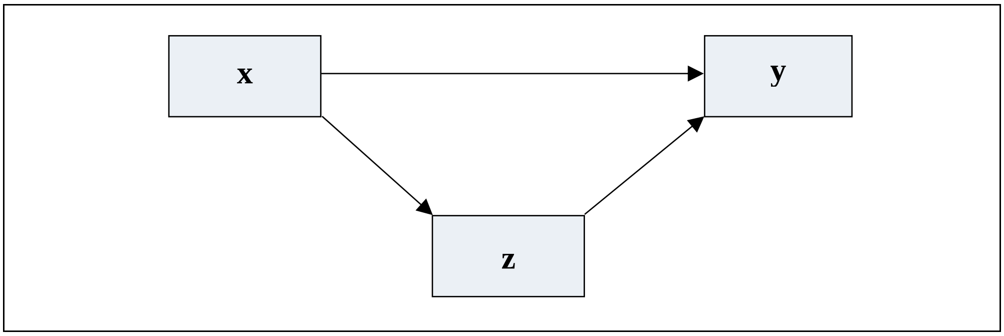
Usually, we work with data that have already been collected – and perhaps for other purposes than what we are interested in. At this stage, you can explore multiple regression analysis with adjustment for confounding, as well as try out stratified analysis and interaction analysis (see Chapter 18). Make sure to adjust for confounding, the best that you can.

Address confounding with specific methods

In addition, there are some specific statistical methods that can be used to handle confounding, such as propensity score matching. This will not be covered in this guide, but if you are interested, we recommend that you explore this further:

More information	help teffects psmatch

9.3.2 Mediating variables



A mediator is a variable that is influenced by the x-variable and influences the y-variable. In other words, some (it could be a little or a lot) of the effect of x on y is mediated through z. For example, let us say that we are interested in the association between parents' educational attainment (x) and children's success on the labour market (y). It could be reasonable to assume that the educational attainment of the parents (x) influences children's own educational attainment (z), which in turn affects their following success on the labour market (y).

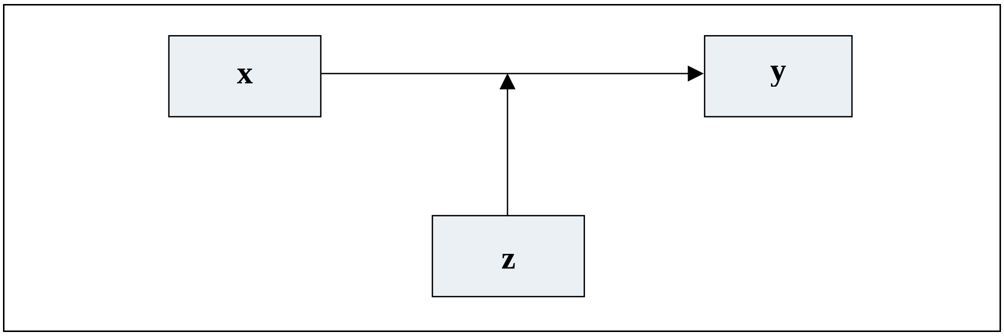
Pathways and mechanisms

In data analysis, we often talk about "explaining" an association by the inclusion of certain mediating variables. Particularly when one has a data material that consists of information collected across several points in time (i.e. longitudinal or life course data), it is common to talk about mediation as "pathways" or "mechanisms".

Mediation analysis

Traditionally, mediators have been treated similarly to confounders in multiple regression analysis. This means that one includes one or more mediators in the model and see how much is explained of the association that we are interested in. This approach has been heavily criticised in the context of *non-linear* regression models (for statistical reasons that we will not discuss here). There are some specific types of mediation analysis that can be used; one of them is the KHB method, which will be explored in Chapter 18.

9.3.3 Moderating (or effect modifying) variables



A moderator (or effect modifier) is a variable that influences the very association between the x-variable and the y-variable. Thus, the association between x and y looks different depending on the value of z. Suppose that we are interested in the association between unemployment (x) and ill-health (y). Here, it could be reasonable to assume that men’s and women’s health is affected differently by unemployment – in that case, gender (sex) would be a moderating variable (z).

Interaction analysis

In data analysis, moderating variables are examined through something called interaction analysis (see Chapter 19).

9.4 A note on causal inference

Earlier in this chapter, we suggested that it is common to focus on associations in quantitative research, and that a statistical effect of one variable on another variable is not the same as a causal effect. Yet, we use concepts throughout the guide that sort of imply causality, such as "exposure", "outcome", "mediator", and "pathway". In this section, we will discuss the issue of causality – or causal inference – in a bit greater depth. Of course, there is not enough space here to address the full complexity of the issue.

Data analysis is often concerned with causal questions. For example, can a given intervention program improve program participant's outcomes? Can a given sickness be prevented? Why do girls typically outperform boys in the educational system? In contrast to statistical inference, in which information obtained from various forms of random sample of observations are used to draw conclusions about the value of some parameter (e.g. a mean or a regression coefficient) in the population from which the sample was drawn (see Chapter 3), causal inference typically refers to the process where multiple sources of information are used to draw reasonable inference about cause and effect.

Causal inference taps into important discussions related to ontology and epistemology, which will not be addressed here. For the purposes of this guide, three broadly defined (and partly overlapping) perspectives on causal inference will be outlined (based on Goldthorpe's "Causation, statistics, and sociology" from 2001), all of which to some extent may relate to the empirical methodologies detailed above: causation as robust association, causation as manipulation, and causation as generative mechanisms.

While it remains true that the widely recognized statement that association (i.e. correlation) does not imply causation, causation must in some way imply association. Causation as robust association comes in many versions but a common denominator is that it emphasizes efforts to ensure that estimated associations are not spurious, i.e. the association cannot be eliminated through one or more other variables being introduced in the analysis. In practice, this approach typically proposes a set of criteria such as temporality and predictive power to assess causal connections between variables.

Causation as manipulation appears to some extent to have emerged in reaction to that of causation as robust association. Here, attention centres on establishing causation through experimental methods. In short, the key idea is that causes can only be those factors that, at least theoretically, can serve as treatments (or more generally exposures) in experiments. This means that causes must in some sense be manipulable, and that causation is determined by comparing what would happened to an observational unit in regard to an outcome if this unit have been vs. not have been exposed to the addressed factor. Since it is not possible for the same unit to be both exposed and not exposed, the solution for estimating a causal effect is to compare the average response for those units exposed to the average response for those units that were not exposed. For this solution to be viable, however, a number of conditions and assumptions need to be met. These conditions are ideally those of randomised controlled trials, but substantial efforts have been made to develop statistical analyses that to a large extent mimic the conditions of such trials (e.g. propensity score matching, endogenous treatment effect regression). Causation as manipulation also comes in different versions and the main difference lies in to what extent there is an emphasis on designing a study or on analysis of already collected data (cf. the potential outcome framework). While the former focusses on removing threats to internal validity by using appropriate experimental designs, the latter focusses on strategies for estimating causal effects using observational data (often in a longitudinal design).

In contrast to the above, causation as generative mechanisms does not focus on relationships between variables but rather on what needs to be added to any criteria before a reasonable argument for causation can be made, namely the agentic capabilities of observational units (typically individuals). Here, actors, their relationships, and the (un)intended outcomes of their actions are emphasized. The properties of actors and their environments can be measured and thereby represented by variables, but the causality does not operate at the variable level. According to this perspective, the actors are the agents of change and the causal process should therefore be specified at the actor level which means that in order to move from association to causation it is not sufficient to just establishing that a given factor precedes the outcome (rather than the other way around), it is also necessary to specify the mechanisms that explain why actors do what they do and how these actions translate into outcomes. In order to do so, proponents of this approach typically suggest that various theories of rational action can be utilized.

Research questions often relates to issues of causality and the perspectives outlined above all have their pros and cons. In any event, estimates of associations alone cannot be used for causal inference. Various sources of information, which includes a theoretical framework of causality combined with the best available research design and data for the research question at hand, are imperative in the process of evaluating whether our estimates may allow for reasonable causal interpretations.

10. (M)AN(C)OVA

Outline

Content

In this chapter, we will discuss and partly also explore the different extensions of ANOVA that are available, including ANCOVA, MANOVA, and MANCOVA.

What do these terms mean?	
ANOVA	Analysis of variance
ANCOVA	Analysis of covariance
MANOVA	Multivariate analysis of variance
MANCOVA	Multivariate analysis of covariance

10.1 ANCOVA

Quick facts	
Number of variables	One group variable (x) One test variable (y) One or more covariates (z)
Scale of variable(s)	Group variable: categorical (nominal/ordinal) Test variable: continuous (ratio/interval) Covariates: categorical (nominal/ordinal) or continuous (ratio/interval)

As discussed in Section 6.4, the one-way ANOVA is the statistical procedure of comparing the means of two or more groups. ANCOVA is very similar to ANOVA. The key difference is that ANCOVA allows you to control for the effects of one or more extraneous variables, known as covariates (also see the discussion on confounding in Chapter 9). These covariates can take any form, i.e. they can be either categorical or continuous – but if you have a non-binary categorical covariate (i.e. one with more than two categories) you need to create dummy variables for this one (see Section 11.2.1).

For example, you could use ANCOVA to see which diet was best for losing weight after controlling for age and body mass index at baseline (i.e. your test variable would be “weight loss”, your group variable would be “type of diet” and your covariates would be “age” and “body mass index at baseline”).

Note In many ways, ANCOVA is equivalent to multiple linear regression (which is described in more detail in Chapter 12). Why then use ANCOVA? Well, the answer depends on what you want to achieve with your analysis, but for most purposes, we would argue that linear regression is more flexible.

Assumptions

First, you have to check your data to see that the assumptions behind ANCOVA hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Continuous and normally distributed test variable	Your test variable should be continuous (i.e. interval/ratio) and normally distributed. For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: "How satisfied with your income are you?" (on a scale 1-10) or "To what extent do you agree with the previous statement?" (on a scale 1-5).
Two or more unrelated categories in the group variable	Your group variable should be categorical (i.e. nominal or ordinal) and consist of two or more groups. Unrelated means that the groups should be mutually excluded: no individual can be in more than one of the groups. For example: low vs. medium vs. high educational level; liberal vs. conservative vs. socialist political views; or poor vs. fair, vs. good vs. excellent health; and so on.
Equal variance	The variance in the test variable should be equal across the groups of the group variable.
No outliers	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.
Homogenetiy of regression slopes	Your test variables and any covariate(s) should have the same slopes across all levels of the categorical group variable.

Function

Basic command	anova testvar groupvar c.covariate	
Explanations	testvar groupvar covariate	Insert the name of the test variable Insert the name of the group variable. Insert the name of the covariate variable
Notes	You need to tell Stata that a variable in your ANOVA statement is continuous or it will treat it as another categorical factor. You denote continuous independent variables within the ANOVA command by placing "c." in front of it.	
More information	help anova	

Practical example

Dataset: StataData1.dta

Name	Label
gpa	Grade point average (Age 15, Year 1985)
bullied	Exposed to bullying (Age 15, Year 1985)
cognitive	Cognitive test scores (Age 15, Year 1985)

```
anova gpa bullied c.cognitive
```

```
           Number of obs =      8,192    R-squared     =  0.4032
           Root MSE      =    .532934    Adj R-squared =  0.3959
    Source | Partial SS         df         MS        F    Prob>F
-----------+----------------------------------------------------
     Model |   1552.841        100    15.52841     54.67  0.0000
           |
   bullied |  4.8369933          1   4.8369933     17.03  0.0000
 cognitive |  1514.3779         99   15.296746     53.86  0.0000
           |
  Residual |  2297.9933      8,091   .28401845
-----------+----------------------------------------------------
     Total |  3850.8343      8,191   .47012993
```

In this example, we are interested in seeing if grade point average (gpa) differs between individuals according to whether they have been exposed to bullying or not (bullied), while controlling for cognitive test scores (cognitive). The null hypothesis is that there is no difference in gpa between unexposed and exposed.

Note Partial SS (SS=sum of squares) refers to variation assigned to one variable while controlling for the other variable.

As can be seen, the F statistic for bullied is 17.03. The corresponding p-value is 0.0000. Since this is below 0.05, it means that there is a statistically significant difference in grade point average between unexposed and exposed to bullying when we control for cognitive test scores. In other words, we can reject the null hypothesis. We can also see that the variable cognitive is statistically significantly related to grade point average (F=53.86, p <0.05).

Postestimation commands

There are many different postestimation commands that you can apply to ANCOVA.

More information	help anova postestimation

For example, we can use the postestimation command contrast to obtain the adjusted mean differences:

```
contrast r.bullied, asobserved
```

```
Contrasts of marginal linear predictions

Margins      : asobserved

------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     bullied |          1        6.19     0.0129
             |
 Denominator |       8189
------------------------------------------------

--------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     bullied |
(Yes vs No)  |  -.0487196   .0195891     -.0871192   -.0103199
--------------------------------------------------------------
```

In the column called Contrast, you see the mean difference (-0.0487) in grade point average between those who were exposed to bullying and those who were not exposed, controlled for cognitive test scores.

We can also use the postestimation command margins, which gives us predicted means for each of the group:

```
margins bullied
```

```
Predictive margins                              Number of obs     =      8,192

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bullied |
         No  |   3.227265   .0062901   513.07   0.000     3.214934    3.239595
        Yes  |   3.178545   .0185206   171.62   0.000      3.14224     3.21485
------------------------------------------------------------------------------
```

Looking at the column called Margin, we see that the predicted mean in grade point average is slightly higher for individuals who were not exposed to bullying (3.227) than for individuals who were exposed to bullying (3.179), controlled for cognitive test scores. Also note that the difference in means between the groups is around 0.0487, which is what we saw with contrast.

10.2 MANOVA

Quick facts	
Number of variables	One group variable (x) Two or more test variables (y)
Scale of variable(s)	Group variable: categorical (nominal/ordinal) Test variables: continuous (ratio/interval)

Like ANOVA, MANOVA is used to test the significance of group differences. However, MANOVA can include several dependent variables, whereas ANOVA can handle only one dependent variable.

For example, you could use a MANOVA to investigate whether salary income and number of weekly work hours differ according to age categories (i.e. your test variables would be "income" and "number of weekly work hours", while "age category" would be your group variable). Alternatively, you could use a MANOVA to investigate whether math and science performance differ based on test anxiety levels amongst students (i.e. your test variables would be "math test score" and "science test score", while your group variable would be "test anxiety level").

Note MANOVA can be seen as a combination of ANOVA and two or more t-tests. Accordingly, advantages are that you can compare more than two groups *and* that the test variables are mutually adjusted for.

Assumptions

First, you have to check your data to see that the assumptions behind MANOVA hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Continuous and normally distributed test variables	Your test variables should be continuous (i.e. interval/ratio) and normally distributed. For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: "How satisfied with your income are you?" (on a scale 1-10) or "To what extent do you agree with the previous statement?" (on a scale 1-5).
Two or more unrelated categories in the group variable	Your group variable should be categorical (i.e. nominal or ordinal) and consist of two or more groups. Unrelated means that the groups should be mutually excluded: no individual can be in more than one of the groups. For example: low vs. medium vs. high educational level; liberal vs. conservative vs. socialist political views; or poor vs. fair, vs. good vs. excellent health; and so on.
Equal variance	The variance in the test variables should be equal across the groups of the group variable.
No outliers	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.
Absence of multicollinearity	Your test variables should not be too correlated. A good rule of thumb is that no correlation should be above r = 0.90.

Function

Basic command	manova testvars = groupvar	
Explanations	testvars	Insert the name of the test variables.
	groupvar	Insert the name of the group variable.
More information	help manova	

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test scores (Age 15, Year 1985)
skipped	Skipped class (Age 15, Year 1985)

```
manova gpa cognitive = skipped
```

```
                Number of obs =      8,689
                W = Wilks' lambda      L = Lawley-Hotelling trace
                P = Pillai's trace     R = Roy's largest root

       Source | Statistic        df     F(df1,     df2) =   F   Prob>F
  ------------+-----------------------------------------------------
      skipped |W   0.9476         2       4.0  17370.0   118.53 0.0000 e
              |P   0.0524                 4.0  17372.0   116.93 0.0000 a
              |L   0.0553                 4.0  17368.0   120.13 0.0000 a
              |R   0.0553                 2.0   8686.0   240.11 0.0000 u
              |-----------------------------------------------------
     Residual |               8686
  ------------+-----------------------------------------------------
        Total |               8688
   ------------------------------------------------------------------
                e = exact, a = approximate, u = upper bound on F
```

In this example, we investigate whether grade point average and cognitive test scores differ between those who never, sometimes, and often have skipped class. Our null hypothesis is that there is no difference.

Stata provides four test statistics by default (listed above the table). The most commonly used criterion is Wilks' Lambda and this is what will be used in this example. Thus, we need to consult the Prob>F column along the Wilks' Lambda (W) row to determine whether the null hypothesis should be rejected.

As can be seen, the F statistic is 118.53. The corresponding p-value is 0.0000 (i.e. below 0.05), meaning that there is statistically significant difference in both grade point average and cognitive test scores between individuals who have never, sometimes, and often skipped class. In other words, we can reject the null hypothesis.

Postestimation commands

There are many different postestimation commands that you can apply to MANOVA.

More information	help manova postestimation

For example, it is probably relevant to obtain the adjusted mean differences between the groups. We can use the postestimation command contrast to achieve this.

First, we can ask for the mean differences in gpa:

```
contrast r.skilled, equation(gpa)
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

---------------------------------------------------------
                      |         df           F        P>F
----------------------+----------------------------------
gpa                   |
              skipped |
(Sometimes vs Never)  |          1      157.68     0.0000
    (Often vs Never)  |          1      281.74     0.0000
               Joint  |          2      167.03     0.0000
                      |
          Denominator |       8686
---------------------------------------------------------

-----------------------------------------------------------------------
                      |   Contrast   Std. Err.     [95% Conf. Interval]
----------------------+------------------------------------------------
gpa                   |
              skipped |
(Sometimes vs Never)  |  -.1975808    .0157347     -.2284246    -.166737
    (Often vs Never)  |  -.3831545    .0228272     -.4279012   -.3384079
-----------------------------------------------------------------------
```

In the column called Contrast, we see that the mean difference in grade point average between those who sometimes have skipped class and those who have never skipped class is -0.198. The mean difference between those who often have skipped class and those who have never skipped class is -0.383.

And then we can obtain the mean differences in cognitive:

```
contrast r.skilled, equation(cognitive)
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

---------------------------------------------------------
                      |         df           F        P>F
----------------------+----------------------------------
cognitive             |
              skipped |
(Sometimes vs Never)  |          1        3.96     0.0467
    (Often vs Never)  |          1        3.43     0.0641
               Joint  |          2        2.75     0.0638
                      |
          Denominator |       8686
---------------------------------------------------------

-----------------------------------------------------------------------
                      |   Contrast   Std. Err.     [95% Conf. Interval]
----------------------+------------------------------------------------
cognitive             |
              skipped |
(Sometimes vs Never)  |  -3.303194   1.660627     -6.558417   -.0479699
    (Often vs Never)  |  -4.460767   2.409153     -9.183278    .2617446
-----------------------------------------------------------------------
```

Here, we see that the mean difference in cognitive test scores between those who sometimes have skipped class versus those who have never skipped class is -3.303. The mean difference between those who often have skipped class versus those who have never skipped class is -4.460.

We can also use the postestimation command margins, which gives us predicted means for each of the groups.

First, we get the predicted means in gpa:

```
margins skipped, predict(equation(gpa))
```

```
Adjusted predictions                            Number of obs     =      8,689

Expression   : Linear prediction, predict(equation(gpa))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     skipped |
       Never |    3.32965   .0111047   299.84   0.000     3.307882    3.351418
   Sometimes |   3.132069   .0111476   280.96   0.000     3.110218    3.153921
       Often |   2.946496   .0199441   147.74   0.000     2.907401    2.985591
------------------------------------------------------------------------------
```

Looking at the column called Margin, we see that the predicted mean in grade point average for individuals who have never skipped class (3.330) is higher than those for individuals who sometimes (3.132), and often have skipped class (2.946), thus confirming what we got with contrast.

And then we can get the predicted means in cognitive:

```
margins skipped, predict(equation(cognitive))
```

```
Adjusted predictions                            Number of obs     =      8,689

Expression   : Linear prediction, predict(equation(cognitive))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     skipped |
       Never |   310.9497   1.171974   265.32   0.000     308.6523     313.247
   Sometimes |   307.6465   1.176503   261.49   0.000     305.3402    309.9527
       Often |   306.4889   2.104874   145.61   0.000     302.3628    310.6149
------------------------------------------------------------------------------
```

Here, we see that the predicted mean in cognitive test scores for individuals who have never skipped class (310.950) is higher than those for individuals who sometimes (307.647), and often have skipped class (306.489), thus confirming what we got with contrast.

10.3 MANCOVA

Quick facts	
Number of variables	One group variable (x) Two test variables (y) One or more covariates (z)
Scale of variable(s)	Group variable: categorical (nominal/ordinal) Test variables: continuous (ratio/interval) Covariates: categorical (nominal/ordinal) or continuous (ratio/interval)

MANCOVA is similar to MANOVA, with the key difference that allows you to control for the effects of one or more extraneous variables, known as covariates (also see the discussion on confounding in Chapter 9). These covariates can take any form, i.e. they can be either categorical or continuous – but if you have a non-binary categorical covariate (i.e. one with more than two categories) you need to create dummy variables for this one (see Section 11.2.1).

For example, you could use a MANCOVA to evaluate the effectiveness of online learning compared to traditional learning on math and reading scores, while controlling for pre-test scores (i.e. your test variables would be “math scores” and “reading scores”, your group variable would be “learning environment” and your covariates would be “pre-test scores”). Alternatively, you could use a MANCOVA to investigate the effect of different sport drinks on athletic performance (as measured by heart rate, blood pressure and blood electrolytes), while controlling for age (i.e. your test variables would be “heart rate”, blood pressure” and “blood electrolytes”, your group variable would be “type of sport drink” and your covariate would be “age”).

Assumptions

First, you have to check your data to see that the assumptions behind MANCOVA hold. If your data "passes" these assumptions, you will have a valid result.

Checklist	
Continuous and normally distributed test variables	Your test variables should be continuous (i.e. interval/ratio) and normally distributed. For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: "How satisfied with your income are you?" (on a scale 1-10) or "To what extent do you agree with the previous statement?" (on a scale 1-5).
Two or more unrelated categories in the group variable	Your group variable should be categorical (i.e. nominal or ordinal) and consist of two or more groups. Unrelated means that the groups should be mutually excluded: no individual can be in more than one of the groups. For example: low vs. medium vs. high educational level; liberal vs. conservative vs. socialist political views; or poor vs. fair, vs. good vs. excellent health; and so on.
Equal variance	The variance in the test variables should be equal across the groups of the group variable.
No outliers	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.
Homogenetiy of regression slopes	Your test variables and any covariate(s) should have the same slopes across all levels of the categorical group variable.
Absence of multicollinearity	Your test variables should not be too correlated to each other. A good rule of thumb is that no correlation should be above r = 0.90.

Function

Basic command	manova testvar testvar = groupvar c.covariate	
Explanations	testvar groupvar covariate	Insert the name of the test variable Insert the name of the group variable. Insert the name of the covariate variable
Notes	You need to tell Stata that a variable in your MANCOVA statement is continuous or it will treat it as another categorical factor. You denote continuous independent variables within the MANCOVA command by placing "c." in front of them.	
More information	help manova	

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test scores (Age 15, Year 1985)
skipped	Skipped class (Age 15, Year 1985)
sex	Sex

```
manova gpa cognitive = skipped sex
```

```
             Number of obs =      8,689

             W = Wilks' lambda      L = Lawley-Hotelling trace
             P = Pillai's trace     R = Roy's largest root

    Source | Statistic        df    F(df1,     df2) =   F   Prob>F
-----------+-------------------------------------------------------
     Model |W   0.9171         3      6.0  17368.0   127.95 0.0000 e
           |P   0.0832                 6.0  17370.0   125.71 0.0000 a
           |L   0.0900                 6.0  17366.0   130.20 0.0000 a
           |R   0.0854                 3.0   8685.0   247.17 0.0000 u
           |-------------------------------------------------------
  Residual |                8685
-----------+-------------------------------------------------------
   skipped |W   0.9749         2      4.0  17368.0    55.52 0.0000 e
           |P   0.0251                 4.0  17370.0    55.19 0.0000 a
           |L   0.0257                 4.0  17366.0    55.86 0.0000 a
           |R   0.0255                 2.0   8685.0   110.73 0.0000 u
           |-------------------------------------------------------
       sex |W   0.9679         1      2.0   8684.0   144.10 0.0000 e
           |P   0.0321                 2.0   8684.0   144.10 0.0000 e
           |L   0.0332                 2.0   8684.0   144.10 0.0000 e
           |R   0.0332                 2.0   8684.0   144.10 0.0000 e
           |-------------------------------------------------------
  Residual |                8685
-----------+-------------------------------------------------------
     Total |                8688
-------------------------------------------------------------------
             e = exact, a = approximate, u = upper bound on F
```

Here, we extend the analysis from the previous section on MANOVA by adding sex as a covariate to the model formula, which gives us a MANCOVA. Specifically, we investigate whether there are differences in grade point average and cognitive test scores across the levels of skipped, while controlling for sex. Again, our null hypothesis is that there are no differences.

The F statistic (based on the Wilks' lambda) for skipped is F=55.52. The corresponding p-value is 0.0000 (i.e. below 0.05), which allows us to reject the null hypothesis. We also see that there are statistically significant differences in gpa and cognitive between men (boys) and women (girls) (F=144.10, p <0.05).

Postestimation commands

There are many different postestimation commands that you can apply to MANCOVA.

More information	help manova postestimation

For example, we might want to obtain the mean differences between the groups. We can use the postestimation command contrast to achieve this.

First, we get the mean differences in gpa:

```
contrast r.skilled, equation(gpa)
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

---------------------------------------------------------
                      |         df           F        P>F
----------------------+----------------------------------
gpa                   |
              skipped |
(Sometimes vs Never)  |          1       86.20     0.0000
    (Often vs Never)  |          1      169.82     0.0000
               Joint  |          2       93.93     0.0000
                      |
          Denominator |       8685
---------------------------------------------------------

-----------------------------------------------------------------------
                      |   Contrast   Std. Err.     [95% Conf. Interval]
----------------------+------------------------------------------------
gpa                   |
              skipped |
(Sometimes vs Never)  |  -.1538159   .0165675      -.186292   -.1213397
    (Often vs Never)  |  -.3154568   .0242069     -.3629081   -.2680055
-----------------------------------------------------------------------
```

In the column called Contrast, we see that the mean difference in grade point average between those who sometimes have skipped class and those who have never skipped class is -0.154, controlled for sex. The mean difference between those who often have skipped class and those who have never skipped class is -0.315, controlled for sex.

And then we can obtain the mean differences in cognitive:

```
contrast r.skilled, equation(cognitive)
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

---------------------------------------------------------
                      |         df           F        P>F
----------------------+----------------------------------
cognitive             |
              skipped |
(Sometimes vs Never)  |          1       14.96     0.0001
    (Often vs Never)  |          1       14.75     0.0001
               Joint  |          2       10.49     0.0000
                      |
          Denominator |       8685
---------------------------------------------------------

-----------------------------------------------------------------------
                      |   Contrast   Std. Err.     [95% Conf. Interval]
----------------------+------------------------------------------------
cognitive             |
              skipped |
(Sometimes vs Never)  |  -6.773942   1.751427     -10.20715   -3.340729
    (Often vs Never)  |  -9.829491   2.559032      -14.8458    -4.81318
-----------------------------------------------------------------------
```

Here, we see that the mean difference in cognitive test scores between those who sometimes have skipped class and those who have never skipped class is -6.774, controlled for sex. The mean difference between those who often have skipped class and those who have never skipped class is -9.829, controlled for sex.

We can also use the postestimation command margins, which gives us predicted means for each of the groups.

First, we get the predicted means in gpa:

```
margins skipped, predict(equation(gpa))
```

```
Predictive margins                              Number of obs     =      8,689

Expression   : Linear prediction, predict(equation(gpa))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     skipped |
      Never  |   3.301672   .0115819   285.07   0.000     3.278968    3.324375
  Sometimes  |   3.147856   .0112729   279.24   0.000     3.125758    3.169953
      Often  |   2.986215   .0204565   145.98   0.000     2.946115    3.026314
------------------------------------------------------------------------------
```

Looking at the column called Margin, we see that the predicted mean in grade point average for individuals who have never skipped class (3.302) is higher than those for individuals who sometimes (3.148), and often have skipped class (2.986), controlled for sex (thus confirming what we got with contrast).

And then we get the predicted means in cognitive:

```
margins skipped, predict(equation(cognitive))
```

```
Predictive margins                              Number of obs     =      8,689

Expression   : Linear prediction, predict(equation(cognitive))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     skipped |
      Never  |   313.1685   1.224382   255.78   0.000     310.7684    315.5686
  Sometimes  |   306.3945   1.191713   257.10   0.000     304.0585    308.7306
      Often  |    303.339   2.162555   140.27   0.000     299.0999    307.5781
------------------------------------------------------------------------------
```

Here, we see that the predicted mean in cognitive test scores for individuals who have never skipped class (313.169) is higher than those for individuals who sometimes (306.395), and often have skipped class (303.339), controlled for sex (thus confirming what we got with contrast).

11. PREPARATIONS FOR REGRESSION ANALYSIS

Outline

Content

Once we have produced suitable descriptive statistics for our data, we often want to move on to more advanced types of analysis, such as regression analysis. Before this, we nonetheless need to consider what kind of analysis that might be suitable, how to design our analysis, and how to deal with missing data. These issues are addressed in the current chapter.

11.1 What type of regression should be used?

There are many different types of regression analysis. Some of the most common types are included in this guide: linear, logistic, ordinal, multinomial, Poisson, and Cox regression. Which one you should choose depends on your outcome (y).

Outcome (y)	**Type of regression**
Continuous (ratio/interval)	**Linear regression**
Nominal with two categories, i.e. binary	**Logistic regression**
Nominal with more than two categories, i.e. non-binary	**Multinomial regression**
Ordinal	**Ordinal regression**
Count	**Poisson regression**
Time-to-event	**Cox regression**

However, your x-variable(s) can take on any form – they can be categorical (i.e. nominal/ordinal) or continuous (i.e. ratio/interval). If you include only one x-variable in your regression analysis, this is called simple (or bivariate) regression analysis. If you include two or more x-variables in your regression analysis, this is called multiple regression analysis. In multiple regression analysis, it is possible to mix different types of x-variables: you can thus use both categorical and continuous x-variables.

11.2 Dummies

When we conduct regression analysis – regardless of the type – we can only analyse x-variables that are continuous (ratio/interval) or binary (i.e. they consist of only two values). A binary variable is sometimes called “dichotomous”, “binomial” or “dummy”. If we have a categorical variable with more than two values, such as in the example below, we need to “trick” the regression analysis to correctly analyse those variables. To do this, there are two alternatives in Stata. These will be presented below.

11.2.1 Dummy variables

The first alternative is to manually create one dummy for each category of the variable.

Example		
Variable	*Categories*	*Dummy*
Educational attainment	1=Compulsory	1=Compulsory, 0=Other
	2=Upper secondary	1=Upper secondary, 0=Other
	3=University	1=University, 0=Other

For example:

```
gen educ_comp=educ

gen educ_uppsec=educ

gen educ_uni=educ

recode educ_comp (1=1) (2=0) (3=0)

recode educ_uppsec (1=0) (2=1) (3=0)

recode educ_uni (1=0) (2=0) (3=1)
```

In the regression analysis, all dummies for the specific variable should be included as x-variables, *except one*. The dummy that you exclude – and it is your own choice which one you exclude – will be the “reference category”. Each of the other dummies will be compared to the dummy that is excluded.

11.2.2 Factor variables

The alternative presented above is quite pedagogical (we think, at least) – but it is quite time consuming to generate dummies. In Stata, there is an easy fix: something called factor variables. When you conduct your regression analysis, you just simply write the prefix "i." before the name of the categorical variable(s), and Stata will include dummies in the analysis automatically, e.g. "i.educ"

Base level

When you include factor variables, the lowest value will automatically be chosen as the reference category. This can be altered by specifying another so-called base level. This is done by adding a "b" to the prefix: "ib." You then also need to specify which category that should be the reference category by adding the value of the category, e.g. "ib3.educ" (which would define "University" as the reference category).

There are also some alternatives to specifying the value of the category. These are the possible so-called base operators:

Base operator	**Explanation**
ib#.	Specifies a specific value as the base. #=the value of the category that we want to choose as the reference category.
ib##.	Specifies the #th ordered value as the base.
ib(first).	Specifies the smallest value as the base. Default.
ib(last).	Specifies the largest value as the base.
ib(freq).	Specifies the most common value as the base.
ibn.	No base level.

11.2.3 A note on the choice of reference category

There are many different ways of choosing a reference category:

Choosing a reference category
The largest category, because we want a stable group to compare the other categories to.
The group in the middle, to represent the average.
The "best off" category – if increasing values of the outcome is more negative.
The "worst off" category – if increasing values of the outcome is more positive.

Note Never choose a very small category – you may end up with very strange estimates.

11.3 Analytical strategy

Regression analysis is of course about data, but it is also about design. The way in which you think your variables are related needs to be translated into an analytical strategy (or modelling strategy). A good way to start is to make a drawing with boxes and arrows: each variable is put into one box and then you put simple-headed or double-headed arrows between the boxes to illustrate how the variables are associated to one another. Remember that the analytical strategy should reflect the aim of the study.

Example

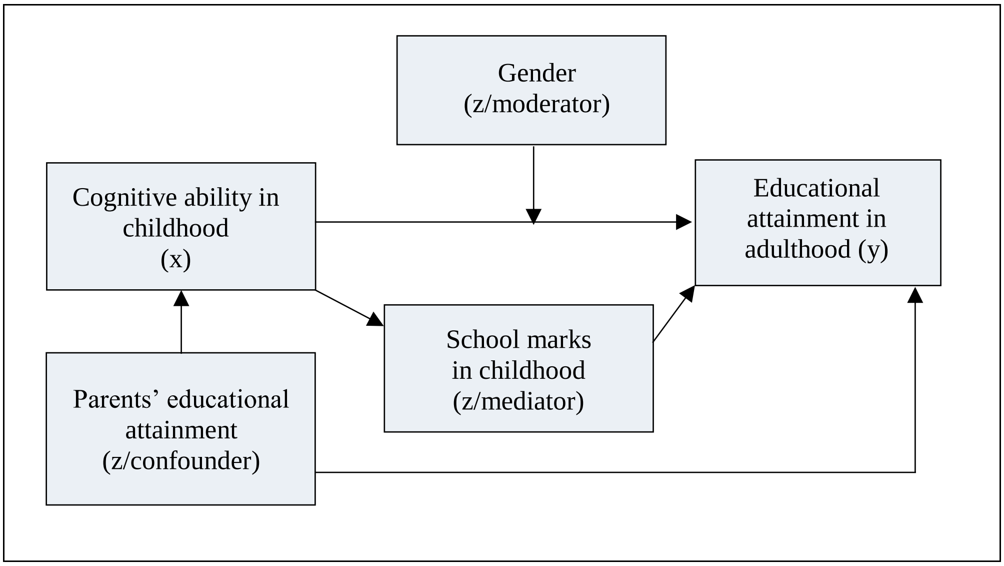
Suppose we are interested in the association between children's cognitive ability and educational attainment in adulthood. To examine this association is thus the aim of the study. We think that this association may be confounded by parents' educational attainment and mediated by children's school marks. Moreover, we suspect that the association may look different depending on the child's gender. The research questions (RQs) can thus be formulated as:

- RQ1. Is children's cognitive ability associated with educational attainment in adulthood?
- RQ2. If so, is this association confounded by parents' educational attainment?
- RQ3. To what extent is the association between children's cognitive ability and educational attainment in adulthood mediated by school marks in childhood?
- RQ4. Is there any gender difference in the association between children's cognitive ability and educational attainment in adulthood?

Accordingly, these are the variables we need to include in our analysis:

Role	Variable	Scale
x	Cognitive ability in childhood	Ratio
Y	Educational attainment in adulthood	Ordinal
z/confounder	Parents' educational attainment	Ordinal
z/mediator	School marks in childhood	Ratio
z/moderator	Child's gender	Nominal (binary)

And this is how we may choose to illustrate our analytical strategy:


Gender
(z/moderator)
Cognitive ability in childhood
(x)
Educational attainment in adulthood (y)
School marks in childhood
(z/mediator)
Parents' educational attainment
(z/confounder)

Often, we want to break down our analysis in different steps – or models. We want our analysis – as a whole – to answer our research questions.

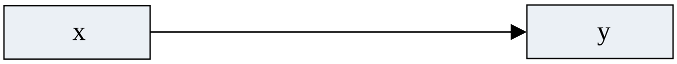
Note that there is no "perfect" way of setting up models. It is often a matter of academic traditions and taste. Some prefer to add variables (confounders, mediators) stepwise, so that each subsequent model becomes more and more complex. Others prefer to do a series of separate models and then finish with "full" model.

We only have some advice:

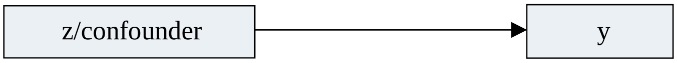
- Always also present an unadjusted analysis for your main association (i.e. simple regression).
- Remember that confounders and mediators play different roles: we are supposed to get rid of the confounding, whereas the mediation could tell us something about possible explanations. In other words, make sure not to mix these up in the analysis (or, in the interpretation and discussion of the results).
- Moderators are a different kind of animal, and are therefore treated and presented in a slightly different way in comparison to confounders and mediators.

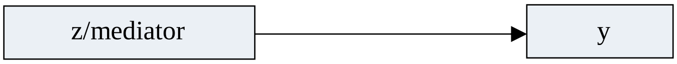
Unadjusted model

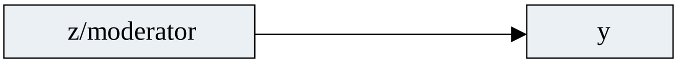
First, start with a simple regression analysis of your main association:



We would also encourage you to do the same for your other variables:



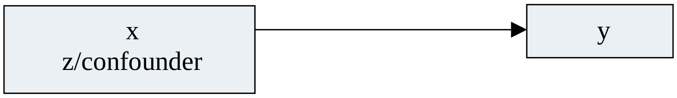




Note If we would have had several confounders, and/or mediators, and/or moderators, these would also have generated their own simple regression model.

Model 1

We continue with multiple regression analysis, by focusing on our main association (x and y) and adding the confounding variable to the model.

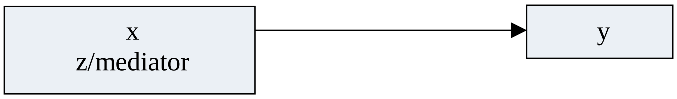


Here, we are interested to see if the estimate(s) for the association between x and y changes when the confounder is added. Does it become weaker (compared to the simple model)?

Note In cases where you have several confounders, you can choose to enter them stepwise one at a time, a few at a time, or all at once. Just remember that if you enter more than one at a time, and you do see a change in the estimate for the association between x and y, you need to check which confounder(s) that might be causing this change.

Model 2

The next step is to add the mediator.



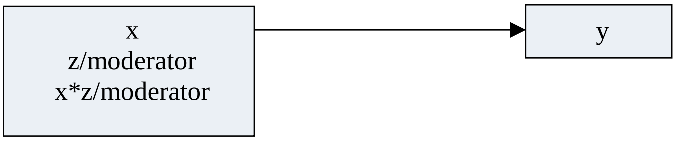
Again, we are interested to see if the estimate(s) for the association between x and y changes when the mediator is added. Does it become weaker (compared to the simple model)?

Note As for cases where you have several mediators: you can choose to enter them stepwise one at a time, a few at a time, or all at once. Just remember that if you enter more than one at a time, and you do see a change in the estimate for the association between x and y, you need to check which mediator(s) that might be causing this change.

Note Remember that this kind of mediation approach might be criticised if you do a *non-linear* (e.g. logistic, ordinal, multinomial, Cox) regression analysis. See Chapter 18 for an alternative approach to mediation analysis.

Model 3

And the final step is to add the moderator. Like it was said earlier, this is more complicated – we will save the details for Chapter 19. But for now, we will just specify this as the following:



11.4 Missing data

As we discussed earlier (see Section 3.2.3), it is common to have missing data. Missing data is sometimes called attrition (particularly in register studies) and sometimes non-response (particularly in survey/questionnaire studies). Missing data can be external or internal:

External or internal?	
External	Occurs when individuals have been sampled from the population but, for various reasons, they do not get included in the register study (they have immigrated, died, moved, are imprisoned, etc.) or do not participate in the survey (they decline, are too sick, cannot be reached, etc.).
Internal	Occurs when individuals who are part of the study, for various reasons (they missed a page of the questionnaire, they refuse to answer specific questions, etc.), have no information for a specific variable or a set or variables.

As shown above, there are many reasons for missing data. If the missingness is problematic or not, depends on what type of missing data we have. In statistical analysis, there are three types of missing data:

Types of missing data	
MCAR	Missing Completely At Random: The probability of missing data is unrelated to both observed and unobserved data; it is completely by chance alone
MAR	Missing At Random: The probability of missing data is unrelated to unobserved data but may be related to observed data
MNAR	Missing Not At Random: The probability of missing data is related to unobserved data

This was probably a bit confusing – let us exemplify the differences between MCAR, MAR and MNAR. Suppose we examine the distribution of income in the Swedish population. If missing data were MCAR, it means that the missingness is unrelated to both observed data (e.g. gender, employment status) and unobserved data (e.g. lower income does not influence the risk of missingness). If missing data were MAR, it would mean that missingness could be related to other variables in the dataset, but the probability of missingness is not increased by certain values of the variable itself (e.g. individuals having lower incomes). Finally, if individuals who had certain values of the variables itself were more likely to be missing, we would have MNAR.

11.4.1 How to deal with missing data?

It is not very easy to statistically address whether missingness is MCAR, MAR or MNAR. The most important advice is that you have to know your data well: produce descriptive statistics for your study variables to see the extent of missingness in the data material. Obviously, if you have a small number of individuals in your data material, a couple of missing values would have more serious consequences than if you have a couple of missing values in a data material based on the total population of a country.

A sound strategy to map out and illustrate potential problems with missingness is first to find out anything you can about the reasons for external attrition. Why are some individuals not included in your dataset? Is it likely that they similar in any important way or is the missingness due to technical reasons?

Then you get into the issue of internal attrition. Analysing internal attrition is simply called attrition analysis or non-response analysis. What you do here is to pick one or more variables for which all individuals in the study sample has information, such as gender, age, or some other socio-demographic variable. Produce descriptive statistics (choice of type of descriptive statistics depends on the measurement scale) for those variables, for all individuals in the sample. Then you produce descriptive statistics for the same variables, but now only for the individuals in the *analytical* sample (Section 11.5 describes how to define an analytical sample).

For example, we have a study sample that contains 5,000 individuals. Approximately 49% are men and 51% are women. The mean age is 38 years. Due to missing data on some of the variables we want to include in our analysis, our analytical sample is reduced to 4,500 individuals. In this sample, 46% are men and 54% are women. The mean age is 40 years. You can illustrate this in a simple descriptive table:

	Sample (n=5,000)	**Analytical sample (n=4,500)**
Gender		
Man	49%	46%
Woman	51%	54%
Age (mean)	38 years	40 years

If we compare the distribution of gender and age in the study sample with the distribution of gender and age in the analytical sample, we can conclude that women and older individuals are more likely to be included in our analysis. This is information that could be important to have when we interpret our results.

11.5 From study sample to analytical sample

This section is an attempt to connect the two previous sections. It is like this: we often split our analysis in different steps or models. Thus, different models include different sets of variables; and different variables have different amount of missing data. The total number of individuals may therefore vary across models, and this makes it difficult to compare the results between the models. In other words, we should ensure that all our analyses – and all steps of analysis – are based on the same individuals. These individuals represent our analytical sample (or effective sample). Put differently: our analytical sample is defined as only those individuals who have valid information (i.e. no missing) for all variables we use in our analysis.

It is good to first check the amount of missing data for each of the variables included in the analysis, to see if any certain variable is particularly problematic in terms of missingness. If a variable has serious problems with missingness, it could be wise to exclude it from the analysis (but it depends on how important the variable is for your study).

The analytical sample should not only be the basis for regression analysis, but all other statistical tests and descriptive statistics should also be based on the analytical sample. Moreover, make sure to state the total number of individuals in the heading of each table and each figure. It could look something like this (see Section 4.8, for more advice on how to write headings):

Some examples

Table 1. Descriptive statistics for all study variables (n=9,451).

Figure 5. Histogram of annual income (n=9,451).

Table 3. The association between educational attainment and mortality. Results from logistic regression analysis, separately for men (n=4,701) and women (n=4,750).

11.5.1 The “pop” variable

It is easy to define an analytical sample in Stata. However, there are some different ways through which you can apply the analytical sample – below, we have described our favourite approach.

You first need to determine exactly which variables are included in the analysis (i.e. all variable you *use*, not all variables in the data material). They should have been properly examined (i.e. reviewed and checked with some initial descriptive statistics) and recoded as you want them.

In the example below, we have chosen four variables that we want to include in our study.

Practical example

Dataset: StataData1.dta

Name	**Label**
sex	Sex
bullied	Exposure to bullying (Age 15, Year 1985)
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)

```
sum sex bullied gpa cognitive
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         sex |     10,000       .4892    .4999083          0          1
     bullied |      8,719    .1076958    .3100137          0          1
         gpa |      9,380    3.178614    .6996298          1          5
   cognitive |      8,879    308.4708    72.18442        100        500
```

Apart from sex, we can see that they all have (different amounts of) missing values.

The first step is to create a “pop” variable – “pop” stands for population – with the `gen` command (see Section 5.2).

```
gen pop=1 if sex!=. & bullied!=. & gpa!=. & cognitive!=.
```

Through this, we specify that the new variable pop is assigned the value 1 if there is no missing information for any of the four variables. Let us check what it looks like:

```
tab pop
```

```
        pop |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      8,192      100.00      100.00
------------+-----------------------------------
      Total |      8,192      100.00
```

We can then apply the pop variable to anything we like, using if. For example:

```
tab sex if pop==1
```

```
        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
        Man |      3,876       47.31       47.31
      Woman |      4,316       52.69      100.00
------------+-----------------------------------
      Total |      8,192      100.00
```

Note Of course, you do not have to call this variable “pop” – choose any name you like.

11.6 Imputation

In the earlier sections, we suggested that the preferable strategy is to exclude individuals with missing data for any of the study variables from our analysis. This is often referred to as complete case analysis. Such an approach might, however, lead to biased estimates, inadequate power, and inaccurate standard errors.

Different types of imputation

An alternative is to apply imputation. Imputation means replacing missing data with substituted values, based on existing values in the data. An assumption is nonetheless that data are MCAR (or at least MAR) – which perhaps seldom is the case.

Types of imputation	
Mean/Median	Calculate the mean or median for the variable and impute that value for all individuals who have missing information for that variable. *A simple approach, but cannot be recommended since it introduces so much bias (e.g. reduces the variance).*
Hot deck/cold deck	Randomly (hot deck) or systematically (cold deck) choose a value from an individual in the sample who has similar values on all other study variables. *Simple, but restricts the range of possible values to the range among observed values.*
Last observation carried forward	Carry forward a value from the last observation for the same individual (works e.g. for repeated measurements) *Simple, but reduces the variance. Yields (potentially too) conservative estimates.*
Regression	Use the predicted value obtained by regressing the missing variable on other variables. *Preserves the relationships between the variables but not the variability around the predicted values.*
Stochastic regression	Use the predicted value obtained by regression the missing variable on other variables, plus a random residual value. *Improves the regression imputation by adding a random component.*
Extrapolation	Estimate a value from other observations for the same individual (works e.g., for repeated measurements). *Might, however, mean that one would estimate values beyond the actual range of data.*

Single versus multiple imputation

Single imputation means coming up with one single value of the missing value – which is simple and therefore quite compelling approach. Unless the data are really MCAR (or at least MAR), single imputation might nevertheless produce bias that is worse than what you would get with a complete case analysis.

The alternative is multiple imputation, which has become a very popular approach. This too assumes that missingness is MCAR or MAR. One starts by creating a number of sets of imputations for the missing values, based on an imputation method with a random component (such as hot deck imputation and stochastic regression imputation). After analysing each completed dataset, the results are combined. If performed well, multiple imputation leads to unbiased estimates and accurate standard errors.

Multiple imputation is not easy, and it requires deep knowledge about the dataset at hand. Therefore, we urge you to think long and hard about whether this is really a good strategy for your analysis. This guide will not cover any practical details about multiple imputation, but feel free to explore it further.

More information	help mi

12. LINEAR REGRESSION

Outline

Content

This chapter starts with an introduction to linear regression and then presents the function in Stata. After this, we offer some practical examples of how to perform simple and multiple linear regression, as well as how to generate and interpret model diagnostics.

12.1 Introduction

Linear regression is used when y is continuous (ratio/interval; see Section 3.3).

A linear regression model generally has the aim to predict or "forecast" the value of y, based on the values of one or more x-variables. Linear regression is concerned with finding the best-fitting straight line through the data points.

The regression line has an intercept (or constant) and a slope. The intercept is where the regression line strikes the y-axis when the value of the x-variable(s) is 0. The slope is basically the steepness of the line; i.e. how much y changes when x increases.

The regression model thus gives us predicted values of y across the values of the x-variable(s). Of course, there is generally a difference between what the model predicts and what the individuals' actual (observed) values are. This difference is called residual and is calculated as the observed value minus the predicted value.

Often, the term error is used instead of residual, and although these terms are closely related, they are not the exact same thing: an error is the difference between the observed value and the population mean (and the population mean is typically unobservable), whereas a residual is the difference between the observed value and the sample mean (and the sample mean is observable).

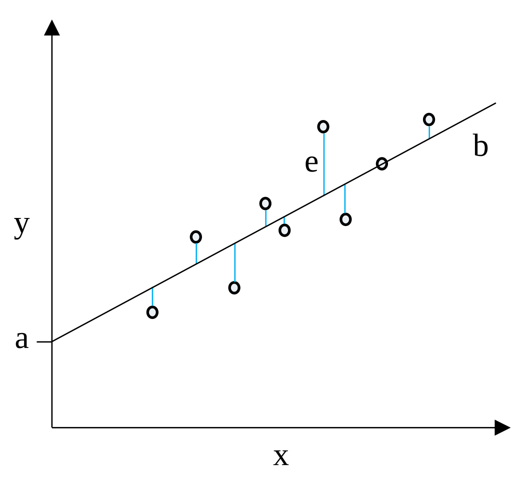
The most common method for fitting the linear equation is the method of ordinary least squares (OLS). It minimises the sum of squared differences between the observed and predicted values.

Just a little bit of maths

We promised to not have (almost) any equations in this guide, but here is a very simple expression of the one for linear regression:

$$y=a+bx+e$$

- y (or rather y hat; $\hat{y}$) is the predicted value of y.
- a is the intercept (or constant), i.e. the value of y when x=0.
- b is the slope (steepness) of the regression line, i.e. how much y changes per unit increase in x.
- x is the value of x.
- e is the error term (or residual), i.e. the error in predicting the value of y given the value of x.



Other names for linear regression

Linear regression is often referred to as OLS regression.

12.1.1 Linear regression in short

If you have only one x, it is called simple regression, and if you have more than one x, it is called multiple regression.

Regardless of whether you are doing a simple or a multiple regression, x-variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from linear regression	
Effect	
B coefficient (B)	The change in y, per unit increase in x
Direction	
Negative	B below 0
Positive	B above 0
Statistical significance	
P-value	$p<0.05$ Statistically significant at the 5% level $p<0.01$ Statistically significant at the 1% level $p<0.001$ Statistically significant at the 0.1% level
95% Confidence intervals	Interval does not include 0: Statistically significant at the 5% level Interval includes 0: Statistically non-significant at the 5% level

B coefficient (B)

In linear regression analysis, the effect that x has on y is reflected by a B coefficient (B):

Negative B coefficient	For every unit increase in x, y decreases by [B].
Positive B coefficient	For every unit increase in x, y increases by [B].

Exactly how one interprets the B coefficient in plain writing depends on the measurement scale of the x-variable. That is why we will present examples later for continuous, binary, and categorical (non-binary) x-variables.

Note What the B coefficient actually stands for depends on the values of x and y.

P-values and confidence intervals

In linear regression analysis you can get information about statistical significance, in terms of both p-values and confidence intervals.

Note The p-values and the confidence intervals will give you partly different information, but they are not contradictory. If the p-value is below 0.05, the 95% confidence interval will not include 0 and, if the p-value is above 0.05, the 95% confidence interval will include 0.

When you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5% level, the 1% level, or the 0.1% level).

When it comes to confidence intervals, Stata will by default choose 95% level confidence intervals. It is however possible to change the confidence level for the intervals. For example, you may instruct Stata to show 99% confidence intervals instead.

For more information about statistical significance, see Chapter 5.

R-Squared

You also get information about something called R-Squared or R2. This term refers to amount of the variance in y that is explained by the inclusion of the x-variable. The R2 value ranges between 0 and 1 – a higher value means a higher amount of explained variance. Generally speaking, the higher the R2 values, the better the model fits the data (i.e. the model has better predictive ability).

Simple versus multiple regression models

The difference between simple and multiple regression models, is that in a multiple regression each x-variable's effect on y is estimated while accounting for the other x-variables' effects on y. We then say that these other x-variables are "held constant", or "adjusted for", or "controlled for". Because of this, multiple regression analysis is a way of dealing with the issue of confounding variables, and to some extent also mediating variables (see Section 9.3).

It is highly advisable to run a simple regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted coefficients with (i.e. what happened to the coefficients when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that they become weaker – which would of course be expected if the x-variables overlapped in their effect on y.

A note

Remember that a regression analysis should always follow from theory as well as a comprehensive set of descriptive statistics and knowledge about the data. In the following sections, we will – for the sake of simplicity – not form any elaborate analytical strategy where we distinguish between x-variables and z-variables (see Chapter 9). However, we will define an analytical sample and use a so-called pop variable (see Section 11.5).

12.2 Function

Basic command	reg depvar indepvars	
Explanations	depvar indepvars	Insert the name of the y-variable. Insert the name of the x-variable(s) that you want to use.
Short names	reg	Regress
More information	help regress	

A walk-through of the output

When we perform a linear regression in Stata, the table looks like this:

```
      Source |       SS           df       MS      Number of obs   =     8,239
-------------+----------------------------------   F(2, 8236)      =   1392.53
       Model |  12513022.3         2  6256511.15   Prob > F        =    0.0000
    Residual |  37003480.9     8,236  4492.89472   R-squared       =    0.2527
-------------+----------------------------------   Adj R-squared   =    0.2525
       Total |  49516503.2     8,238  6010.74329   Root MSE        =    67.029

------------------------------------------------------------------------------
        yvar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xvar1 |   7.631827   1.476975     5.17   0.000     4.736583    10.52707
       xvar2 |   5.815392   .1108175    52.48   0.000     5.598162    6.032622
       _cons |   162.8872   3.044564    53.50   0.000     156.9191    168.8553
------------------------------------------------------------------------------
```

In this example, yvar ranges between 0 and 500, whereas xvar1 is a binary (0/1) variable and xvar2 is a continuous variable ranging between 1 and 40.

The upper left part of the table is an ANOVA table which shows distribution of variance. This is what the different columns mean:

Column	**Explanation**
Source	The Total variance is partitioned into Model and Residual. The former is the variance that can be explained by the Model, i.e. the x-variable(s) that we include. The latter is the variance which cannot be explained by the model.
SS	The sum of squares (SS) associated with the sources of variance.
Df	The degrees of freedom (df) associated with the sources of variance.
MS	The mean squares (MS), which is the sum of squares divided by the degrees of freedom.

The upper right part shows the overall model fit. This is what the different rows mean:

Row	**Explanation**
Number of obs	The number of observations included in the model.
F	F-value, calculated as the mean square model divided by the mean square residual.
Prob > F	The p-value associated with the F-value. If the p-value is below 0.05, it means that the x-variable(s) reliably predict the y-variable.
R-squared	The proportion of variance in the y-var that can be explained by the x-variable(s).
Adj R-squared	Same as R-squared, but accounts for the overlap in the variance explain by each x-variable.
Root MSE	Root mean square error (RMSE). This can be seen as a measure of accuracy (the lower the RMSE, the less errors, i.e. the better the predictive power).

The lower part of the table presents the parameter estimates from the analysis.

Column	**Explanation**
	The first column lists the y-variable on top, followed by our x-variable(s). The last row represents the constant (intercept).
Coef.	These are the B coefficients.
Std. Err.	The standard errors associated with the B coefficients.
t	T-value (B coefficient divided by its standard error).
P>|t|	P-value.
[95% Conf. Interval]	95% confidence intervals (lower limit and upper limit).

The analytical sample used for the examples

In the subsequent sections, we will use the following variables:

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)
bullied	Exposure to bullying (Age 15, Year 1985)
skipped	Skipped class (Age 15, Year 1985)

```
sum gpa cognitive bullied skipped
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      9,380    3.178614    .6996298          1          5
   cognitive |      8,879    308.4708    72.18442        100        500
     bullied |      8,719    .1076958    .3100137          0          1
     skipped |      8,843    1.701911    .6934793          1          3
```

We define our analytical sample through the following command:

```
gen pop_linear=1 if gpa!=. & cognitive!=. & bullied!=. & skipped!=.
```

This means that new the variable pop_linear gets the value 1 if the four variables do not have missing information. In this case, we have 8,136 individuals that are included in our analytical sample.

```
tab pop_linear
```

```
pop_linear |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      8,136      100.00      100.00
------------+-----------------------------------
      Total |      8,136      100.00
```

12.3 Simple linear regression

Quick facts	
Number of variables	One dependent (y) One independent (x)
Scale of variable(s)	Dependent: continuous (ratio/interval) Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

12.3.1 Simple linear regression with a continuous x

Theoretical examples

Example 1

Suppose we want to examine the association between unemployment days (x) and income (y). Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Income is measured in thousands of Swedish crowns per month and ranges between 20 and 40. Let us say that we get a B coefficient that is -0.13. That would mean that for each unit increase in unemployment days, income would (on average) decrease by 0.13. Given the values of our variables, we can conclude that for each additional day in unemployment, monthly income would decrease by 130 SEK on average.

Example 2

In another example, we may examine the association between time spent reading at home (x) and cognitive test scores (y). Time spent reading at home is a continuous variable measured in hours per week, and ranges between 0 and 10. Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Here, we get a B coefficient that is 5.499 Given the values of our variables, we can conclude that for each hour spent reading at home, the cognitive test score increases (on average) by almost six points.

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)

```
sum gpa cognitive if pop_linear==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      8,136    3.223144    .6860155          1          5
   cognitive |      8,136    312.7443    69.53904        100        500
```

```
reg gpa cognitive if pop_linear==1
```

```
      Source |       SS           df       MS      Number of obs   =     8,136
-------------+----------------------------------   F(1, 8134)      =   5060.73
       Model |  1468.37846         1  1468.37846   Prob > F        =    0.0000
    Residual |  2360.09351     8,134  .290151648   R-squared       =    0.3835
-------------+----------------------------------   Adj R-squared   =    0.3835
       Total |  3828.47197     8,135  .470617329   Root MSE        =    .53866

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cognitive |   .0061096   .0000859    71.14   0.000     .0059412    .0062779
       _cons |   1.312407   .0275152    47.70   0.000      1.25847    1.366343
------------------------------------------------------------------------------
```

R-squared is 0.38. Thus, cognitive explains 38% of the variance in gpa.

The B coefficient for cognitive is 0.006. In other words, for each point increase in the cognitive test score, the grade point average increases by 0.006. Although this is a very low estimate, we have to keep in mind that cognitive is a continuous variable ranging between 100 and 500. A unit increase in cognitive test scores is therefore not that much.

Note You can rescale continuous x-variables to make the interpretation more reasonable. For example, by multiplying the B coefficient for cognitive by 50, we would end up with the following interpretation: for every 50-point increase in the cognitive test score, the grade point average increases by 0.3. Still a quite low estimate, but slightly more sensible.

Concerning statistical significance, there is a statistically significant association between cognitive and gpa, as reflected by the p-value (0.000) and the 95% confidence interval (0.006 to 0.006).

Summary
There is a positive (B=0.006) and statistically significant (95% CI=0.006-0.006) association between cognitive test score and grade point average at age 15. In other words, the higher the cognitive test score, the higher the grade point average.

12.3.2 Simple linear regression with a binary x

Theoretical examples

Example 1

Suppose we want to examine the association between gender (x) and income (y). Gender has the values 0=Man and 1=Woman. Income is measured in thousands of Swedish crowns per month and ranges between 20 and 40. Let us assume that we get a B coefficient that is -1.3. That means that women have (on average) 1300 SEK less in monthly income compared to men.

Example 2

Suppose we want to examine the association between having young children (x) and the number of furry pets (y). Having young children is measured as either 0=No young children and 1=Young children. The number of furry pets is measured as the number of cats, dogs, or other furry animals living in the household, and ranges between 0 and 10. We get a B coefficient that is 0.98. In other words, those who have young children have (on average) almost one additional furry pet compared to those without young children.

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
bullied	Exposure to bullying (Age 15, Year 1985)

```
sum gpa bullied if pop_linear==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      8,136    3.223144    .6860155          1          5
     bullied |      8,136    .1039823    .3052563          0          1
```

The variable bullied is a binary variable with two categories: 0=No, 1=Yes. When we add it to the model, the category with the lowest value will be the reference category (i.e. No).

```
reg gpa bullied if pop_linear==1
```

```
      Source |       SS           df       MS      Number of obs   =     8,136
-------------+----------------------------------   F(1, 8134)      =     82.48
       Model |  38.4306413         1  38.4306413   Prob > F        =    0.0000
    Residual |  3790.04133     8,134  .465950495   R-squared       =    0.0100
-------------+----------------------------------   Adj R-squared   =    0.0099
       Total |  3828.47197     8,135  .470617329   Root MSE        =    .68261

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bullied |  -.2251621   .0247928    -9.08   0.000    -.2737624   -.1765618
       _cons |   3.246557   .0079948   406.08   0.000     3.230885    3.262229
------------------------------------------------------------------------------
```

R-squared is 0.01. Thus, bullied only explains 1% of the variance in gpa.

The B coefficient for bullied is -0.23. In other words, those who have been exposed to bullying have, on average, a 0.23 point lower grade point average compared to those who have not been exposed to bullying. This is not a very high estimate.

Nonetheless, there is a statistically significant association between bullied and gpa, as reflected by the p-value (0.000) and the 95% confidence interval (-0.27 to -0.18).

Summary

At age 15, there is a negative (B=-0.23) and statistically significant (95% CI=-0.27 to -0.18) association between exposure to bullying and grade point average. Put differently, individuals who were exposed to bullying received a lower grade point average compared to those who were not exposed.

12.3.3 Simple linear regression with a categorical (non-binary) x

Theoretical examples

Example 1

We want to investigate the association between educational attainment (x) and income (y). Educational attainment has the values: 1=Compulsory, 2=Upper secondary, and 3=University. We choose Compulsory as our reference category. Income is measured in thousands of Swedish crowns per month and ranges between 20 and 40. Let us say that we get a B coefficient for Upper secondary that is 2.1 and we get a B coefficient for University that is 3.4. In other words, those with upper secondary education have 2100 SEK higher income compared to those with compulsory education, and those with university education have 3400 SEK higher income compared to those with compulsory education.

Example 2

Suppose we are interested in the association between family type (x) and children’s average school marks (y). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. We choose Two-parent household as our reference category. Children’s average school marks range from 1 to 5. The analysis results in a B coefficient of -0.1 for joint custody and a B coefficient of -0.9 for single-parent household. That would mean that children living in joint custody families have a 0.1 point lower score for average school marks compared to those living in two-parent households. Moreover, children living in single-parent households have a 0.9 point lower score for average school marks compared to those living in two-parent households.

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
skipped	Skipped class (Age 15, Year 1985)

```
sum gpa skipped if pop_linear==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      8,136    3.223144    .6860155          1          5
     skipped |      8,136    1.685226    .6906936          1          3
```

The variable skipped has three categories: 1=Never, 2=Sometimes, and 3=Often. Here, we (with ib1) specify that the first category (Never) will be the reference category.

```
reg gpa ib1.skipped if pop_linear==1
```

```
      Source |       SS           df       MS      Number of obs   =     8,136
-------------+----------------------------------   F(2, 8133)      =    147.39
       Model |  133.908916         2  66.9544579   Prob > F        =    0.0000
    Residual |  3694.56305     8,133  .454268173   R-squared       =    0.0350
-------------+----------------------------------   Adj R-squared   =    0.0347
       Total |  3828.47197     8,135  .470617329   Root MSE        =    .67399

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     skipped |
  Sometimes  |  -.1792257   .0160331   -11.18   0.000    -.2106546   -.1477967
      Often  |   -.376323   .0235095   -16.01   0.000    -.4224076   -.3302385
             |
       _cons |   3.348289    .011196   299.06   0.000     3.326342    3.370236
------------------------------------------------------------------------------
```

R-squared is 0.04. Thus, skipped only explains 4% of the variance in gpa.

With regard to the B coefficient, we get two: one for skipped: Sometimes and one for skipped: Often. They are compared to the reference category skipped: Never. The refeence group in linear regression always has a B coefficient of 0.00. In this case we can see that the B coefficient for Sometimes is -0.18, and for Often it is -0.38. Put differently, the more the individuals have skipped class, the lower the grade point average.

Both Sometimes and Often have p-values that are below 0.05 (0.000) and the 95% confidence intervals are -0.21 to -0.15 and -0.42 to -0.33, respectively. Thus, there is a statistically significant difference in gpa between Sometimes and Never, and between Often and Never.

Test the overall effect

The output presented and interpreted above, is based on the coefficients for the dummy variables of skipped. But what about the overall statistical effect of skipped on gpa? We can assess it through contrast, which is a postestimation command.

```
contrast p.skipped, noeffects
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     skipped |
   (linear)  |          1      256.23     0.0000
(quadratic)  |          1        0.30     0.5865
      Joint  |          2      147.39     0.0000
             |
 Denominator |       8133
------------------------------------------------
```

Here, we focus on the row for linear, which shows a p-value (P>chi2) below 0.05. This suggests that we have a statistically significant trend in gpa according to skipped.

More information	help contrast

We will also produce a graph of the trend. First, however, we need to apply the post-estimation command margins.

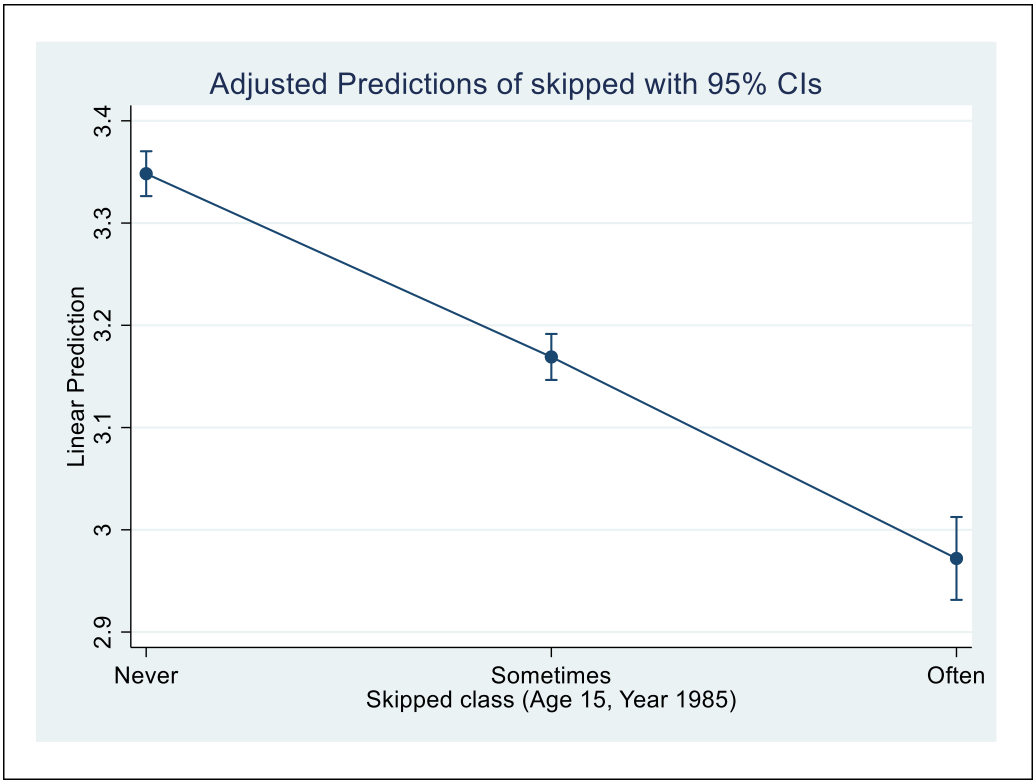
Note This command can also be used for variables that are continuous or binary, but is particularly useful for categorical, non-binary (i.e. ordinal) variables.

```
margins skipped
```

```
Adjusted predictions                            Number of obs     =      8,136
Model VCE    : OLS

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     skipped |
      Never  |   3.348289    .011196   299.06   0.000     3.326342    3.370236
   Sometimes |   3.169063   .0114765   276.13   0.000     3.146567     3.19156
      Often  |   2.971966   .0206723   143.77   0.000     2.931443    3.012489
------------------------------------------------------------------------------
```

Note that the estimate for Never in the column Margin is exactly reflecting the constant from the linear regression analysis (3.348289). Adding the B coefficient for Sometimes (-0.1792257), we end up with the estimate for Sometimes in this table (3.169063). Adding the B coefficient for Often (-0.376323), we get the estimate for Often in this table (2.971966).

```
marginsplot
```



Note The y-axis shows predicted values (i.e. not B coefficients).

More information	help marginsplot

Summary
Among 15-year-olds, there is a negative and statistically significant association between having skipped class and grade point average. The association is graded: those who skipped class sometimes have a lower grade point average (B=-0.18, 95% CI=-0.21 to -0.15) and those who skipped class often have even lower (B=-0.38, 95% CI=-0.42 to -0.33), compared to those who never skipped class.

12.4 Multiple linear regression

Quick facts	
Number of variables	One dependent (y) At least two independent (x)
Scale of variable(s)	Dependent: continuous (ratio/interval) Independent: categorical (nominal/ordinal) and/or continuous (ratio/interval)

Theoretical example

Example

Suppose we are interested to see if young children (x), residential area (x), and income (x) are related to the number of furry pets (y).

Having young children is measured as either 0=No young children and 1=Young children. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. We choose Metropolitan as our reference category. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700). The number of furry pets is measured as the number of cats, dogs or other furry animals living in the household, and ranges between 0 and 10.

We get a B coefficient for having young children that is 0.51. That means that the number of furry pets is higher among those who have young children. This association is adjusted for residential area and income.

With regards to residential area, the B coefficient for Smaller city is 2.02 whereas the B coefficient for Rural is 4.99. That suggests, firstly, that the number of furry pets is higher (about two more pets, on average) among individuals living in smaller cities compared to metropolitan areas. Secondly, the number of furry pets is much higher (almost five more pets, on average) among individuals living in rural areas compared to metropolitan areas. This association is adjusted for having young children and income.

Finally, the B coefficient for income is -0.1. This suggests that for every unit increase in income (i.e. for every additional one thousand SEK), the number of furry pets decrease by 0.1. This association is adjusted for having young children and residential area.

Practical example

Dataset: StataData1.dta

Name	**Label**
gpa	Grade point average (Age 15, Year 1985)
cognitive	Cognitive test score (Age 15, Year 1985)
bullied	Exposure to bullying (Age 15, Year 1985)
skipped	Skipped class (Age 15, Year 1985)

```
sum gpa cognitive bullied skipped if pop_linear==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         gpa |      8,136    3.223144    .6860155          1          5
   cognitive |      8,136    312.7443    69.53904        100        500
     bullied |      8,136    .1039823    .3052563          0          1
     skipped |      8,136    1.685226    .6906936          1          3
```

In this model, we have three x-variables: cognitive, bullied, and skipped. When we put them together, their statistical effect on gpa is mutually adjusted.

```
reg gpa cognitive bullied ib1.skipped if pop_linear==1
```

```
      Source |       SS           df       MS      Number of obs   =     8,136
-------------+----------------------------------   F(4, 8131)      =   1465.17
       Model |  1603.62921         4  400.907303   Prob > F        =    0.0000
    Residual |  2224.84276     8,131   .27362474   R-squared       =    0.4189
-------------+----------------------------------   Adj R-squared   =    0.4186
       Total |  3828.47197     8,135  .470617329   Root MSE        =    .52309

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cognitive |   .0060635   .0000841    72.13   0.000     .0058987    .0062283
     bullied |  -.0737877    .019177    -3.85   0.000    -.1113795   -.0361959
             |
     skipped |
  Sometimes  |  -.1813156   .0124573   -14.56   0.000    -.2057351   -.1568962
      Often  |  -.3741848   .0182611   -20.49   0.000    -.4099812   -.3383884
             |
       _cons |   1.460238   .0280723    52.02   0.000     1.405209    1.515266
------------------------------------------------------------------------------
```

In the simple regression models, we had R-squared values of 0.3835 (for cognitive), 0.0100 (for bullied), and 0.0350 (for skipped). Now that we have a multiple regression analysis, it is better to look at the adjusted R-squared, which in this case is 0.4186. This means that 42% of the variance in gpa is explained by our three x-variables.

When it comes to the B coefficients, they are roughly the same or somewhat lower (i.e. closer to 0) in comparison to the simple regression models. For example, the B coefficient for cognitive is still 0.006. The B coefficient for bullied is lower: -0.07 here instead of -0.23. Concerning the categories of skipped, we see that the B coefficient for Sometimes is still -0.18 and the B coefficient for Often is -0.37 instead of -0.38.

The associations between the x-variables and gpa are still statistically significant ($p< 0.05$) after mutual adjustment.

Summary
In the fully adjusted model, it can be observed that the associations with grade point average are not altered in any substantial way in comparison to the simple models. To conclude, cognitive test scores, exposure to bullying, and having skipped class are associated with grade point average at a statistically significant level (all: p=0.000). Nonetheless, the associations are generally rather weak.

Estimates table and coefficients plot

If we have multiple models, we can facilitate comparisons between the regression models by asking Stata to construct estimates tables and coefficients plots. What we do is to run the regression models one-by-one, save the estimates after each, and than use the commands estimates table and coefplot.

The coefplot option is not part of the standard Stata program, so unless you already have added this package, you need to install it:

```
ssc install coefplot
```

As an example, we can include the three simple regression models as well as the multiple regression model. The quietly option is included in the beginning of the regression commands to suppress the output.

Run and save the first simple regression model:

```
quietly reg gpa cognitive if pop_linear==1
```

```
estimates store model1
```

Run and save the second simple regression model:

```
quietly reg gpa bullied if pop_linear==1
```

```
estimates store model2
```

Run and save the third simple regression model:

```
quietly reg gpa ib1.skipped if pop_linear==1
```

```
estimates store model3
```

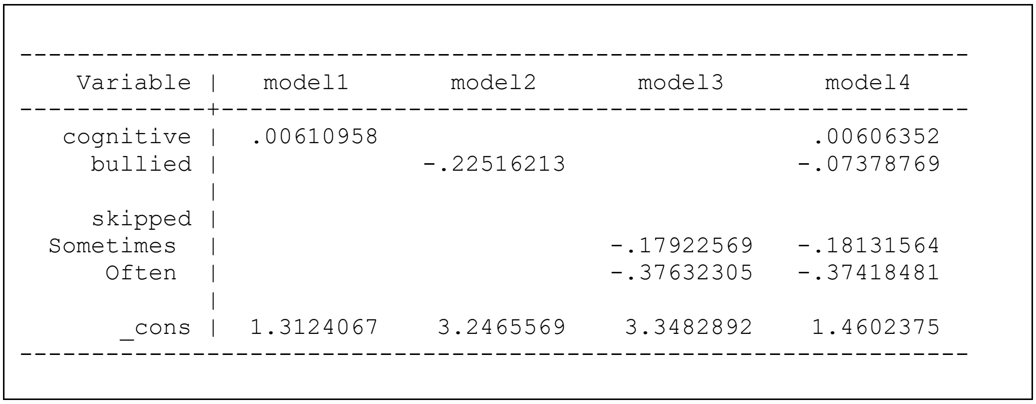
Run and save the multiple regression model:

```
quietly reg gpa cognitive bullied ib1.skipped if pop_linear==1
```

```
estimates store model4
```

Produce the estimates table:

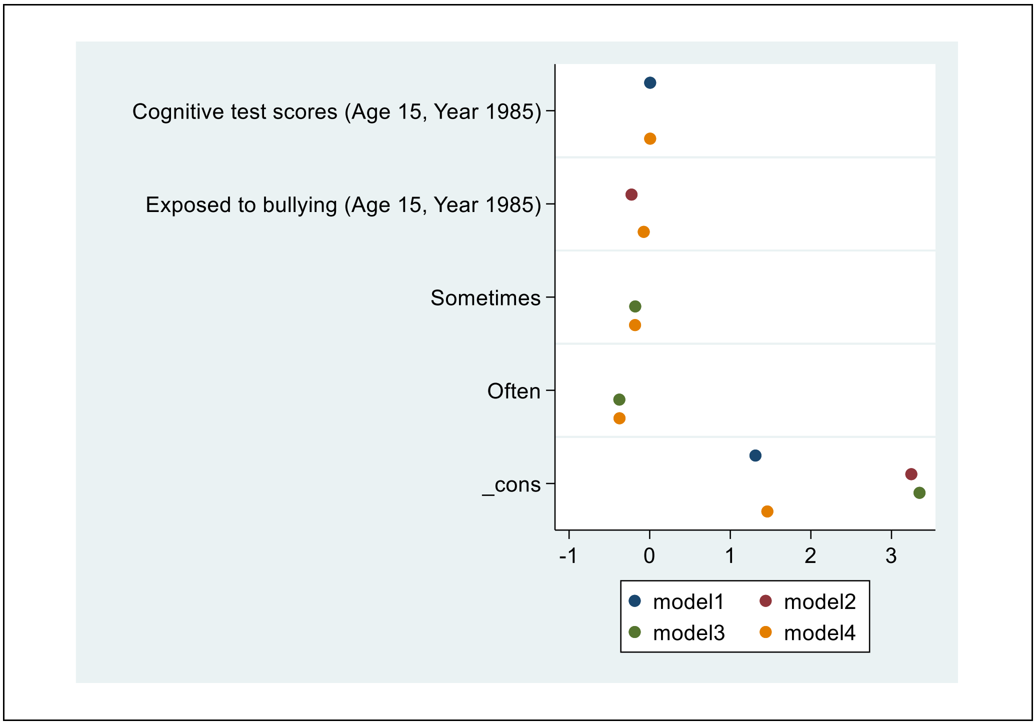
```
estimates table model1 model2 model3 model4
```



```
---------------------------------------------------------------------
    Variable |   model1       model2       model3       model4
-------------+-------------------------------------------------------
   cognitive |  .00610958                              .00606352
     bullied |              -.22516213                -.07378769
             |
     skipped |
  Sometimes  |                           -.17922569   -.18131564
      Often  |                           -.37632305   -.37418481
             |
       _cons |  1.3124067    3.2465569    3.3482892    1.4602375
---------------------------------------------------------------------
```

Produce the coefficients plot:

```
coefplot model1 model2 model3 model4
```



Note You can improve the graph by using the Graph Editor to delete “_cons” as well as to adjust the category and label names.

12.5 Model diagnostics

Before we can trust the results from our linear regression analysis to be valid, we need to assess our model to check that it does not violate any of the fundamental assumptions of linear regression.

More information	help reg postestimation

Checklist	
Continuous and normally distributed outcome	The y-variable has to be continuous. It should also be normally distributed. Check this with a histogram. If it is not normally distributed, you might need to consider another alternative. For example, you can transform your y-variable (e.g. through categorisation, or log transformation).
Correct model specification	Your model should be correctly specified. This means that the x-variables that are included should be meaningful and contribute to the model. No important (confounding) variables should be omitted (often referred to as omitted variable bias).
No outliers	Outliers are individuals who do not follow the overall pattern of data. Sometimes referred to as influential observations (however, not all outliers are influential).
Homoscedasticity	The variance around the regression line should be constant across all values of the x-variable(s).
Normality	The residuals for our x-variables should be normally distributed.
Linearity	The effect of x on y should be linear.
No multicollinearity	Multicollinearity may occur when two or more x-variables that are included simultaneously in the model are strongly correlated with each another. Actually, this does not violate the assumptions, but is does create greater standard errors which makes it harder to reject the null hypothesis.

Types of model diagnostics	
Link test	Assess model specification
Residual plot	Check for linearity, homoscedasticity, and outliers
Breusch-Pagan/Cook-Weisberg test	Check for homoscedasticity
Density plot	Check for normality
Normal probability plot	Check for normality
Normal quantile plot	Check for normality
Variance inflation factor	Check for multicollinearity
Correlation matrix	Check for multicollinearity

12.5.1 Link test

With the command linktest, we can assess whether our model is correctly specified. This test uses the linear predicted value (called _hat) and the linear predicted value squared (_hatsq) to rebuild the model. We expect _hat to be statistically significant, and _hatsq to be statistically non-significant. If one or both of these expectations are not met, the model is mis-specified.

However, do not rely too much on this test – remember that you should also use theory and common sense to guide your decisions. It is very seldom relevant to focus on this test if our ambition is to investigate associations (and not to make the best possible prediction of the outcome).

More information	help linktest

Practical example

We perform this test for the full model, so let us go back to the example from the multiple linear regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly reg gpa cognitive bullied ib1.skipped if pop_linear==1
```

And then we run the test:

```
linktest
```

```
      Source |       SS           df       MS      Number of obs   =     8,136
-------------+----------------------------------   F(2, 8133)      =   3032.82
       Model |  1635.5178         2  817.758901   Prob > F        =    0.0000
    Residual |  2192.95417     8,133  .269636563   R-squared       =    0.4272
-------------+----------------------------------   Adj R-squared   =    0.4271
       Total |  3828.47197     8,135  .470617329   Root MSE        =    .51927

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |  -.5712423   .1450631    -3.94   0.000    -.8556031   -.2868815
      _hatsq |   .2478382   .0227898    10.87   0.000     .2031644     .292512
       _cons |   2.440784   .2283713    10.69   0.000     1.993118     2.88845
------------------------------------------------------------------------------
```

The p-value for _hat is below 0.05, but since the p-value for the variable _hatsq is also below 0.05, it means that our model is not correctly specified. We could try to amend this by transforming any of the included variables (e.g. through categorisation, or log transformation), excluding any of the included variables, or adding more variables to the model (other x-variables or e.g. interactions between the included variables).

12.5.2 Residual plot

A residual plot graphs the residuals (on the y-axis) against the fitted values (on the x-axis). Residual plots can be produced with the rvfplot command. This is a postestimation command, so you need to order it right after your regression analysis.

If the points in the plot are evenly/randomly dispersed around the x-axis, it means that a linear regression is appropriate. If not – and there is some type of pattern (e.g. cone-shaped) emerging in the plot – then you most likely have problems with heteroskedasticity. If the pattern is such that the points are not following the regression line (e.g. showing a curve-linear pattern), you may have problems with non-linearity. Moreover, you will quite clearly see if there are any outliers in the plot.

More information	help rvfplot

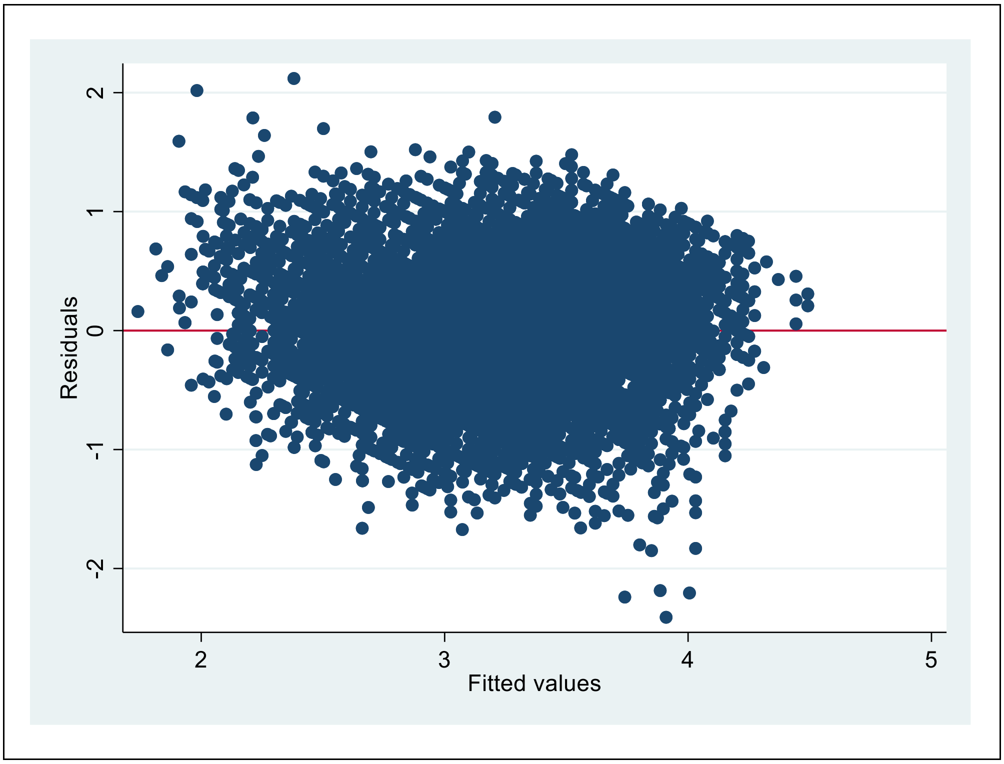
Practical example

Usually, we would conduct model diagnostics for the full model, so we go back to the example from the multiple linear regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly reg gpa cognitive bullied ib1.skipped if pop_linear==1
```

We then order the residual plot:

```
rvfplot, yline(0)
```



It looks pretty OK – apart from the points in the upper left corner and the lower right corner. This suggests that this model might not have any massive problems with heteroskedasticity, non-linearity, or outliers.

12.5.3 Breusch-Pagan/Cook-Weisberg test

There are some tests that can be used to assess whether our assumption of homoskedasticity holds or not. One of them is the Breusch-Pagan/Cook-Weisberg test of heteroskedasticity (estat hettest command). It uses the fitted values of the y-variable and produces a p-value: if p<0.05, it means that our assumption is violated.

More information	help estat hettest

Practical example

The first step is re-run the multiple linear regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly reg gpa cognitive bullied ib1.skipped if pop_linear==1
```

Then we run the test.

```
estat hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of gpa

         chi2(1)      =     2.90
         Prob > chi2  =   0.0884
```

Since the p-value is above 0.05 (0.0884), we can conclude that our model does not have problems with heteroskedasticity.

12.5.4 Density plot, normal probability plot, and normal quantile plot

A density plot is a graph of the residuals with a normal distribution curve superimposed. It can be used to check whether the normality assumption holds. In Stata, kdensity (k=kernel) can be used to generate the density plot.

More information	help kdensity

The normal probability plot (pnorm) constitutes another a way of testing whether the residuals are normally distributed. Compared to the normal quantile plot, it is more sensitive to anomalies in the middle of the distribution.

More information	help pnorm

Yet another alternative for checking the normality assumption is the normal quantile plot (qnorm). Compared to the normal probability plot, it is more sensitive to anomalies in the tails of the distribution.

More information	help pnorm

Practical example

The first step is re-run the multiple linear regression model. The quietly option is included in the beginning of the command to suppress the output.

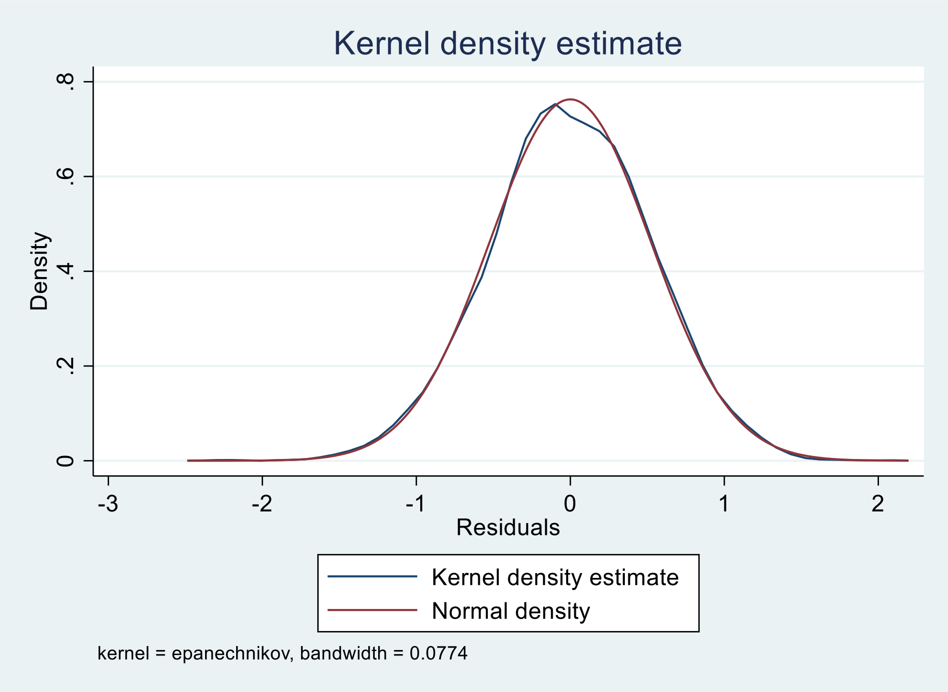
```
quietly reg gpa cognitive bullied ib1.skipped if pop_linear==1
```

Then we have to save the residuals from the model by creating a new variable, here called res, using the predict command.

```
predict res, resid
```

The next step is to produce the density plot (the option "normal" means that we include a normal distribution curve in the graph).

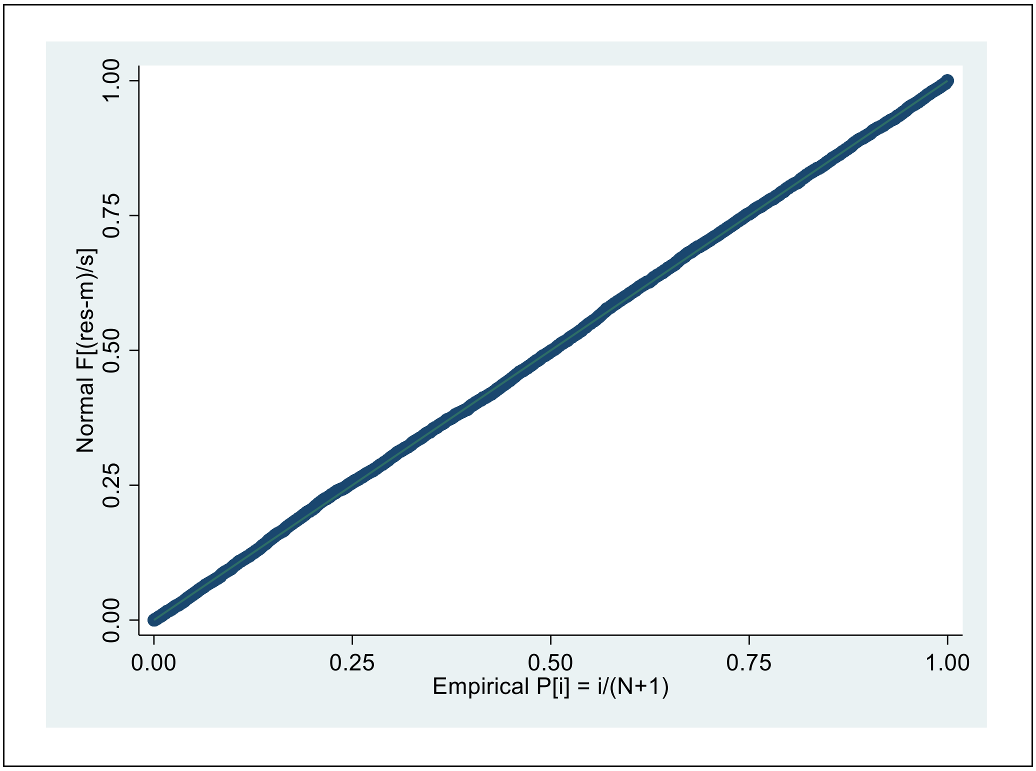
```
kdensity res, normal
```



In this example, the residuals seem to be pretty normally distributed (apart from the small dip at the top of the peak).

Now we will generate a normal probability plot. We can re-use the variable res for this.

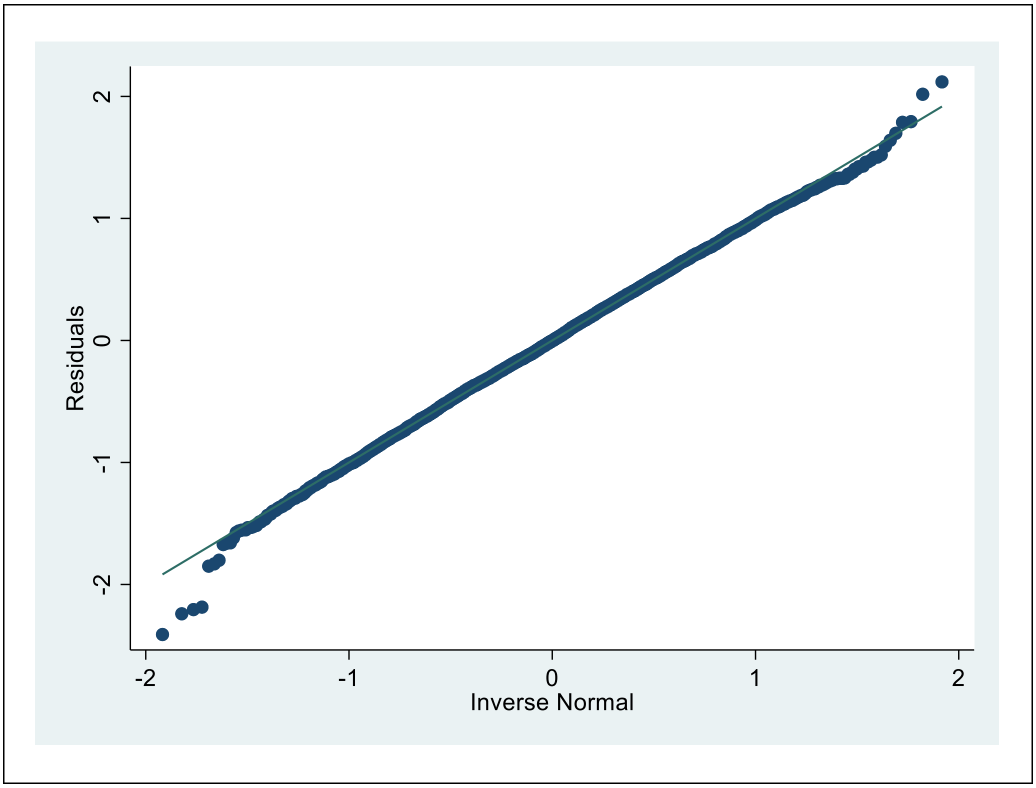
```
pnorm res
```



A straight, diagonal line like this means that the residuals are normally distributed.

The final step is to create the normal quantile plot.

```
qnorm res
```



This plot too looks good. Deviation at the tails is almost inevitable – it is more problematic if the points are distributed in a wider s-shaped pattern and deviate from the diagonal over the whole range of values.

12.5.5 Variance inflation factor and correlation matrix

As the x-variables become more strongly correlated, it becomes more difficult to determine which of the variables are actually producing the statistical effect on the y-variable. This is a problem of multicollinearity.

One way of assessing problems with multicollinearity is through the estat vif command (vif=variance inflation factor). This tells us how much of the variance that is being inflated by multicollinearity. As a rule of thumb, a vif-value that is near 10 or higher calls for concern.

More information	help estat vif

Another way of assessing multicollinearity is using the estat vce command, with the corr (short for correlation) option.

More information	help estat vce

Practical example

The first step is re-run the multiple linear regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly reg gpa cognitive bullied ib1.skipped if pop_linear==1
```

Next, we try the estat vif command.

```
estat vif
```

```
    Variable |       VIF       1/VIF
-------------+----------------------
   cognitive |      1.02    0.984328
     bullied |      1.02    0.981538
     skipped |
          2  |      1.13    0.887427
          3  |      1.13    0.887926
-------------+----------------------
    Mean VIF |      1.07
```

We get a mean vif-value of 1.07, which tells us that we do not seem to have any problems with multicollinearity in this model.

Let us also try the estat vce command. By adding the corr (=correlation) option, we will get a correlation matrix instead of a covariance matrix.

```
estat vce, corr
```

```
Correlation matrix of coefficients of regress model

             |                              2.         3.
        e(V) |  cognit~e    bullied    skipped    skipped      _cons
-------------+-------------------------------------------------------
   cognitive |    1.0000
     bullied |    0.1251     1.0000
   2.skipped |    0.0061     0.0472     1.0000
   3.skipped |    0.0088     0.0406     0.3338     1.0000
       _cons |   -0.9473    -0.2005    -0.2255    -0.1589     1.0000
```

The table shows the correlations between the different variables/categories. In line with the earlier sections on correlation analysis (see Chapter 7.2), we can conclude that the coefficients suggest (very) weak correlations here.

13. LOGISTIC REGRESSION

Outline

Content

This chapter starts with an introduction to logistic regression and then presents the function in Stata. After this, we offer some practical examples of how to perform simple and multiple logistic regression, as well as how to generate and interpret model diagnostics.

13.1 Introduction

Logistic regression is used when y is categorical with only two categories, i.e. dichotomous/binary (see Section 3.3).

Cases and non-cases

A logistic regression is based on the fact that the outcome has only two possible values: 0 or 1. Often, the value 1 is used to denote a “case” whereas the value 0 is then a “non-case”. What is meant by case or non-case depends on how the hypothesis is formulated.

Example

a. We want to investigate the association between educational attainment (x) and employment (y). Our hypothesis is that educational attainment is positively associated with employment (i.e. higher educational attainment = more likely to be employed).
Coding of employment: 0=Unemployment (non-case); 1=Employment (case).

b. We want to investigate the association between educational attainment (x) and unemployment (y). Our hypothesis is that educational attainment is negatively associated with unemployment (i.e. higher educational attainment = less likely to be unemployed).
Coding of unemployment: 0=Employment (non-case); 1=Unemployment (case).

Logistic regression is used to predict the odds of being a case, compared to not being a case, based on the values of x. We get a coefficient – called log odds – that shows the effect of x on y. The log odds are the natural logarithm of the odds. These coefficients are not easy to interpret. Instead, we usually focus on something called the odds ratio (OR). The OR is calculated by taking the exponent of the coefficient. This part is further explained below.

Why not use linear regression?

With linear regression, we model the mean outcome. When we have a binary outcome, the mean is a probability.

What is a probability?

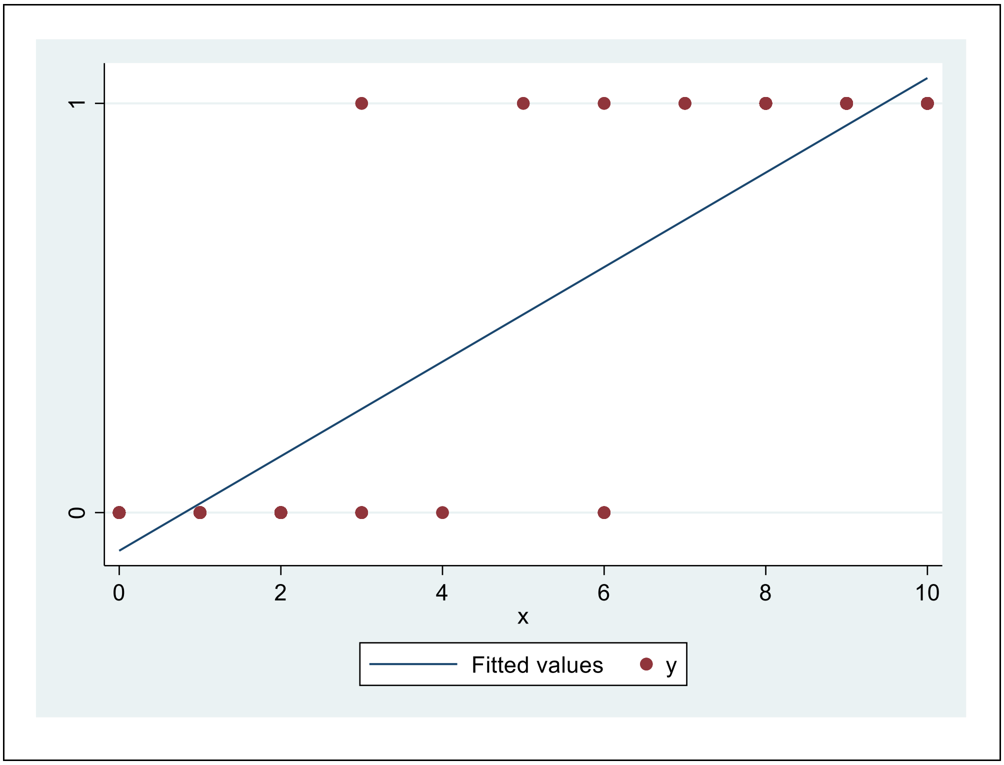
- The extent to which an event is likely to occur. Or, if we stick to the terminology presented earlier: the extent to which the outcome is likely to be a case.
- If the probability of the outcome being a case is p, then the probability of the outcome being a non-case is 1-p.
- The formula can be expressed as: p(case)=number of cases/total number of cases+non-cases.
- Probabilities always range between 0 and 1.

Example

We have a sample of 10 individuals, of which 3 are diagnosed with depression (cases), and 7 are not (non-cases). The probability of depression in the sample is thus 3/10=0.3 (can also be expressed as percentages, which would be 30%).

Moreover, 5 of the individuals are men, of which 1 is a case and 4 are non-cases. The remaining 5 individuals are women, of which 2 are cases and 3 are non-cases. The probability of depression among men is thus 1/5=0.2 (20%) whereas the probability of depression among women is 2/5=0.4 (40%).

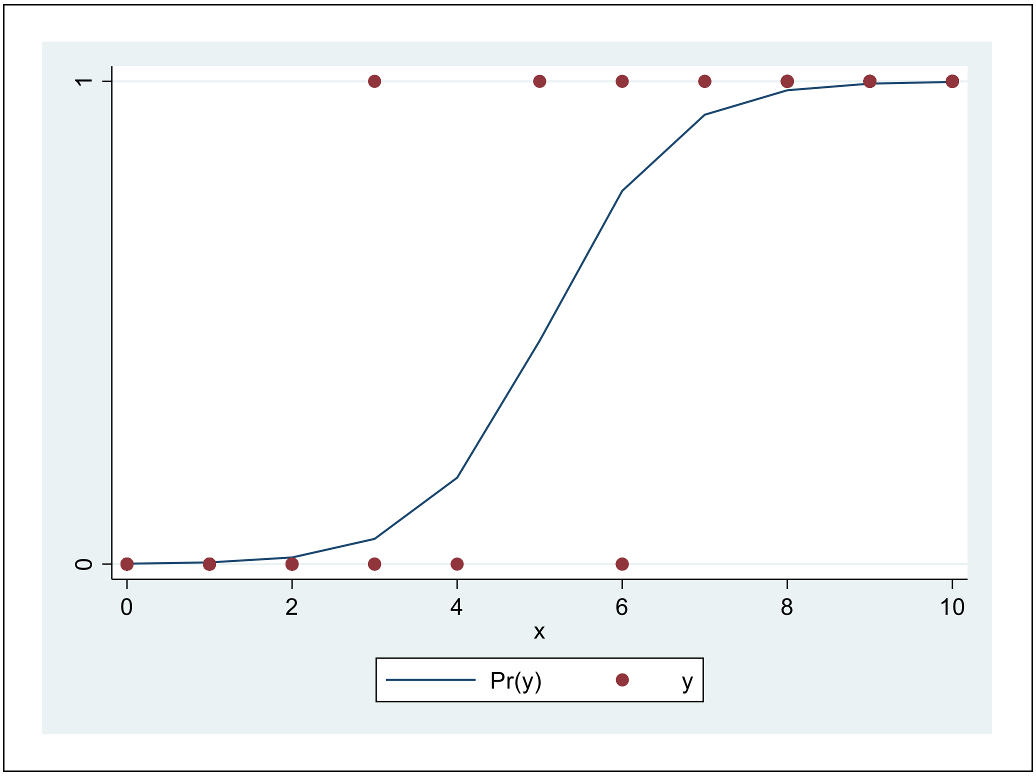
If we were to fit a linear regression for a binary outcome, it is fully possible that we will have predicted values that are outside of the range of probabilities (i.e. below 0 and/or above 1). See for example the figure below, where a binary outcome is modelled together with a continuous x-variable, using linear regression.



Apart from this, applying a linear regression to a binary outcome will violate several of the other assumptions of linear regression analysis (normality, homoscedasticity).

The logistic function

Instead, we can apply a generalised linear model (GLM) which uses a link function that allows the outcome to vary linearly with the predicted values instead of varying linearly with the x-variable(s). For logistic regression, the link function that we choose is the logistic function. Through this, we restrict the probabilities to vary between 0 and 1. See for example the figure below, where a binary outcome is modelled together with a continuous x-variable, using logistic regression.



But how does the logistic function work? Well, it transforms probabilities to log odds, using maximum likelihood estimation. For logistic regression, the maximum likelihood method is the equivalent to ordinary least squares (OLS).

What are odds?

- The probability that the outcome will be a case, divided by the probability that the outcome will be a non-case.
- Can take any value from zero to infinity.
- If the probability of the outcome being a case is p, then the odds of the outcome being a case is p/(1-p).

Example: In our sample, the probability of having been diagnosed with depression is 0.3. This means that the odds of depression are 0.3/1-0.3=0.4286 (rounded value).

For men, the probability of having been diagnosed with depression is 0.2. Their odds of depression are thus 0.2/1-0.2=0.25. For women, the probability of having been diagnosed with depression is 0.4. Their odds of depression are thus 0.4/1-0.4=0.6667 (rounded value).

What are log odds?

- The logarithm of the odds (the logarithm is the power to which a number must be raised in order to produce some other number).
- Also referred to as the logit of the probability.
- Can take any value.
- Is symmetric around zero.
- Estimated as: log(p/1-p).

Example: In our sample, the odds of depression are 0.4286. This corresponds to log odds of -0.8472. For men, the odds of depression are 0.25, which means that the log odds are -1.3863. The odds among women are 0.667 and their log odds are thus -0.4054. That is a difference of 0.9808 (rounded value) between men and women (women have 0.9808 higher log odds than men).

So far, so good! But as we previously mentioned, the log odds are not easy to interpret. That is why it is very common to convert them to odds ratios.

What is an odds ratio (OR)?

- The exponent of the log odds (the exponent is a special way expressing repeated multiplications).
- Can take any value from zero to infinity.
- Estimated as: exp(log odds).

Example: A difference of 0.9808 between men and women, corresponds to an OR of 2.67 (rounded value). Thus, women have 2.67 times the odds of depression compared to men.

Before we continue, let us revisit the mathematical expression for linear regression:

$$y=a+bx+e$$

- y (or rather y hat; ŷ) is the predicted value of y.
- a is the intercept (or constant), i.e. the value of y when x=0.
- b is the slope (steepness) of the regression line, i.e. how much y changes per unit increase in x.
- x is the value of x.
- e is the error term (or residual), i.e. the error in predicting the value of y given the value of x.

For logistic regression, the formula is:

$$\log(p/1-p)=a+bx+e$$

- log(p/1-p) is the log transformation of the probability that the outcome will be a case, divided by the probability that the outcome will be a non-case.
- a is the intercept (or constant), i.e. the log odds of y when x=0.
- b is the change in log odds per unit increase in x. Can be transformed to odds ratio by taking the exponent of b. Can be transformed back to log odds by taking the log of the odds ratio.
- x is the value of x.
- e is the error term (or residual), i.e. the error in predicting the probability of y given the value of x.

Other names for logistic regression

We have chosen to use the term logistic regression when we refer to binary logistic regression or binomial regression (in reality, ordinal regression and multinomial regression are also types of logistic regressions, see Chapters 14-15). Other names for this type of regression model are, e.g., logit regression and generalized linear model (GLM) with logit link function.

13.1.1 Logistic regression in short

If you have only one x, it is called simple regression, and if you have more than one x, it is called multiple regression.

Regardless of whether you are doing a simple or a multiple regression, x-variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from logistic regression		
Effect		
Odds ratio (OR)	The exponent of log odds	
	Log odds	The logarithm of odds
	Odds	The probability of the outcome being case divided by the probability of the outcome being a non-case
	Probability	The probability of an event happening
Direction		
Negative	OR below 1	
Positive	OR above 1	
Statistical significance		
P-value	p<0.05 Statistically significant at the 5% level p<0.01 Statistically significant at the 1% level p<0.001 Statistically significant at the 0.1% level	
95% Confidence intervals	Interval does not include 1: Statistically significant at the 5% level Interval includes 1: Statistically non-significant at the 5% level	

Odds ratio (OR)

In logistic regression analysis, the effect that x has on y is reflected by an odds ratio (OR):

OR below 1	For every unit increase in x, the odds of y decreases.
OR above 1	For every unit increase in x, the odds of y increases.

Exactly how one interprets the OR in plain writing depends on the measurement scale of the x-variable. That is why we will present examples later for continuous, binary, and categorical (non-binary) x-variables.

Note Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for logistic regression is 1.

Note An OR can never be negative – it can range between 0 and infinity.

How to *not* interpret odds ratios

Odds ratios are not the same as risk ratios (see Section 4.7.6). ORs tend to be inflated when they are above 1 and understated when they are below 1. This becomes more problematic the more common the outcome is (i.e. the more “cases” we have). However, the rarer the outcome is (<10% is usually considered a reasonable cut-off here), the closer odds ratios and risks ratios become.

Many would find it compelling to interpret ORs in terms of percentages. For example, an OR of 1.20 might lead to the interpretation that the odds of the outcome increase by 20%. If the OR is 0.80, some would then suggest that the odds decrease by 20%. We would to urge you to carefully reflect upon the latter kind of interpretation since odds ratios are not symmetrical: it can take any value above 1 but cannot be below 0. Thus, the choice of reference category might lead to quite misleading conclusions about effect size. The former kind of interpretation is usually considered reasonable when ORs are below 2. If they are above 2, it is better to refer to “times”, i.e. an OR of 4.07 could be interpreted as “more than four times the odds of…”.

Take home messages
Do not interpret odds ratios as risk ratios, unless the outcome is rare (<10%, but even then, be careful).
It is completely fine to discuss the results more generally in terms of higher or lower odds/risks. However, if you want to give exact numbers to exemplify, you need to consider the asymmetry of odds ratios as well as the size of the OR.

Some examples

- The results suggest that women (OR=0.84) are less likely than men to subscribe to a daily newspaper.
- Based on logistic regression analysis, it may be concluded that individuals with more behavioural problems in childhood have a greater risk of drug abuse in adulthood (OR=1.49).
- There is a negative association between educational attainment and number of children: the higher the educational attainment, the lower the number of children (OR=0.90).
- Individuals living in urban areas (OR=0.33) are less likely compared to those living in rural areas to own a horse.

Note Do you have a binary outcome and would like to produce risk ratios instead of odds ratios? Perform a Poisson regression analysis for your binary outcome. The coefficients that you produce will be equivalent to risk ratios (and not incidence rate ratios). For more information on how to conduct Poisson regression, see Chapter 15.

P-values and confidence intervals

In logistic regression analysis you can get information about statistical significance, in terms of both p-values and confidence intervals (also see Section 5.2).

Note The p-values and the confidence intervals will give you partly different information, but they are not contradictory. If the p-value is below 0.05, the 95% confidence interval will not include 1 and, if the p-value is above 0.05, the 95% confidence interval will include 1.

When you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5% level, the 1% level, or the 0.1% level).

Concerning confidence intervals, Stata will by default choose 95% level confidence intervals. It is however possible to change the confidence level for the intervals. For example, you may instruct Stata to show 99% confidence intervals instead.

R-Squared

R-Squared (or R2) does not work very well due to the assumptions behind logistic regression. Stata produces a pseudo R2, but due to inherent bias this is seldom used.

Simple versus multiple regression models

The difference between simple and multiple regression models, is that in a multiple regression each x-variable's effect on y is estimated while accounting for the other x-variables' effects on y. We then say that these other x-variables are "held constant", or "adjusted for", or "controlled for". Because of this, multiple regression analysis is a way of dealing with the issue of confounding variables, and to some extent also mediating variables (see Section 9.3).

It is highly advisable to run a simple regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted coefficients with (i.e. what happened to the coefficients when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that they become weaker – which would of course be expected if the x-variables overlapped in their effect on y.

A note

Remember that a regression analysis should follow from theory as well as a comprehensive set of descriptive statistics and knowledge about the data. In the following sections, we will – for the sake of simplicity – not form any elaborate analytical strategy where we distinguish between x-variables and z-variables (see Section 9.3). However, we will define an analytical sample and use a so-called pop variable (see Section 11.5).

13.2 Function

Basic command	logistic depvar indepvars	
Explanations	depvar indepvars	Insert the name of the y-variable. Insert the name of the x-variable(s) that you want to use.
More information	help logistic	

Note The logistic command automatically produces odds ratios. If you, for some reason, want to produce log odds instead, try logit.

A walk-through of the output

When we perform a logistic regression in Stata, the table looks like this:

```
Logistic regression                             Number of obs     =      8,886
                                                LR chi2(2)        =      89.41
                                                Prob > chi2       =     0.0000
Log likelihood = -2694.8097                     Pseudo R2         =     0.0163

------------------------------------------------------------------------------
        yvar | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xvar1 |   .5758398   .0436167    -7.29   0.000     .4963954    .6679986
       xvar2 |   .9695844   .0051825    -5.78   0.000     .9594798    .9797953
       _cons |   .2809438   .0389031    -9.17   0.000     .2141663    .3685427
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

In this example, yvar is a binary (0/1) variable, whereas xvar1 is a binary (0/1) variable and xvar2 is a continuous variable ranging between 1 and 40.

The upper part of the table shows a model summary. This is what the different rows mean:

Row	**Explanation**
Log likelihood	This value does not mean anything in itself, but can be used if we would like compare nested models.
Number of obs	The number of observations included in the model.
LR chi2(x)	The likelihood ratio (LR) chi-square test. The number within the brackets shows the degrees of freedom (one per variable).
Prob >chi2	Shows the probability of obtaining the chi-square statistic given that there is no statistical effect of the x-variables on y. If the p-value is below 0.05, we can conclude that the overall model is statistically significant.
Pseudo R2	A type of R-squared value. Seldom used.

The lower part of the table presents the parameter estimates from the analysis.

Column	**Explanation**
	The first column lists the y-variable on top, followed by our x-variable(s). The last row represents the constant (intercept).
Odds ratio	These are the odds ratios.
Std. Err.	The standard errors associated with the coefficient.
Z	Z-value (coefficient divided by the standard error of the coefficient).
P>|z|	P-value.
[95% Conf. Interval]	95% confidence intervals (lower limit and upper limit).

The analytical sample used for the examples

In the subsequent sections, we will use the following variables:

Dataset: StataData1.dta

Name	**Label**
earlyret	Early retirement (Age 50, Year 2020)
bmi	Body mass index (Age 20, Year 1990)
sex	Sex
educ	Educational level (Age 40, Year 2010)

```
sum earlyret bmi sex educ
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earlyret |      8,773    .1371253    .3439992          0          1
         bmi |      8,385    22.64526     3.50581   10.97624   39.25653
         sex |     10,000       .4892    .4999083          0          1
        educ |      9,183    2.173691    .7263263          1          3
```

We define our analytical sample through the following command:

```
gen pop_logistic=1 if earlyret!=. & bmi!=. & sex!=. & educ!=.
```

This means that new the variable pop_logistic gets the value 1 if the four variables do not have missing information. In this case, we have 7,406 individuals that are included in our analytical sample.

```
tab pop_logistic
```

```
pop_logisti |
          c |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      7,406      100.00      100.00
------------+-----------------------------------
      Total |      7,406      100.00
```

13.3 Simple logistic regression

Quick facts	
Number of variables	One dependent (y) One independent (x)
Scale of variable(s)	Dependent: binary Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

13.3.1 Simple logistic regression with a continuous x

Theoretical examples

Example 1

Suppose we want to examine the association between unemployment days (x) and depression (y) by means of simple logistic regression analysis. Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Depression has the values 0=No and 1=Yes. Let us say that we get an OR that is 1.03. That would mean that we have a positive association: the higher the number of unemployment days, the higher the risk of depression.

Example 2

In another example, we may examine the association between intelligence scores (x) and drug use (y). Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Drug use has the values 0=No and 1=Yes. Here, we get an OR of 0.91. We can thus conclude that the risk of using drugs decreases for every unit increase in intelligence scores.

Practical example

Dataset: StataData1.dta

Name	**Label**
earlyret	Early retirement (Age 50, Year 2020)
bmi	Body mass index (Age 20, Year 1990)

```
sum earlyret bmi if pop_logistic==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earlyret |      7,406    .1259789     .331848          0          1
         bmi |      7,406    22.62361    3.506361   11.11549   39.25653
```

```
logistic earlyret bmi if pop_logistic==1
```

```
Logistic regression                             Number of obs     =      7,406
                                                LR chi2(1)        =       0.27
                                                Prob > chi2       =     0.6024
Log likelihood = -2804.2996                     Pseudo R2         =     0.0000

------------------------------------------------------------------------------
    earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   1.005211   .0100237     0.52   0.602     .9857555    1.02505
       _cons |   .1281317   .0293128    -8.98   0.000     .0818326   .2006257
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

When we look at the results for bmi, we see that the odds ratio (OR) is 1.00 or, more precisely, 1.005211. Thus, one unit increase in bmi does almost not change the odds of earlyret at all.

The association between bmi and earlyret is not statistically significant, as reflected in the p-value (0.60) and the 95% confidence intervals (0.99-1.03).

Summary

There is a positive association between body mass index at age 15 and early retirement at age 50. The association is nonetheless very weak (OR=1.005) and statistically non-significant (95% CI=0.99-1.03).

13.3.2 Simple logistic regression with a binary x

Theoretical examples

Example 1
Suppose we want to examine the association between gender (x) and alcohol abuse (y). Gender has the values 0=Man and 1=Woman, whereas alcohol abuse has the values 0=No and 1=Yes. Now, we get an OR of 0.66. This would mean that women are less likely compared to men to abuse alcohol.

Example 2
Here, we want to examine the association between having young children (x) and owning a pet (y). Having young children is measured as either 0=No young children and 1=Young children. Owning a pet has the values 0=No and 1=Yes. Let us say that we get an OR that is 1.49. We can hereby conclude that it is more common to own a pet in families with young children compared to families without young children.

Practical example

Dataset: StataData1.dta

Name	**Label**
earlyret	Early retirement (Age 50, Year 2020)
sex	Sex

```
sum earlyret sex if pop_logistic==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earlyret |      7,406    .1259789     .331848          0          1
         sex |      7,406    .5213341    .4995784          0          1
```

The variable sex is binary: 0=Man, 1=Woman. When we add it to the model, the category with the lowest value will be the reference category (i.e. Man).

```
logistic earlyret sex if pop_logistic==1
```

```
Logistic regression                             Number of obs     =      7,406
                                                LR chi2(1)        =      55.53
                                                Prob > chi2       =     0.0000
Log likelihood = -2776.6704                     Pseudo R2         =     0.0099

------------------------------------------------------------------------------
    earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.701551   .1231687     7.34   0.000     1.476487    1.960921
       _cons |   .1064295   .0060624   -39.33   0.000     .0951866    .1190003
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

When we look at the results for sex, we see that the odds ratio (OR) is 1.70. Now, it is important to remember the coding of sex: 0=Man, 1=Woman. Thus, a unit increase in sex is the same as being a woman compared to a man. Since men are the reference category, they automatically get the OR 1.00. The specific odds ratio of 1.70 can be interpreted as women having higher odds of earlyret compared to men.

There is a statistically significant association between sex and earlyret, as reflected in the p-value (0.000) and the 95% confidence intervals (1.48-1.96).

Summary

Women are more likely to have experienced early retirement at the age of 50, as compared to men (OR=1.70, 95% CI=1.48 to 1.96).

13.3.3 Simple logistic regression with a categorical (non-binary) x

Theoretical examples

Example 1

We want to investigate the association between educational attainment (x) and divorce (y). Educational attainment has the values: 1=Compulsory, 2=Upper secondary, and 3=University. We choose Compulsory as our reference category. Let us say that we get an OR for upper secondary education that is 0.82 and we get an OR for university education that is 0.69. We can thus conclude – based on the direction of the estimates – that higher levels of educational attainment are associated with a lower risk of divorce.

Example 2

Suppose we are interested in the association between family type (x) and children's average school marks (y). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. We choose Two-parent household as our reference category. Children's average school marks are categorised into 0=Above average and 1=Below average. The analysis results in an OR of 1.02 for Joint custody and an OR of 1.55 for Single-parent household. That would mean that children living in family types other than two-parent households are more likely to have school marks below average.

Practical example

Dataset: StataData1.dta

Name	**Label**
earlyret	Early retirement (Age 50, Year 2020)
educ	Educational level (Age 40, Year 2010)

```
sum earlyret educ if pop_logistic==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earlyret |      7,406    .1259789     .331848          0          1
        educ |      7,406     2.20902    .7122387          1          3
```

The variable educ has three categories: 1=Compulsory, 2=Upper secondary, and 3=University. Here, we (with ib1) specify that the first category (Compulsory) will be the reference category.

```
logistic earlyret ib1.educ if pop_logistic==1
```

```
Logistic regression                             Number of obs     =      7,406
                                                LR chi2(2)        =     138.66
                                                Prob > chi2       =     0.0000
Log likelihood = -2735.1032                     Pseudo R2         =     0.0247

---------------------------------------------------------------------------------
       earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
           educ |
Upper secondary  |   .7062803   .0611141    -4.02   0.000     .5961053    .8368184
     University  |   .3291693   .0334052   -10.95   0.000     .2697966    .4016077
                |
          _cons |   .2399608   .0170711   -20.06   0.000       .20873    .2758645
---------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

When we look at the results for educ, we see two odds ratios: one for Upper secondary and one for University. They are compared to the reference category, which in this case in Compulsory (OR=1.00). The odds ratio for Upper secondary is 0.71, meaning that those with upper secondary education have lower odds of earlyret, compared to those with compulsory education. The odds ratio for University is 0.33, which suggests that these individuals are even less likely to having retired at age 50, compared to those with compulsory education.

The dummies for educ are both significantly different from the reference category, as reflected in the p-value (0.000) and the 95% confidence intervals (0.60-0.84, and 0.27-0.40 respectively).

Test the overall effect

The output presented and interpreted above, is based on the odds ratios for the dummy variables of educ. But what about the overall statistical effect of educ on earlyret? We can assess it through contrast, which is a postestimation command.

```
contrast p.educ, noeffects
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
        educ |
    (linear) |          1      119.89     0.0000
 (quadratic) |          1        8.64     0.0033
       Joint |          2      127.75     0.0000
------------------------------------------------
```

Here, we focus on the row for linear, which shows a p-value (P>chi2) below 0.05. This suggests that we have a statistically significant trend in earlyret according to educ.

More information	help contrast

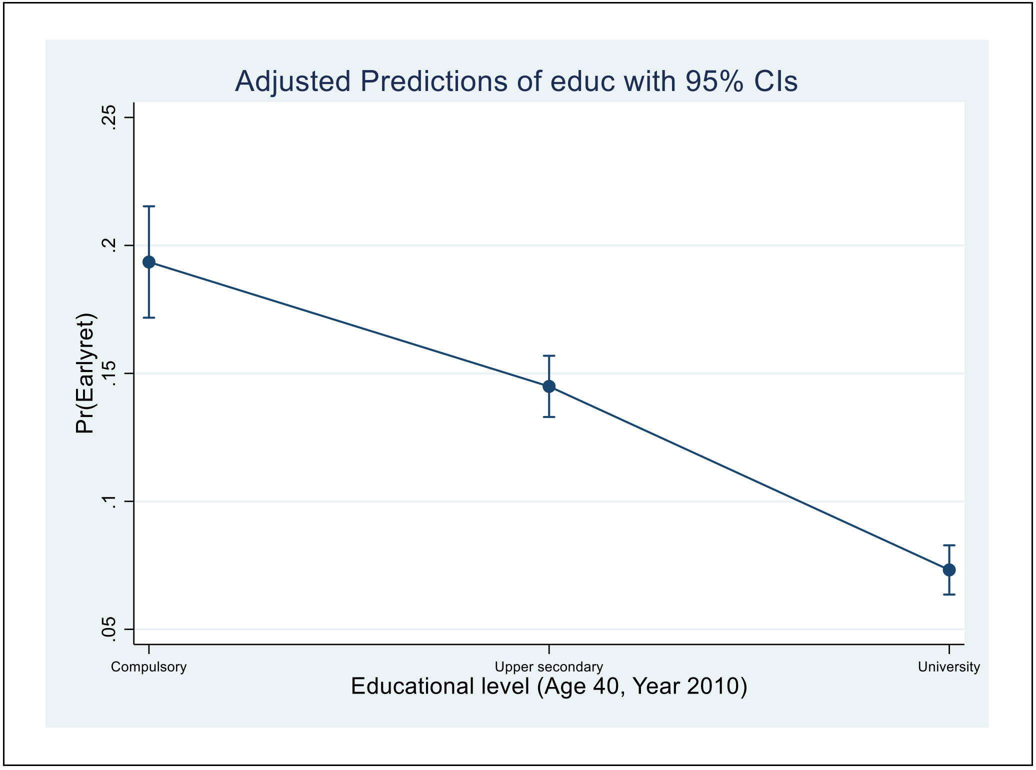
We will also produce a graph of the trend. First, however, we need to apply the post-estimation command margins.

Note This command can also be used for variables that are continuous or binary, but is particularly useful for categorical, non-binary (i.e. ordinal) variables.

```
margins educ
```

```
Adjusted predictions                            Number of obs     =      7,406
Model VCE    : OIM

Expression   : Pr(earlyret), predict()

---------------------------------------------------------------------------------
                 |            Delta-method
                 |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+---------------------------------------------------------------
            educ |
     Compulsory  |   .1935229   .0111031    17.43   0.000     .1717612    .2152846
Upper secondary  |   .1449188   .0061039    23.74   0.000     .1329555    .1568822
     University  |   .0732054   .0049102    14.91   0.000     .0635815    .0828293
---------------------------------------------------------------------------------
```

```
marginsplot
```



This is our marginsplot. A quite clear trend is shown here.

Note The y-axis shows predicted probabilities (i.e. not log odds or odds ratios).

More information	help marginsplot

Summary
There is a negative association between educational level and early retirement; the higher the educational level, the lower the odds of early retirement (Upper secondary vs Compulsory: OR=0.71, 95% CI=0.60-0.84; University vs Compulsory: OR=0.33, 95% CI=0.27-0.40).

13.4 Multiple logistic regression

Quick facts	
Number of variables	One dependent (y) At least two independent (x)
Scale of variable(s)	Dependent: binary Independent: categorical (nominal/ordinal) and/or continuous (ratio/interval)

Example

Suppose we are interested to see if having young children (x), residential area (x), and income (x) are related to owning a pet (y). Having young children is measured as either 0=No young children and 1=Young children. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. We choose Metropolitan as our reference category. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK). Owning a pet has the values 0=No and 1=Yes.

We get an OR for Young children that is 1.30. That means that those who have young children are more likely to also own a pet, compared to those who do not have young children. This association is adjusted for residential area and income.

With regards to residential area, we get an OR for Smaller city of 1.78, whereas the OR for Rural is 4.03. This suggests that those who live in a smaller city are more likely to own a pet, and so are those living in rural areas. These results are adjusted for having young children and income.

Finally, the OR for income is 0.93. This suggests that for every unit increase in income (i.e. for every additional one thousand SEK), the likelihood of owning a pet decreases. This association is adjusted for having young children as well as residential area.

Practical example

Dataset: StataData1.dta

Name	**Label**
earlyret	Early retirement (Age 50, Year 2020)
bmi	Body mass index (Age 20, Year 1990)
sex	Sex
educ	Educational level (Age 40, Year 2010)

```
sum earlyret bmi sex educ if pop_logistic==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earlyret |      7,406    .1259789     .331848          0          1
         bmi |      7,406    22.62361    3.506361   11.11549   39.25653
         sex |      7,406    .5213341    .4995784          0          1
        educ |      7,406     2.20902    .7122387          1          3
```

```
logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

```
Logistic regression                             Number of obs     =      7,406
                                                LR chi2(4)        =     205.06
                                                Prob > chi2       =     0.0000
Log likelihood = -2701.9047                     Pseudo R2         =     0.0366

----------------------------------------------------------------------------------
        earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
             bmi |   1.013217   .0101501     1.31   0.190     .9935174    1.033308
             sex |   1.815529   .1349486     8.02   0.000     1.569398     2.10026
                 |
            educ |
Upper secondary  |   .6751302   .0590041    -4.50   0.000     .5688469    .8012715
     University  |   .3129199   .0320419   -11.35   0.000     .2560195    .3824664
                 |
           _cons |   .1316595   .0330562    -8.08   0.000     .0804893    .2153607
----------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

In this model, we have three x-variables: bmi, sex, and educ. When we put them together, their statistical effect on earlyret is mutually adjusted.

When it comes to the odds ratios, they have changed in comparison to the simple regression models. For example, the odds ratio for bmi has increased slightly from 1.005 to 1.013. The odds ratio for sex is also higher now: 1.82 instead of 1.70. Concerning the categories of educ, we see that the odds ratio for Upper secondary has also become slightly stronger (i.e. become further from 1) from 0.71 to 0.68, and the odds ratio for University is 0.31 instead of 0.33.

The associations between the sex and earlyret on the one hand, and between educ and earlyret on the other hand, are still statistically significant ($p<0.05$) after mutual adjustment. The association between bmi and earlyret remains statistically non-significant.

Note A specific odds ratio from a simple logistic regression model can increase when other x-variables are included. Usually, it is just “noise”, i.e. not any large increases, and therefore not much to be concerned about. But it can also reflect that there is something going on that we need to explore further. There are many possible explanations for increases in multiple regression models: a) We actually adjust for a confounder and then “reveal” the “true” statistical effect. b) There are interactions among the x-variables in their effect on the y-variable. c) There is something called collider bias (which we will not address in this guide) which basically mean that both the x-variable and the y-variable causes another x-variable in the model. d) The simple regression models and the multiple regression model are based on different samples. e) It can be due to rescaling bias (see Chapter 18).

Summary
In the fully adjusted model, it can be observed that odds ratios for body mass index at age 20, sex, and educational level at age 40, with regard to early retirement at age 50, become slightly stronger in comparison to the simple (unadjusted) models.

Estimates table and coefficients plot

If we have multiple models, we can facilitate comparisons between the regression models by asking Stata to construct estimates tables and coefficients plots. What we do is to run the regression models one-by-one, save the estimates after each, and than use the commands estimates table and coefplot.

The coefplot option is not part of the standard Stata program, so unless you already have added this package, you need to install it:

```
ssc install coefplot
```

As an example, we can include the three simple regression models as well as the multiple regression model. The quietly option is included in the beginning of the regression commands to suppress the output.

Run and save the first simple regression model:

```
quietly logistic earlyret bmi if pop_logistic==1
```

```
estimates store model1
```

Run and save the second simple regression model:

```
quietly logistic earlyret sex if pop_logistic==1
```

```
estimates store model2
```

Run and save the third simple regression model:

```
quietly logistic earlyret ib1.educ if pop_logistic==1
```

```
estimates store model3
```

Run and save the multiple regression model:

```
quietly logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

```
estimates store model4
```

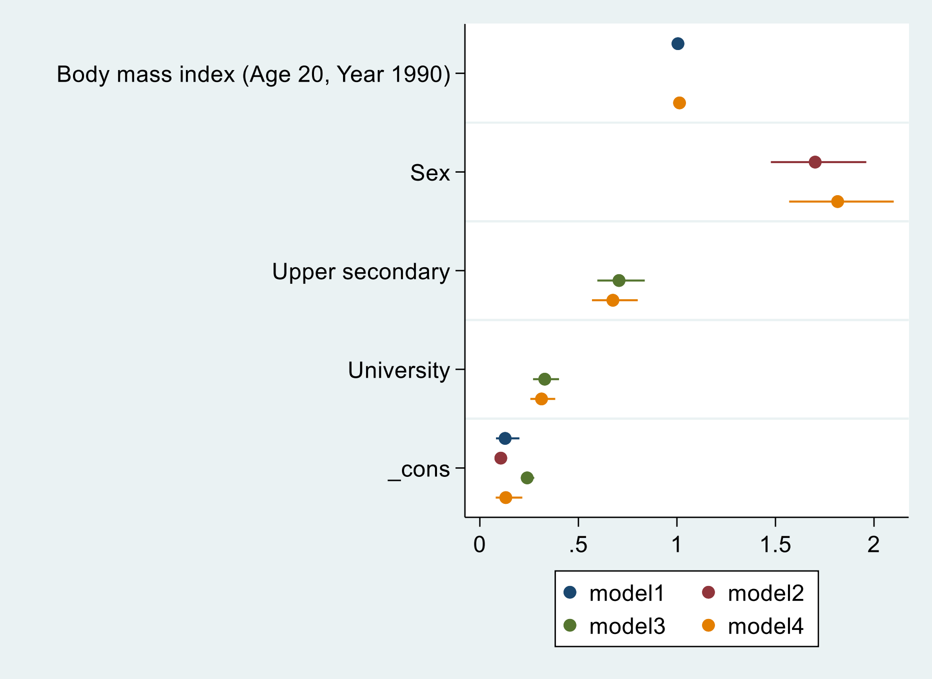
Produce the estimates table (include the option eform to show odds ratios):

```
estimates table model1 model2 model3 model4, eform
```

```
-----------------------------------------------------------------
    Variable |   model1       model2       model3       model4
-------------+---------------------------------------------------
         bmi |  1.0052109                              1.0132172
         sex |               1.7015507                 1.8155285
             |
        educ |
Upper sec..  |                            .70628032    .67513018
 University  |                            .32916928    .31291986
             |
       _cons |  .12813166    .10642947    .23996082    .13165953
-----------------------------------------------------------------
```

Produce the coefficients plot (include the option eform to show odds ratios):

```
coefplot model1 model2 model3 model4, eform
```



Note You can improve the graph by using the Graph Editor to delete "_cons" as well as to adjust the category and label names.

13.5 Model diagnostics

The assumptions behind logistic regression are different from linear regression. For example, we do not need to assume that the effect of the x-variable(s) on y is linear, homoscedasticity or normality.

More information	help logistic postestimation

Checklist	
Binary outcome	The y-variable has to be binary. Also double-check that the proportion of "cases" (or "non-cases", for that matter) is not too small.
Independence of errors	Data should be independent, i.e. not derived from any dependent samples design, e.g. before-after measurements/paired samples.
Correct model specification	Your model should be correctly specified. This means that the x-variables that are included should be meaningful and contribute to the model. No important (confounding) variables should be omitted (often referred to as omitted variable bias).
Linear relationship	There has to be a linear relationship between any continuous x-variable(s) and the log odds of the y-variable (not the same as the linearity assumed for linear regression).
No outliers	Outliers are individuals who do not follow the overall pattern of data. Sometimes referred to as influential observations (however, not all outliers are influential). Only relevant for continuous x-variables.
No multicollinearity	Multicollinearity may occur when two or more x-variables that are included simultaneously in the model are strongly correlated with each another. Actually, this does not violate the assumptions, but is does create greater standard errors which makes it harder to reject the null hypothesis.

Most importantly, the model should fit the data. There are several tests to determine "goodness of fit" or, put differently, if the estimated model (i.e. the model with one or more x-variables) predicts the outcome better than the null model (i.e. a model without any x-variables).

Before going into any specific tests, we need to address the issues of "sensitivity" and "specificity". By comparing the cases and non-cases predicted by the model with the cases and non-cases actually present in the outcome, we can draw a conclusion about the proportion of correctly predicted cases (sensitivity) and the proportion of correctly classified non-cases (specificity).

Sensitivity and specificity			
		Estimated model	
		Non-case	**Case**
"Truth"	**Non-case**	*True negative*	*False positive*
	Case	*False negative*	*True positive*

A general comment about model fit: if the main interest was to identify the best model to predict a certain outcome, that would solely guide which x-variables we put into the analysis. For example, we would exclude x-variables that do not contribute to the model's predictive ability. However, research is typically guided by theory and by the interest of examining associations between variables. If we thus have good theoretical reasons for keeping a certain x-variable or sticking to a certain model, we should most likely do that (but still, the model should not fit the data horribly). Model diagnostics will then be a way of showing others the potential problems with the model we use.

Types of model diagnostics	
Link test	Assess model specification
Box-Tidwell and exponential regression models	Check for linearity
Deviance and leverage	Check for influential observations
Correlation matrix	Check for multicollinearity
The Hosmer and Lemeshow test	Asses goodness of fit
ROC curve	Assess goodness of fit

13.5.1 Link test

With the command linktest, we can assess whether our model is correctly specified. This test uses the linear predicted value (called _hat) and the linear predicted value squared (_hatsq) to rebuild the model. We expect _hat to be statistically significant, and _hatsq to be statistically non-significant. If one or both of these expectations are not met, the model is mis-specified.

However, do not rely too much on this test – remember that you should also use theory and common sense to guide your decisions. It is very seldom relevant to focus on this test if our ambition is to investigate associations (and not to make the best possible prediction of the outcome).

More information	help linktest

Practical example

We perform this test for the full model, so let us go back to the example from the multiple regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

And then we run the test:

```
linktest
```

```
Logistic regression                             Number of obs     =      7,406
                                                LR chi2(2)        =     212.82
                                                Prob > chi2       =     0.0000
Log likelihood = -2698.0244                     Pseudo R2         =     0.0379

------------------------------------------------------------------------------
    earlyret |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |  -.4574558   .5367606    -0.85   0.394    -1.509487    .5945756
      _hatsq |  -.3739188    .136949    -2.73   0.006    -.6423339   -.1055036
       _cons |  -1.330509   .5051462    -2.63   0.008    -2.320578   -.3404411
------------------------------------------------------------------------------
```

Since the p-value for the variable _hat is above 0.05 and the p-value for _hatsq is below 0.05, it means that our model is completely mis-specified. This was not surprising, given our problems with the multiple regression analysis earlier.

We could try to amend this by transforming any of the included variables (e.g. through categorisation, or log transformation), excluding any of the included variables, or adding more variables to the model (other x-variables or e.g. interactions between the included variables).

Of course, this should be explored before we continue to assess model fit – but for the sake of simplicity, we will ignore this problem in the following sections.

13.5.2 Box-Tidwell and exponential regression models

The command boxtid might be helpful in checking for linearity in the effect of any continuous x-variable(s) on the log odds of the y-variable. Note that this can also provide some clues as to why the link test produced such poor results.

This command requires that you install a user-written package first. So, if you have not installed it already, type:

```
ssc install boxtid
```

More information	help boxtid

Practical example

Let us apply boxtid to our multiple regression model:

```
boxtid logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

```
Logistic regression                             Number of obs     =      7,406
                                                LR chi2(5)        =     211.62
                                                Prob > chi2       =     0.0000
Log likelihood = -2698.6253                     Pseudo R2         =     0.0377

----------------------------------------------------------------------------------
        earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
         Ibmi__1 |   1.000026   .0000832     0.31   0.758     .9998625    1.000189
         Ibmi_p1 |   .9999966   .0000672    -0.05   0.960     .9998649    1.000128
             sex |    1.80429   .1334461     7.98   0.000     1.560814    2.085747
                 |
            educ |
Upper secondary  |   .6765092   .0591659    -4.47   0.000     .5699407    .8030042
     University  |   .3135918   .0321475   -11.31   0.000     .2565102    .3833759
                 |
           _cons |   .1728802   .0145507   -20.85   0.000     .1465896    .2038861
----------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
------------------------------------------------------------------------------
bmi          |    .0131306    .0100177     1.31 Nonlin. dev. 6.556   (P = 0.010)
          p1 |    8.513755    3.082357
------------------------------------------------------------------------------
Deviance: 5397.253.
```

The test of non-linearity for our continuous variable bmi is statistically significant (p=0.010), suggesting that the assumption of a linear effect is violated. Although we will not explore this further here, we could consider transformations of this variable (e.g. through categorisation, or log transformation). For example, we could categorise bmi into underweight, normal weight, overweight, and obesity.

13.5.3 Deviance and leverage

We will explore three complementary ways of identifying influential observations. Remember that this is only relevant for the continuous x-variables in our model.

	Explanation	**Rule of thumb**
Standardised Pearson residuals	The relative deviations between the observed and fitted values.	Statistic >+/-2
Deviance residuals	The difference between the maxima of the observed and the fitted log likelihood functions.	Statistic >+/-2
Leverage	How far that an x-variable deviates from its mean.	Statistic >3 times of the average of leverage

More information	help logistic postestimation

Practical example

The first step is to re-run our multiple regression model. The quietly option is included in the beginning of the command to suppress the output.

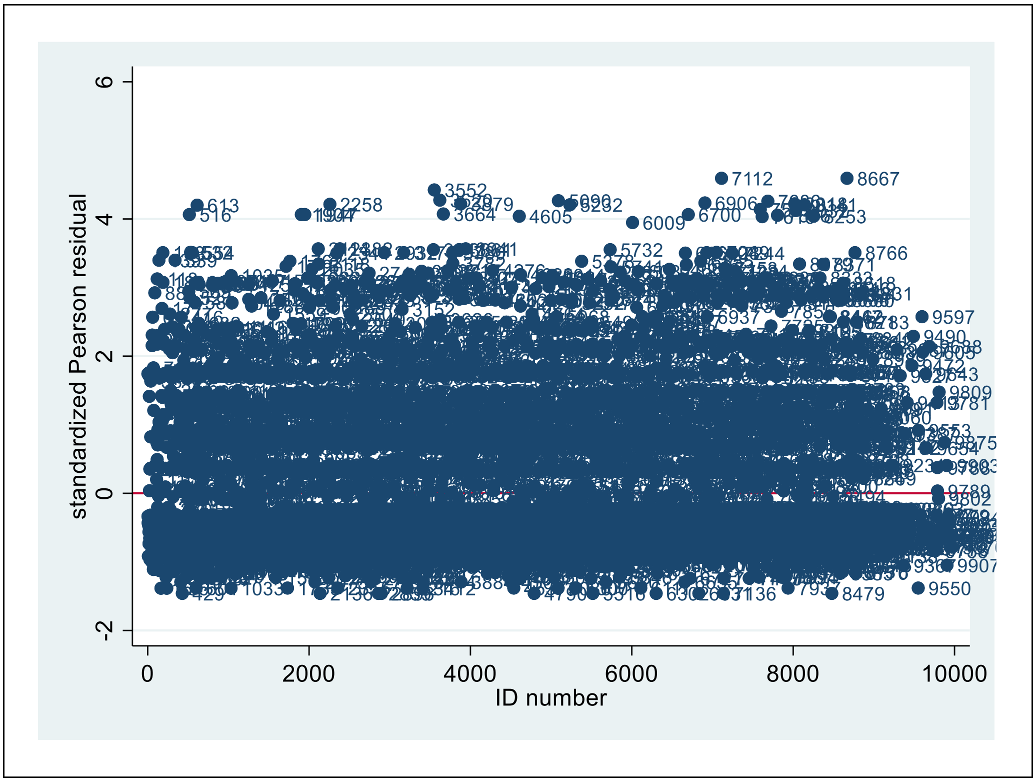
```
quietly logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

Then we generate a new variable – rstandard – that contains the standardised Pearson residuals.

```
predict rstandard, rstandard
```

Next, we generate a scatterplot for rstandard, displaying the id variable on the x-axis. We also include the so-called marker labels (the values of id, in this case), and a regression line at y=0.

```
graph twoway scatter rstandard id, mlab(id) yline(0)
```



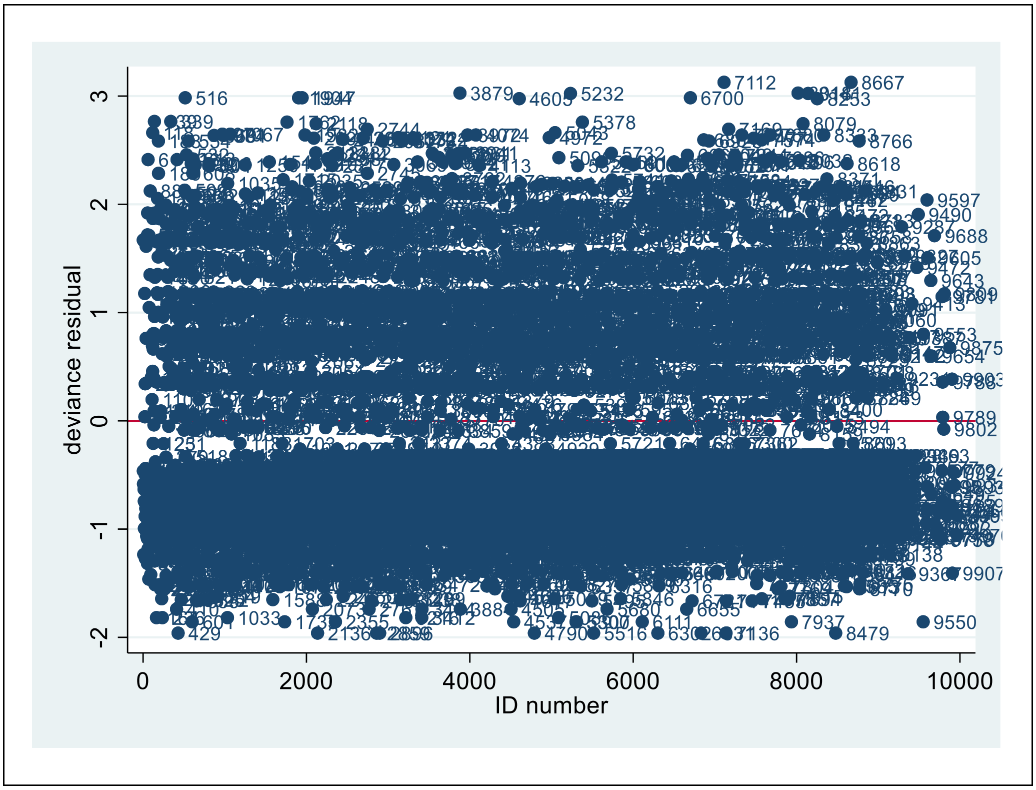
Well, we can see that there are plenty of observations that have residuals greater than 2.

Let us continue with the deviance residuals. We start with generating a new variable – deviance – which contains the deviance residuals.

```
predict deviance, deviance
```

Next, we generate a scatterplot for deviance, displaying the id variable on the x-axis. We also include the so-called marker labels (the values of id, in this case), and a regression line at y=0.

```
graph twoway scatter deviance id, mlab(id) yline(0)
```



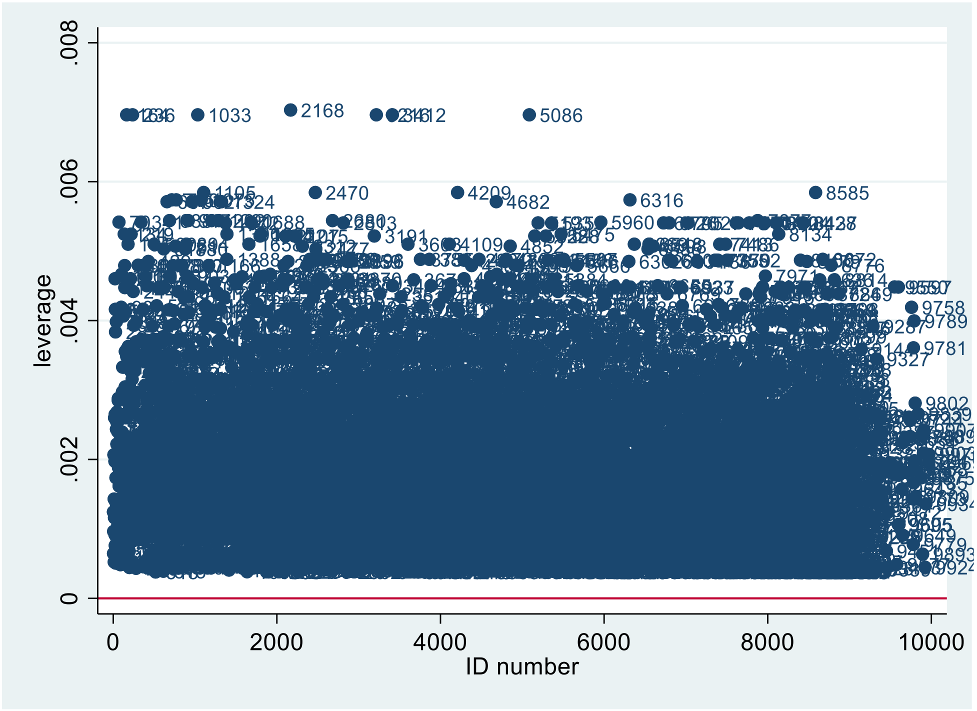
This graph too shows that there are a lot of observations that have higher deviance residuals than we would like (>2).

Next, we consider the leverage. We begin by generating a new variable – hat – which contains the leverage values.

```
predict hat, hat
```

Next, we generate a scatterplot for hat, displaying the id variable on the x-axis. We also include the so-called marker labels (the values of id, in this case), and a regression line at y=0.

```
graph twoway scatter hat id, mlab(id) yline(0)
```



In order to know which observations that display problematic values for leverage, we need to know what the mean leverage is:

```
mean hat
```

```
Mean estimation                   Number of obs   =      7,406

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         hat |   .0017924    .0000123      .0017682    .0018166
--------------------------------------------------------------
```

Mean leverage x 3 (our preferred cut-off value, specified earlier), equals 0.0053772 (i.e. 0.017924 x 3).

There are some observations with higher values than this, but it is a bit tricky to see how many. Let explore this further.

```
sum id if hat>0.0053772 & hat!=.
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |         43    3770.721    2986.881         70       8585
```

We thus have 43 observations with leverage values that might be considered too high. To see their id number, we can order another table (output will be omitted due to how long it gets…):

```
tab id if hat>0.0053772 & hat!=.
```

So, what should we do with all of this information? There as some additional commands that can be used to explore the importance of each (potentially) influential observation further. However, once again, our advice would be to give up on the continuous version of bmi and use a categorised one instead.

13.5.4 Correlation matrix

As the x-variables become more strongly correlated, it becomes more difficult to determine which of the variables are actually producing the statistical effect on the y-variable. This is the problem with multicollinearity.

One way of assessing multicollinearity is using the estat vce command, with the corr (short for correlation) option.

More information	help estat vce

Practical example

The first step is re-run the multiple logistic regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

Next, we try the estat vce command. By adding the corr (=correlation) option, we will get a correlation matrix instead of a covariance matrix.

```
estat vce, corr
```

```
Correlation matrix of coefficients of logistic model

             | earlyret
             |                                 2.         3.
        e(V) |       bmi        sex        educ       educ      _cons
-------------+----------------------------------------------------
earlyret     |
         bmi |    1.0000
         sex |    0.1609     1.0000
      2.educ |    0.0398    -0.0692     1.0000
      3.educ |    0.0438    -0.0725     0.5812     1.0000
       _cons |   -0.9452    -0.3089    -0.2593    -0.2282     1.0000
```

The table shows the correlations between the different variables/categories. In line with the earlier sections on correlation analysis (see Chapter 7.2), we can conclude that the coefficients suggest (very) weak correlations here. The only exceptions are two of the dummies for educ, which is irrelevant since they reflect the same variable.

13.5.5 The Hosmer and Lemeshow test

This test is a type of chi-square test. It indicates the extent to which the estimated model provides a better fit to the data (i.e. has better predictive power) than the null model. The test will produce a p-value: if the p-value is above 0.05 (statistically non-significant) the estimated model has adequate fit, and if the p-value is below 0.05 (statistically significant) the estimated model does not adequately fit the data.

More information	help estat gof

Practical example

Let us first go back to the example from the multiple linear regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

And then we run the test:

```
estat gof
```

```
Logistic model for earlyret, goodness-of-fit test

       number of observations =      7406
 number of covariate patterns =      3782
             Pearson chi2(3777) =      3778.63
                    Prob > chi2 =         0.4895
```

The p-value for the test (Prob > chi2) is above 0.05, suggesting that the estimated model has adequate fit.

13.5.6 ROC curve

The ROC curve is a graph that shows how well the estimated model predicts cases (sensitivity) and non-cases (specificity). What we are interested in here is the "area under the curve" (AUC). The AUC ranges between 0.5 and 1.0. The nearer the AUC is to 1, the better the predictive power. On the other hand, a value of 0.5 suggests that we may just flip a coin to decide on whether the outcome is a case or non-case. Here are some commonly used cut-off points when it comes to AUC:

Area under the curve (AUC)	
0.5-0.6	Fail
0.6-0.7	Poor
0.7-0.8	Fair
0.8-0.9	Good
0.9-1.0	Excellent

More information	help estat

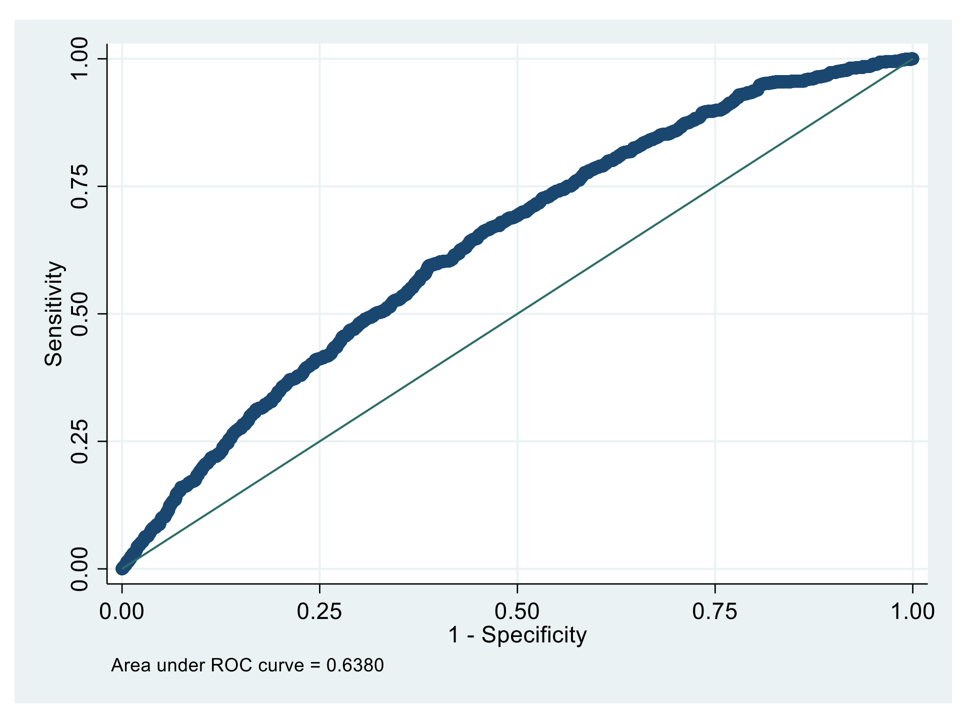
Practical example

Let us first go back to the example from the multiple linear regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly logistic earlyret bmi sex ib1.educ if pop_logistic==1
```

Then we order the ROC curve:

```
lroc
```



The AUC value is 0.64, suggesting that our model has poor predictive power.

13.6 Linear probability modelling

Finally, we would like to make you aware that a viable alternative to the logistic regression model is the linear probability model (LPM). Estimating an LPM means that you enforce a linear regression model (following the instructions in Chapter 12) on your binary outcome. The coefficients (estimates) that are derived from this analysis would then be interpreted as the mean difference in the outcome, i.e. difference in probabilities. The coefficients can thus be interpreted as risk differences.

As long as we are interested in estimating and interpreting associations, and have a strong interest in comparing crude (i.e. unadjusted) and adjusted coefficients between models and/or across samples, the LPM has clear advantages over the logistic regression model. Apart from the fact that we do not have to bother with the interpretation of odds ratios, the potential problem of rescaling bias when we perform mediation analysis (see Chapter 18) is obliterated. We also retain statistical power for interaction analysis (see Chapter 19).

A clear disadvantage with LPM, as highlighted in the introduction of this chapter, is that we might end up with predicted probabilities that are larger than 1 or smaller than 0. This might not be a problem if the goal with our analysis is - as mentioned above - to examine associations rather than making predictions. Another disadvantage is that the interpretations of coefficients for continuos x-variables become problematic (the slope of the linear equation does not approximate well for values at the beginning and the end of the range of values). Consequently, if we have a strong interest in interpreting such associations we need to recode the continuos variables into groups and use dummy variables in our regression. Also, an additional disadvantage is that the error term is not normally distributed, but this is really only a problem with small samples.

14. ORDINAL REGRESSION

Outline

Content

This chapter starts with an introduction to ordinal regression and then presents the function in Stata. After this, we offer some practical examples of how to perform simple and multiple ordinal regression, as well as to generate and interpret model diagnostics.

14.1 Introduction

Ordinal regression is used when y is ordinal. This means that the outcome consists of three or more categories that are possible to rank (i.e. ordered categories; see Section 3.3).

Some examples
Educational level (1=Compulsory; 2=Upper secondary; 3=University)
School marks (1=Low; 2=Average; 3=High)
Self-rated health (1=Excellent; 2=Good; 3=Fair; 4=Poor)
Statement: "Eurovision Song Contest is entertaining" (1=Strongly agree; 2=Agree; 3=Neither agree nor disagree; 4=Disagree; 5=Strongly disagree)

Proportional odds

Ordinal regression is a type of logistic regression that can handle the fact that the outcome has multiple (ordered) outcome categories. Instead of modelling the probability of the outcome being a case, we consider the cumulative probability across the outcome categories. This means that we estimate the odds of being at or above a given threshold across all cumulative splits.

In the model, each outcome category has its own intercept (at each threshold) but the same coefficient for the overall x-variable. Because of this, we have to assume that the effect of x on the odds of the outcome being a case for each subsequent category is the same for every category. This reflects the notion of proportional odds (sometimes referred to as parallel lines), which is a key assumption behind ordinal regression analysis. Put differently, the proportional odds assumption means that the estimate between each pair of outcome categories are assumed to be the same regardless of which pair is considered.

Other names for ordinal regression

Sometimes, ordinal regression analysis is referred to as, e.g., ordered logit regression, ordinal logistic regression, or proportional odds regression.

14.1.1 Ordinal regression in short

If you have only one x, it is called simple regression, and if you have more than one x, it is called multiple regression.

Regardless of whether you are doing a simple or a multiple regression, x-variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from ordinal regression		
Effect		
Odds ratio (OR)	The exponent of log odds	
	Log odds	The logarithm of odds
	Odds	The probability of the outcome being case divided by the probability of the outcome being a non-case
	Probability	The probability of an event happening
Direction		
Negative	OR below 1	
Positive	OR above 1	
Statistical significance		
P-value	p<0.05 Statistically significant at the 5% level p<0.01 Statistically significant at the 1% level p<0.001 Statistically significant at the 0.1% level	
95% Confidence intervals	Interval does not include 1: Statistically significant at the 5% level Interval includes 1: Statistically non-significant at the 5% level	

Odds ratio (OR)

In ordinal regression analysis, the effect that x has on y is reflected by an odds ratio (OR):

OR below 1	For every unit increase in x, the odds of being in a higher ordered category of y decreases.
OR above 1	For every unit increase in x, the odds of being in a higher ordered category of y increases.

Exactly how one interprets the OR in plain writing depends on the measurement scale of the x-variable. That is why we will present examples later for continuous, binary, and categorical (non-binary) x-variables.

Note Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for ordinal regression is 1.

Note An OR can never be negative – it can range between 0 and infinity.

How to *not* interpret odds ratios

Odds ratios are not the same as risk ratios (see Section 4.7.6). ORs tend to be inflated when they are above 1 and understated when they are below 1. This becomes more problematic the more common the outcome is (i.e. the more "cases" we have). However, the rarer the outcome is (<10% is usually considered a reasonable cut-off here), the closer odds ratios and risks ratios become.

Many would find it compelling to interpret ORs in terms of percentages. For example, an OR of 1.20 might lead to the interpretation that the odds of being in a higher ordered category of the outcome increase by 20%. If the OR is 0.80, some would then suggest that the odds of being in a higher ordered category of the outcome decrease by 20%. We would to urge you to carefully reflect upon the latter kind of interpretation since odds ratios are not symmetrical: it can take any value above 1 but cannot be below 0. Thus, the choice of reference category might lead to quite misleading conclusions about effect size. The former kind of interpretation is usually considered reasonable when ORs are below 2. If they are above 2, it is better to refer to "times", i.e. an OR of 4.07 could be interpreted as "more than four times the odds of…".

Take home messages
Do not interpret odds ratios as risk ratios, unless the outcome is rare (<10%, but even then, be careful).
It is completely fine to discuss the results more generally in terms of higher or lower odds/risks. However, if you want to give exact numbers to exemplify, you need to consider the asymmetry of odds ratios as well as the size of the OR.

P-values and confidence intervals

In ordinal regression analysis you can get information about statistical significance, in terms of both p-values and confidence intervals (also see Section 5.2).

Note The p-values and the confidence intervals will give you partly different information, but they are not contradictory. If the p-value is below 0.05, the 95% confidence interval will not include 1 and, if the p-value is above 0.05, the 95% confidence interval will include 1.

When you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5% level, the 1% level, or the 0.1% level).

When it comes to confidence intervals, Stata will by default choose 95% level confidence intervals. It is however possible to change the confidence level for the intervals. For example, you may instruct Stata to show 99% confidence intervals instead.

R-Squared

R-Squared (or R2) does not work very well due to the assumptions behind ordinal regression. Stata produces a pseudo R2, but due to inherent bias this is seldom used.

Simple versus multiple regression models

The difference between simple and multiple regression models, is that in a multiple regression each x-variable's effect on y is estimated while accounting for the other x-variables' effects on y. We then say that these other x-variables are "held constant", or "adjusted for", or "controlled for". Because of this, multiple regression analysis is a way of dealing with the issue of confounding variables, and to some extent also mediating variables (see Section 9.3).

It is highly advisable to run a simple regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted coefficients with (i.e. what happened to the coefficients when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that they become weaker – which would of course be expected if the x-variables overlapped in their effect on y.

A note

Remember that a regression analysis should follow from theory as well as a comprehensive set of descriptive statistics and knowledge about the data. In the following sections, we will – for the sake of simplicity – not form any elaborate analytical strategy where we distinguish between x-variables and z-variables (see Section 9.3). However, we will define an analytical sample and use a so-called pop variable (see Section 11.5).

14.2 Function

Basic command	ologit depvar indepvars	
Useful options	ologit depvar indepvars, or	
Explanations	depvar	Insert the name of the y-variable.
	indepvars	Insert the name of the x-variable(s) that you want to use.
	or	Produces odds ratios.
More information	help ologit	

Note The ologit command produces log odds, unless otherwise specified.

A walk-through of the output

When we perform an ordinal regression in Stata, the table looks like this:

```
Ordered logistic regression                     Number of obs     =      8,291
                                                LR chi2(2)        =     457.54
                                                Prob > chi2       =     0.0000
Log likelihood = -10785.127                     Pseudo R2         =     0.0208

------------------------------------------------------------------------------
        yvar | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xvar1 |     .80466   .0322698    -5.42   0.000     .7438342    .8704597
       xvar2 |   1.065727   .0033413    20.30   0.000     1.059198    1.072296
-------------+----------------------------------------------------------------
       /cut1 |  -.2824731   .0865445                     -.4520973   -.1128489
       /cut2 |   .8467749    .085935                      .6783453    1.015204
       /cut3 |   2.431372   .0897729                      2.255421    2.607324
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

In this example, yvar is an ordinal variable with four categories, whereas xvar1 is a binary (0/1) variable and xvar2 is a continuous variable ranging between 1 and 40.

The upper part of the table shows a model summary. This is what the different rows mean:

Row	Explanation
Log likelihood	This value does not mean anything in itself, but can be used if we would like compare nested models.
Number of obs	The number of observations included in the model.
LR chi2(x)	The likelihood ratio (LR) chi-square test. The number within the brackets shows the degrees of freedom (one per variable).
Prob >chi2	Shows the probability of obtaining the chi-square statistic given that there is no statistical effect of the x-variables on y. If the p-value is below 0.05, we can conclude that the overall model is statistically significant.
Pseudo R2	A type of R-squared value. Seldom used.

The lower part of the table presents the parameter estimates from the analysis.

Column	Explanation
	The first column lists the y-variable on top, followed by our x-variable(s). We also get the cut points for the levels of the y-variable. These are usually not interpreted.
Odds ratio	These are the odds ratios.
Std. Err.	The standard errors associated with the coefficient.
Z	Z-value (coefficient divided by the standard error of the coefficient).
P>|z|	P-value.
[95% Conf. Interval]	95% confidence intervals (lower limit and upper limit).

The analytical sample used for the examples

In the subsequent sections, we will use the following variables:

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)
bullied	Exposure to bullying (Age 15, Year 1985)
bestfriends	Number of best friends (Age 15, Year 1985)

```
sum educ gpa bullied bestfriends
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      9,183    2.173691    .7263263          1          3
         gpa |      9,380    3.178614    .6996298          1          5
     bullied |      8,719    .1076958    .3100137          0          1
 bestfriends |      8,714    2.852536     1.11414          1          5
```

We define our analytical sample through the following command:

```
gen pop_ordinal=1 if educ!=. & gpa!=. & bullied!=. & bestfriends!=.
```

This means that new the variable pop_ordinal gets the value 1 if the four variables do not have missing information. In this case, we have 7,986 individuals that are included in our analytical sample.

```
tab pop_ordinal
```

```
pop_ordinal |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      7,986      100.00      100.00
------------+-----------------------------------
      Total |      7,986      100.00
```

14.3 Simple ordinal regression

Quick facts	
Number of variables	One dependent (y) One independent (x)
Scale of variable(s)	Dependent: ordinal Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

14.3.1 Simple ordinal regression with a continuous x

Theoretical examples

Example 1

Suppose we want to examine the association between unemployment days (x) and self-rated health (y). Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Self-rated health has the values 1=Poor; 2=Fair; and 3=Good. Let us say that we get an OR that is 0.93. That would mean that we have a negative association: the higher the number of unemployment days, the lower the odds of having better health.

Example 2

In another example, we may examine the association between intelligence scores (x) and the number of books read per month (y). Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Book reading has the values 1=0 books; 2=1-3 books; and 3=4 or more books. Here, we get an OR of 1.18. We can thus conclude that a higher intelligence score is associated with more reading of books.

Practical example

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)

```
sum educ gpa if pop_ordinal==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      7,986    2.203231    .7143889          1          3
         gpa |      7,986    3.214425    .6854603        1.1          5
```

```
ologit educ gpa if pop_ordinal==1, or
```

```
Ordered logistic regression                     Number of obs     =      7,986
                                                LR chi2(1)        =    2024.97
                                                Prob > chi2       =     0.0000
Log likelihood = -7227.5609                     Pseudo R2         =     0.1229

------------------------------------------------------------------------------
        educ | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   4.629276    .171564    41.35   0.000     4.304939    4.978049
-------------+----------------------------------------------------------------
       /cut1 |   3.035526   .1118891                      2.816228    3.254825
       /cut2 |   5.531312   .1249445                      5.286425    5.776198
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

When we look at the results for gpa, we see that the odds ratio (OR) is 4.63. In other words, a unit increase in gpa (e.g. going from a grade point average of 2 to 3, or from 4 to 5) is associated with higher educational level.

The association between gpa and educ is statistically significant, as reflected in the p-value (0.000) and the 95% confidence intervals (4.30-4.98).

Summary

There is a positive association between grade point average at age 15 and educational level at age 40 (OR=4.63). The association is statistically significant (95% CI=4.30-4.98).

14.3.2 Simple ordinal regression with a binary x

Theoretical examples

Example 1

Suppose we want to examine the association between gender (x) and educational level (y) by means of a simple ordinal regression analysis. Gender has the values 0=Man and 1=Woman, whereas educational level has the values 1=Low, 2=Medium, and 3=High. Now, we get an OR of 1.62. This would mean that women have higher educational attainment compared to men.

Example 2

Here we want to examine the association between having young children (x) and number of pets (y). Having young children is measured as either 0=No young children and 1=Young children. Number of pets has the values 1=No pet, 2=1-2 pets, and 3=3 or more pets. Let us say that we get an OR that is 1.29. We can hereby conclude that families with young children own more pets than families without young children.

Practical example

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)
bullied	Exposure to bullying (Age 15, Year 1985)

```
sum educ bullied if pop_ordinal==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      7,986    2.203231    .7143889          1          3
     bullied |      7,986    .1033058    .3043769          0          1
```

```
ologit educ bullied if pop_ordinal==1, or
```

```
Ordered logistic regression                     Number of obs     =      7,986
                                                LR chi2(1)        =      23.74
                                                Prob > chi2       =     0.0000
Log likelihood = -8228.1754                     Pseudo R2         =     0.0014

------------------------------------------------------------------------------
        educ | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bullied |   .7140462   .0493383    -4.87   0.000      .623607    .8176014
-------------+----------------------------------------------------------------
       /cut1 |  -1.594357   .0306383                     -1.654407   -1.534307
       /cut2 |   .4674008   .0240525                      .4202587    .5145428
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

When we look at the results for bullied, we see that the odds ratio (OR) is 0.71. Put differently, a unit increase in bullied is associated with lower educational level. This means that those who were exposed to bullying are less likely to reach a higher level of educational attainment.

The association between bullied and educ is statistically significant, as reflected in the p-value (0.000) and the 95% confidence intervals (0.63-0.82).

Summary

Those who were exposed to bullying at age 15 are less likely to reach higher levels of educational attainment at age 40 (OR=0.71; 95% CI=0.62-0.82).

14.3.3 Simple ordinal regression with a categorical (non-binary) x

Theoretical examples

Example 1

We want to investigate the association between educational attainment (x) and happiness (y). Educational attainment has the values: 1=Compulsory, 2=Upper secondary, and 3=University. We choose Compulsory as our reference category. Happiness has the values 1=Happy, 2=Neither happy not unhappy; 3=Unhappy. Let us say that we get an OR for Upper secondary that is 0.87 and we get an OR for University that is 0.66. We can thus conclude that higher educational attainment is associated with less unhappiness (or more happiness).

Example 2

Suppose we are interested in the association between family type (x) and adolescent smoking (y). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. We choose Two-parent household as our reference category. Adolescent smoking has the values 1=No, 2=Occasionally, and 3=Frequently. The analysis results in an OR of 1.33 for Joint custody and an OR of 3.01 for Single-parent household. That would mean that adolescents living in family types other than two-parent households smoke more.

Practical example

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)
bestfriends	Number of best friends (Age 15, Year 1985)

```
sum educ bestfriends if pop_ordinal==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      7,986    2.203231    .7143889          1          3
 bestfriends |      7,986    2.865515    1.109054          1          5
```

The variable bestfriends has five categories: 1=No best friends, 2=One best friend, 3=Two best friends, 4=Three best friends, and 5=Four or more best friends. Here, we (with ib1) specify that the first category (No best friends) will be the reference category.

```
ologit educ ib1.bestfriends if pop_ordinal==1, or
```

```
Ordered logistic regression                     Number of obs     =      7,986
                                                LR chi2(4)        =     259.65
                                                Prob > chi2       =     0.0000
Log likelihood = -8110.2232                     Pseudo R2         =     0.0158

------------------------------------------------------------------------------------
              educ | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
       bestfriends |
   One best friend  |   1.182315   .0859868     2.30   0.021     1.025244    1.363449
  Two best friends  |   1.592136   .1040421     7.12   0.000     1.400736    1.809689
Three best frie..  |     2.2905   .1618419    11.73   0.000      1.99428    2.630719
Four or more be..  |   3.599876   .3881071    11.88   0.000     2.914194    4.446892
-------------------+----------------------------------------------------------------
             /cut1 |   -1.12101   .0573805                     -1.233473   -1.008546
             /cut2 |    .986406   .0569578                      .8747709    1.098041
------------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

When we look at the results for the dummies for bestfriends, we see that the odds ratios range from 1.18 for One best friend, to 3.60 for Four or more best friends. Put differently, having more best friends is associated with higher levels of educational attainment.

All dummies for bestfriends are significantly different from the reference category, as reflected in the p-values and the 95% confidence intervals.

Test the overall effect

The output presented and interpreted above, is based on the odds ratios for the dummy variables of bestfriends. Let us also assess the overall statistical effect of bestfriends on educ? We can assess it through contrast, which is a postestimation command.

```
contrast p.bestfriends, noeffects
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
educ         |
 bestfriends |
   (linear)  |          1      203.83     0.0000
(quadratic)  |          1        7.33     0.0068
    (cubic)  |          1        0.06     0.8038
  (quartic)  |          1        0.06     0.8047
      Joint  |          4      253.38     0.0000
------------------------------------------------
```

Here, we focus on the row for linear, which shows a p-value (P>chi2) below 0.05. This suggests that we have a statistically significant trend in educ according to bestfriends.

More information	help contrast

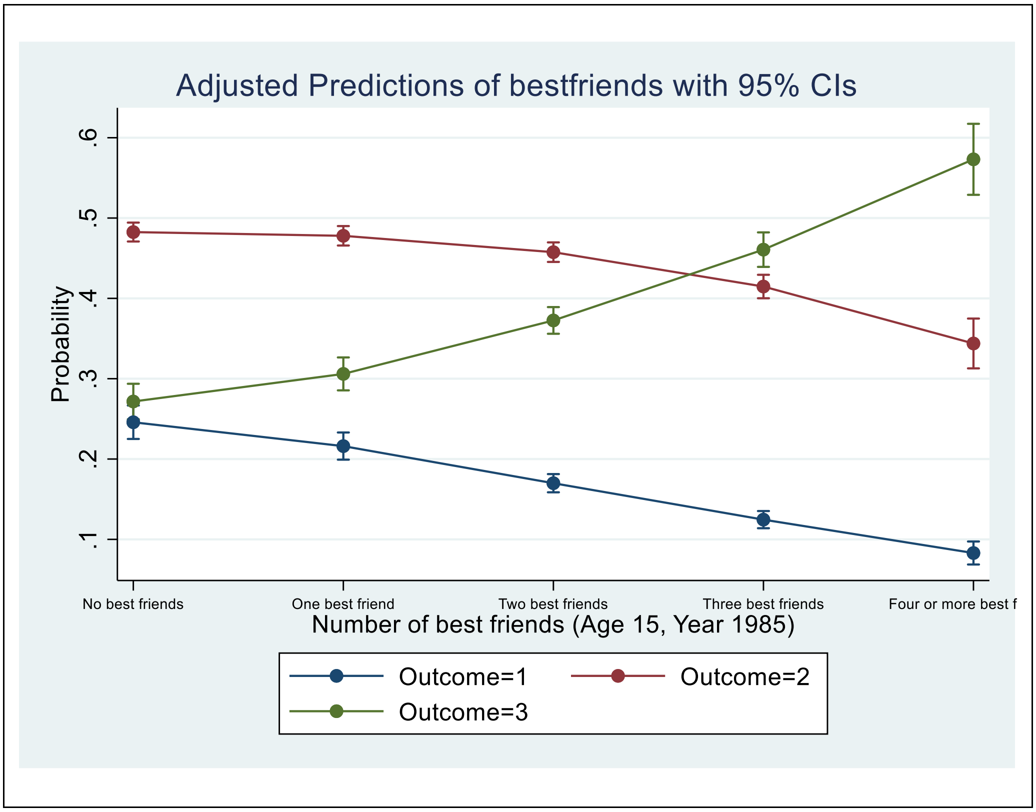
We will also produce a graph of the trend. First, however, we need to apply the post-estimation command margins.

Note This command can also be used for variables that are continuous or binary, but is particularly useful for categorical, non-binary (i.e. ordinal) variables.

```
margins bestfriends
```

```
Adjusted predictions                            Number of obs     =      7,986
Model VCE    : OIM

1._predict   : Pr(educ==1), predict(pr outcome(1))
2._predict   : Pr(educ==2), predict(pr outcome(2))
3._predict   : Pr(educ==3), predict(pr outcome(3))

----------------------------------------------------------------------------------------------
                             |            Delta-method
                             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------------+----------------------------------------------------------------
        _predict#bestfriends |
          1#No best friends  |    .245824    .010638    23.11   0.000     .2249739    .2666742
          1#One best friend  |   .2161095   .0086269    25.05   0.000      .199201     .233018
         1#Two best friends  |   .1699353   .0057761    29.42   0.000     .1586143    .1812562
       1#Three best friends  |   .1245774   .0054542    22.84   0.000     .1138873    .1352675
1#Four or more best friends  |   .0830272   .0073103    11.36   0.000     .0686992    .0973553
          2#No best friends  |   .4825534   .0059961    80.48   0.000     .4708013    .4943055
          2#One best friend  |      .4779   .0061759    77.38   0.000     .4657955    .4900045
         2#Two best friends  |   .4575234   .0062073    73.71   0.000     .4453573    .4696894
       2#Three best friends  |   .4147502   .0074558    55.63   0.000     .4001371    .4293633
2#Four or more best friends  |   .3438769   .0158128    21.75   0.000     .3128844    .3748695
          3#No best friends  |   .2716225   .0112687    24.10   0.000     .2495362    .2937089
          3#One best friend  |   .3059905   .0104606    29.25   0.000     .2854881    .3264928
         3#Two best friends  |   .3725414   .0084756    43.95   0.000     .3559295    .3891532
       3#Three best friends  |   .4606724   .0109558    42.05   0.000     .4391996    .4821453
3#Four or more best friends  |   .5730958   .0225505    25.41   0.000     .5288977     .617294
----------------------------------------------------------------------------------------------
```

```
marginsplot
```



Note The y-axis shows predicted probabilities (i.e. not log odds or odds ratios). The different colours reflect the different levels of the y-variable.

This graph is quite interesting. It shows that the greater the number of best friends, the lower the probabilities of compulsory education, and the higher the probabilities of university education. The trend for upper secondary education is rather unexpected it is quite flat first, and then decreasing.

More information	help marginsplot

Summary
There seem to be a quite clear, and statistically significant, trend in level of educational attainment at age 40 according to the number of best friends at age 15. Having more friends is particularly associated with higher probabilities of university education.

14.4 Multiple ordinal regression

Quick facts	
Number of variables	One dependent (y) At least two independent (x)
Scale of variable(s)	Dependent: ordinal Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

Theoretical examples

Example

Suppose we are interested to see if having young children (x), residential area (x), and income (x) is related to alcohol consumption (y). Having young children is measured as either 0=No young children and 1=Young children. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. We choose Metropolitan as our reference category. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK). Alcohol consumption has the values 1=None/low, 2=Medium, 3=High.

In the regression analysis, we get an OR for Young children that is 0.65. That means that those who have young children drink less alcohol. This association is adjusted for residential area and income.

With regards to residential area, we get an OR for Smaller city of 1.32, whereas the OR for Rural is 2.44. This suggests that those who live in a smaller city drink more alcohol, and so do those living in rural areas. These results are adjusted for having young children and income.

Finally, the OR for income is 0.95. This suggests that for every unit increase in income (i.e. for every additional one thousand SEK), the consumption of alcohol decreases. This association is adjusted for having young children and residential area.

Practical example

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)
bullied	Exposure to bullying (Age 15, Year 1985)
bestfriends	Number of best friends (Age 15, Year 1985)

```
sum educ gpa bullied bestfriends if pop_ordinal==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      7,986    2.203231    .7143889          1          3
         gpa |      7,986    3.214425    .6854603        1.1          5
     bullied |      7,986    .1033058    .3043769          0          1
 bestfriends |      7,986    2.865515    1.109054          1          5
```

```
ologit educ gpa bullied ib1.bestfriends if pop_ordinal==1, or
```

```
Ordered logistic regression                     Number of obs     =      7,986
                                                LR chi2(6)        =    2027.37
                                                Prob > chi2       =     0.0000
Log likelihood = -7226.3627                     Pseudo R2         =     0.1230

--------------------------------------------------------------------------------------------
                       educ | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------------------+---------------------------------------------------------------
                        gpa |   4.675346   .1850569    38.97   0.000     4.326353    5.052491
                    bullied |   .9650606   .0901579    -0.38   0.703     .8035884    1.158979
                            |
                bestfriends |
            One best friend  |   1.035933   .0916915     0.40   0.690     .8709449    1.232175
           Two best friends  |   1.047919   .0901688     0.54   0.586     .8852896    1.240423
         Three best friends  |   .9814308    .091428    -0.20   0.841     .8176427    1.178029
  Four or more best friends  |   .9443533   .1222976    -0.44   0.658     .7326566    1.217218
----------------------------+---------------------------------------------------------------
                      /cut1 |   3.079316   .1299124                      2.824693     3.33394
                      /cut2 |    5.57622   .1415543                      5.298779    5.853662
--------------------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

In this model, we have three x-variables: gpa, bullied, and bestfriends. When we put them together, their statistical effect on educ is mutually adjusted.

When it comes to the odds ratios, they have changed in comparison to the simple regression models. For example, the odds ratio for gpa has increased from 4.63 to 4.68 – however, this is really a minor change. The odds ratio for bullied has become close to 1 (from 0.71 to 0.97). Concerning the dummies of bestfriends, we see that all odds ratios are more or less around 1.

The association between the gpa and educ remains statistically significant (p<0.05) after mutual adjustment. The associations between bullied and educ on the one hand, and between bestfriends and educ on the other hand, have now reached statistically non-significant levels.

Note A specific odds ratio from a simple ordinal regression model can increase when other x-variables are included. Usually, it is just “noise”, i.e. not any large increases, and therefore not much to be concerned about. But it can also reflect that there is something going on that we need to explore further. There are many possible explanations for increases in multiple regression models: a) We actually adjust for a confounder and then “reveal” the “true” statistical effect. b) There are interactions among the x-variables in their effect on the y-variable. c) There is something called collider bias (which we will not address in this guide) which basically mean that both the x-variable and the y-variable causes another x-variable in the model. d) The simple regression models and the multiple regression model are based on different samples. e) It can be due to rescaling bias (see Chapter 18).

Summary

In the fully adjusted model, it can be observed that the association between grade point average at age 15 and the level of educational attainment at age 40, remains strong and statistically significant (OR=4.68; 95% CI=4.33-5.05). Exposure to bullying and number of best friends are, however, no longer associated with the outcome.

Estimates table and coefficients plot

If we have multiple models, we can facilitate comparisons between the regression models by asking Stata to construct estimates tables and coefficients plots. What we do is to run the regression models one-by-one, save the estimates after each, and than use the commands estimates table and coefplot.

The coefplot option is not part of the standard Stata program, so unless you already have added this package, you need to install it:

```
ssc install coefplot
```

As an example, we can include the three simple regression models as well as the multiple regression model. The quietly option is included in the beginning of the regression commands to suppress the output.

Run and save the first simple regression model:

```
quietly ologit educ gpa if pop_ordinal==1, or
```

```
estimates store model1
```

Run and save the second simple regression model:

```
quietly ologit educ bullied if pop_ordinal==1, or
```

```
estimates store model2
```

Run and save the third simple regression model:

```
quietly ologit educ ib1.bestfriends if pop_ordinal==1, or
```

```
estimates store model3
```

Run and save the multiple regression model:

```
quietly ologit educ gpa bullied ib1.bestfriends if pop_ordinal==1, or
```

```
estimates store model4
```

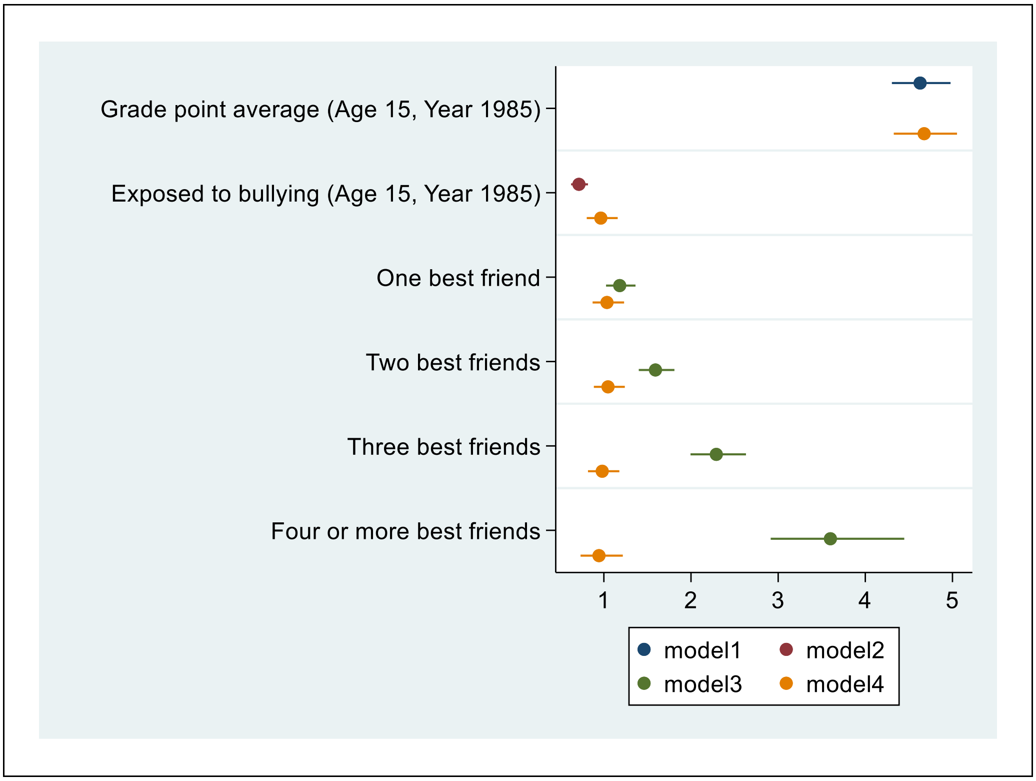
Produce the estimates table (include the option eform to show odds ratios):

```
estimates table model1 model2 model3 model4, eform
```

```
------------------------------------------------------------------
    Variable |   model1       model2       model3       model4
-------------+----------------------------------------------------
educ         |
          gpa |  4.6292758                                 4.6753457
      bullied |               .71404623                  .96506061
             |
 bestfriends |
One best ..  |                             1.182315    1.0359326
Two best ..  |                            1.5921357    1.0479185
Three bes..  |                               2.2905     .9814308
Four or m..  |                            3.5998758    .94435326
-------------+----------------------------------------------------
        /cut1 |  20.811928    .20303902    .32595053     21.74353
        /cut2 |  252.47482    1.5958408    2.6815796    264.07166
------------------------------------------------------------------
```

Produce the coefficients plot (include the option eform to show odds ratios):

```
coefplot model1 model2 model3 model4, eform
```



Note You can improve the graph by using the Graph Editor to adjust the category and label names.

14.5 Model diagnostics

The assumptions behind ordinal regression are similar to the ones for logistic regression.

More information	help ologit postestimation

Checklist	
Ordinal outcome	The y-variable has to be ordinal.
Independence of errors	Data should be independent, i.e. not derived from any dependent samples design, e.g. before-after measurements/paired samples.
Correct model specification	Your model should be correctly specified. This means that the x-variables that are included should be meaningful and contribute to the model. No important (confounding) variables should be omitted (often referred to as omitted variable bias).
No multicollinearity	Multicollinearity may occur when two or more x-variables that are included simultaneously in the model are strongly correlated with each another. Actually, this does not violate the assumptions, but is does create greater standard errors which makes it harder to reject the null hypothesis.
Parallel lines/Proportional odds	The relationship between each pair of outcome groups is the same, i.e. the coefficients that describe the relationship between, for example, the lowest versus all higher categories of the outcome variable are the same as those that describe the relationship between the next lowest category and all higher categories, and so on.

Types of model diagnostics	
Link test	Assess model specification
Correlation matrix	Check for multicollinearity
Brant test	Check parallel lines/proportional odds assumption

14.5.1 Link test

With the command linktest, we can assess whether our model is correctly specified. This test uses the linear predicted value (called _hat) and the linear predicted value squared (_hatsq) to rebuild the model. We expect _hat to be statistically significant, and _hatsq to be statistically non-significant. If one or both of these expectations are not met, the model is mis-specified.

However, do not rely too much on this test – remember that you should also use theory and common sense to guide your decisions. It is very seldom relevant to focus on this test if our ambition is to investigate associations (and not to make the best possible prediction of the outcome).

More information	help linktest

Practical example

We perform this test for the full model, so let us go back to the example from the multiple regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly ologit educ gpa bullied ib1.bestfriends if pop_ordinal==1
```

And then we run the test:

```
linktest
```

```
Ordered logistic regression                     Number of obs     =      7,986
                                                LR chi2(2)        =    2051.87
                                                Prob > chi2       =     0.0000
Log likelihood = -7214.1101                     Pseudo R2         =     0.1245

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .1419645   .1752914     0.81   0.418    -.2016003    .4855293
      _hatsq |   .0880618    .017913     4.92   0.000      .052953    .1231705
-------------+----------------------------------------------------------------
       /cut1 |   1.082786   .4177953                       .263922    1.901649
       /cut2 |   3.568574   .4231193                      2.739276    4.397873
------------------------------------------------------------------------------
```

Since the p-value for the variable _hat is above 0.05 and the p-value for _hatsq is below 0.05, it means that our model is mis-specified.

We could try to amend this by transforming any of the included variables (e.g. through categorisation, or log transformation), excluding any of the included variables, or adding more variables to the model (other x-variables or e.g. interactions between the included variables).

Of course, this should be explored before we continue to assess model fit – but for the sake of simplicity, we will ignore this problem in the following sections.

14.5.2 Correlation matrix

As the x-variables become more strongly correlated, it becomes more difficult to determine which of the variables are actually producing the statistical effect on the y-variable. This is the problem with multicollinearity.

One way of assessing multicollinearity is using the estat vce command, with the corr (short for correlation) option.

More information	help estat vce

Practical example

The first step is re-run the multiple ordinal regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly ologit educ gpa bullied ib1.bestfriends if pop_ordinal==1
```

Next, we try the estat vce command. By adding the corr (=correlation) option, we will get a correlation matrix instead of a covariance matrix.

```
estat vce, corr
```

```
Correlation matrix of coefficients of ologit model

             | educ                                                       | /
             |                            2.        3.        4.        5.|
        e(V) |      gpa   bullied  bestfr~s  bestfr~s  bestfr~s  bestfr~s |     cut1      cut2
-------------+------------------------------------------------------------+--------------------
educ         |                                                            |
         gpa |   1.0000                                                   |
     bullied |  -0.0581    1.0000                                         |
2.bestfrie~s |  -0.0607    0.5189    1.0000                               |
3.bestfrie~s |  -0.1409    0.6053    0.7457    1.0000                     |
4.bestfrie~s |  -0.2393    0.5757    0.7015    0.7811    1.0000           |
5.bestfrie~s |  -0.2642    0.4171    0.5097    0.5750    0.5672    1.0000 |
-------------+------------------------------------------------------------+--------------------
/            |                                                            |
        cut1 |   0.7919    0.3552    0.4390    0.4118    0.2933    0.1334 |   1.0000
        cut2 |   0.8264    0.3250    0.4052    0.3812    0.2700    0.1207 |   0.9645    1.0000
```

The table shows the correlations between the different variables/categories. In line with the earlier sections on correlation analysis (see Chapter 7.2), we can conclude that the coefficients suggest weak to moderate correlations here (with the exception of the dummies of bestfriends, which is not a problem since they reflect the same underlying variable).

14.5.3 Brant test

One critical thing that we need to consider is called the proportional odds assumption (or parallel lines assumption). This means that the coefficients that describe the relationship between, for example, the lowest versus all higher categories of the outcome variable are the same as those that describe the relationship between the next lowest category and all higher categories, and so on. Because the relationships between all pairs of categories are assumed to be the same, we only get one estimate for each x-variable.

For this purpose, we can use the brant command. This command requires that you install a user-written package first. So, if you have not installed it already, type:

```
ssc install spost13
```

More information	help brant

Practical example

The first step is re-run the multiple ordinal regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly ologit educ gpa bullied ib1.bestfriends if pop_ordinal==1
```

Then we use the brant command, which produces the following output:

```
brant, detail
```

```
Estimated coefficients from binary logits

------------------------------------
    Variable |   y_gt_1     y_gt_2
-------------+----------------------
         gpa |    1.347      1.699
             |    24.34      34.81
     bullied |   -0.015     -0.045
             |    -0.12      -0.38
             |
 bestfriends |
One best ..  |    0.104     -0.009
             |     0.92      -0.08
Two best ..  |    0.121      0.003
             |     1.09       0.03
Three bes..  |    0.058     -0.088
             |     0.46      -0.77
Four or m..  |    0.209     -0.200
             |     0.97      -1.35
             |
       _cons |   -2.603     -6.057
             |   -15.09     -34.10
------------------------------------
                         legend: b/t

Brant test of parallel regression assumption

                |       chi2     p>chi2      df
 ---------------+------------------------------
            All |      31.32      0.000       6
 ---------------+------------------------------
            gpa |      29.89      0.000       1
        bullied |       0.04      0.835       1
  2.bestfriends |       0.69      0.407       1
  3.bestfriends |       0.80      0.373       1
  4.bestfriends |       1.02      0.312       1
  5.bestfriends |       3.21      0.073       1

A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

We can see from the test that we have an occurrence of a significant test statistic: gpa has a p-value (p>chi2) which is below 0.05 (0.000), thus violating the proportional odds assumption. While we will not explore this matter further here, a possible solution would be to transform this variable (e.g. through categorisation, or log transformation), or to choose some other type of analysis (e.g. transforming the outcome into a binary version and conduct logistic regression instead).

Additional alternatives

If the proportional odds assumption is violated, it might be interesting to explore other alternatives to ologit. Among these, we have the gologit2 and omodel commands. Both of them are user-written packages.

15. MULTINOMIAL REGRESSION

Outline

Content

This chapter starts with an introduction to multinomial regression and then present the function in Stata. After this, we offer some practical examples of how to perform simple and multiple multinomial regression, as well as how to generate and interpret model diagnostics.

15.1 Introduction

Multinomial regression is used when y is nominal with more than two categories, i.e. polytomous (see Section 3.3). However, it is a good idea not to have too many categories because the interpretation quickly gets quite messy (if you have more than 5-6, try to collapse some of the categories).

Multinomial regression analysis can be seen as an extension of logistic regression. The most complicated part about the multinomial regression is that we decide on a reference category in the outcome variable as well (for linear, logistic and ordinal regression, we only had to deal with reference categories for the x-variables). To make it easier to distinguish between reference categories in x on the one hand, and in y on the other hand, this chapter will use the term reference category when x-variables are concerned, but use base outcome with regard to the y-variable.

Our outcome should have a base outcome – what is that? Let us use an example:

Example
We want to investigate the association between gender (x) and preferred ice-cream flavour (y). Gender has the values 0=Man and 1=Woman. Preferred ice-cream flavour has the values: 1=Vanilla, 2=Chocolate, 3=Strawberry. We choose the first category (vanilla) as our base outcome. When we run the multinomial regression analysis, we will get two relative risk ratios; one for the risk of the outcome being chocolate instead of vanilla depending on the values of the x-variable, and one for the outcome being strawberry instead of vanilla depending on the values of the x-variable.

Other names for multinomial regression

Multinomial regression analysis is also called multinomial logistic regression.

15.1.1 Multinomial regression in short

If you have only one x, it is called simple regression, and if you have more than one x, it is called multiple regression.

Regardless of whether you are doing a simple or a multiple regression, x-variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from multinomial regression		
Effect		
Relative risk ratio (RRR)	The exponent of log relative risk	
	Log relative risk	The logarithm of relative risk
	Relative risk	The probability of the outcome being case divided by the probability of the outcome being a non-case
	Probability	The probability of an event happening
Direction		
Negative	RRR below 1	
Positive	RRR above 1	
Statistical significance		
P-value	p<0.05 Statistically significant at the 5% level p<0.01 Statistically significant at the 1% level p<0.001 Statistically significant at the 0.1% level	
95% Confidence intervals	Interval does not include 1: Statistically significant at the 5% level Interval includes 1: Statistically non-significant at the 5% level	

Relative risk ratio (RRR)

In multinomial regression analysis, the effect that x has on y is reflected by a relative risk ratio (RRR):

RRR below 1	For every unit increase in x, the relative risk of ending up in a certain category of y, compared to the base outcome, decreases.
RRR above 1	For every unit increase in x, the relative risk of ending up in a certain category of y, compared to the base outcome, increases.

A relative risk ratio is not an odds ratio – it is rather a relative odds ratio since it estimates the risk relative to the base outcome. For the sake of simplicity, we will leave it at that.

Exactly how one interprets the RRR in plain writing depends on the measurement scale of the x-variable. That is why we will present examples later for continuous, binary, and categorical (non-binary) x-variables.

Note Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for multinomial regression is 1.

Note An RRR can never be negative – it can range between 0 and infinity.

How to *not* interpret relative risk ratios

The relative risk ratios produced with multinomial regression analysis are not the same as risk ratios (see Section 4.7.6). RRRs tend to be inflated when they are above 1 and understated when they are below 1. This becomes more problematic the more common the outcome is (i.e. the more "cases" we have). However, the rarer the outcome is (<10% is usually considered a reasonable cut-off here), the closer odds ratios and risks ratios become.

Many would find it compelling to interpret RRRs in terms of percentages. For example, an RRR of 1.20 might lead to the interpretation that the relative risk of ending up in a certain category of the outcome, instead of the base outcome, increase by 20%. If the RRR is 0.80, some would then suggest that the relative risk of ending up in a certain category of the outcome, instead of the base outcome, decrease by 20%. We would to urge you to carefully reflect upon the latter kind of interpretation since relative risk ratios are not symmetrical: it can take any value above 1 but cannot be below 0. Thus, the choice of reference category might lead to quite misleading conclusions about effect size. The former kind of interpretation is usually considered reasonable when RRRs are below 2. If they are above 2, it is better to refer to "times", i.e. an RRR of 4.07 could be interpreted as "more than four times the relative risk of…".

Take home messages
Do not interpret relative risk ratios as risk ratios, unless the outcome is very rare (<10%, but even then, be careful).
It is completely fine to discuss the results more generally in terms of higher or lower relative risks/risks. However, if you want to give exact numbers to exemplify, you need to consider the asymmetry of relative risk ratios as well as the size of the RRR.

P-values and confidence intervals

In multinomial regression analysis you can get information about statistical significance, in terms of both p-values and confidence intervals (also see Section 5.2).

Note The p-values and the confidence intervals will give you partly different information, but they are not contradictory. If the p-value is below 0.05, the 95% confidence interval will not include 1 and, if the p-value is above 0.05, the 95%

confidence interval will include 1.

When you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5% level, the 1% level, or the 0.1% level).

When it comes to confidence intervals, Stata will by default choose 95% level confidence intervals. It is however possible to change the confidence level for the intervals. For example, you may instruct Stata to show 99% confidence intervals instead.

R-Squared

R-Squared (or R2) does not work very well due to the assumptions behind multinomial regression. Stata produces a pseudo R2, but due to inherent bias this is seldom used.

Simple versus multiple regression models

The difference between simple and multiple regression models, is that in a multiple regression each x-variable's effect on y is estimated while accounting for the other x-variables' effects on y. We then say that these other x-variables are "held constant", or "adjusted for", or "controlled for". Because of this, multiple regression analysis is a way of dealing with the issue of confounding variables, and to some extent also mediating variables (see Section 9.3).

It is highly advisable to run a simple regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted coefficients with (i.e. what happened to the coefficients when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that they become weaker – which would of course be expected if the x-variables overlapped in their effect on y.

15.2 Function

Basic command	mlogit depvar indepvars	
Useful options	mlogit depvar indepvars, rrr mlogit depvar indepvars, rrr b(x)	
Explanations	depvar indepvars rrr b(x)	Insert the name of the y-variable. Insert the name of the x-variable(s) that you want to use. Produces relative risk ratios. Specify the value of the base outcome. By default, the category with the most observations is chosen.
Short names	b	Base outcome
More information	help mlogit	

Note The mlogit command produces log relative risk, unless otherwise specified.

A walk-through of the output

When we perform an ordinal regression in Stata, the table looks like this:

```
Multinomial logistic regression                 Number of obs     =      8,236
                                                LR chi2(4)        =    1441.30
                                                Prob > chi2       =     0.0000
Log likelihood = -7843.4583                     Pseudo R2         =     0.0841

------------------------------------------------------------------------------
        yvar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
       xvar1 |   .4091507   .0624971     6.55   0.000     .2866587    .5316428
       xvar2 |   .0594214   .0047894    12.41   0.000     .0500344    .0688084
       _cons |  -.6833348   .1172279    -5.83   0.000    -.9130974   -.4535723
-------------+----------------------------------------------------------------
3            |
       xvar1 |   .4501309   .0682677     6.59   0.000     .3163288     .583933
       xvar2 |   .1780728   .0056612    31.45   0.000      .166977    .1891686
       _cons |  -4.067188   .1494159   -27.22   0.000    -4.360038   -3.774338
------------------------------------------------------------------------------
```

In this example, yvar is a nominal variable with three categories, whereas xvar1 is a binary (0/1) variable and xvar2 is a continuous variable ranging between 1 and 40.

The upper part of the table shows a model summary. This is what the different rows mean:

Row	**Explanation**
Log likelihood	This value does not mean anything in itself, but can be used if we would like compare nested models.
Number of obs	The number of observations included in the model.
LR chi2(x)	The likelihood ratio (LR) chi-square test. The number within the brackets shows the degrees of freedom (one per variable).
Prob >chi2	Shows the probability of obtaining the chi-square statistic given that there is no statistical effect of the x-variables on y. If the p-value is below 0.05, we can conclude that the overall model is statistically significant.
Pseudo R2	A type of R-squared value. Seldom used.

The lower part of the table presents the parameter estimates from the analysis.

Column	**Explanation**
	The first column lists the y-variable on top, followed by our x-variable(s). We get one set of x-variables per level of the y-variable (always in comparison to the base outcome).
RRR	These are the relative risk ratios.
Std. Err.	The standard errors associated with the coefficient.
Z	Z-value (coefficient divided by the standard error of the coefficient).
P>|z|	P-value.
[95% Conf. Interval]	95% confidence intervals (lower limit and upper limit).

The analytical sample used for the examples

In the subsequent sections, we will use the following variables:

Dataset: StataData1.dta

Name	**Label**
marstat40	Marital status (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)
sex	Sex
educ	Educational level (Age 40, Year 2010)

```
sum marstat40 gpa sex educ
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   marstat40 |      8,950     1.69933    .8147083          1          4
         gpa |      9,380    3.178614    .6996298          1          5
         sex |     10,000       .4892    .4999083          0          1
        educ |      9,183    2.173691    .7263263          1          3
```

We define our analytical sample through the following command:

```
gen pop_multinom=1 if marstat40!=. & gpa!=. & sex!=. & educ!=.
```

This means that new the variable pop_multinom gets the value 1 if the four variables do not have missing information. In this case, we have 8,409 individuals that are included in our analytical sample.

```
tab pop_multinom
```

```
pop_multino |
          m |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      8,409      100.00      100.00
------------+-----------------------------------
      Total |      8,409      100.00
```

15.3 Simple multinomial regression

Quick facts	
Number of variables	One dependent (y) One independent (x)
Scale of variable(s)	Dependent: nominal (with more than two categories) Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

15.3.1 Simple multinomial regression with a continuous x

Theoretical examples

Example 1

Suppose we want to examine the association between unemployment days (x) and type of health care visit (y). Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Type of health care visit has the values 1=No health care visit, 2=Out-patient care, and 3=In-patient care. We choose No health care visit as our base outcome. Let us say that we get an RRR for unemployment days and Out-patient care that is 2.88. That would mean that for every unit increase of employment days, the risk of experiencing out-patient care compared to having had no health care visit increases. Moreover, we get an RRR for unemployment days and In-patient care that is 4.02. This would suggest that for every unit increase of employment days, the risk of experiencing in-patient care compared to having had no health care visit increases.

Example 2

In another example, we examine the association between intelligence scores (x) and the preferred type of books (y). Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Preferred type of books has the values 1=Fiction, 2=Non-fiction, 3=Comic books. We choose Fiction as our base outcome. Here, we get an RRR of 1.40 for intelligence scores and Non-fiction, meaning that for every unit increase of intelligence, the likelihood of preferring non-fiction books over fiction books increases. For intelligence scores and Comic books, the RRR is 0.92. This suggests that for every unit increase of intelligence, the likelihood of preferring comic books over fiction books decreases.

Practical example

Dataset: StataData1.dta

Name	**Label**
marstat40	Marital status (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)

```
sum marstat40 gpa if pop_multinom==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   marstat40 |      8,409     1.69378    .8148308          1          4
         gpa |      8,409    3.184861    .6931467          1          5
```

```
mlogit marstat40 gpa if pop_multinom==1, rrr b(1)
```

```
Multinomial logistic regression                 Number of obs     =      8,409
                                                LR chi2(3)        =     109.57
                                                Prob > chi2       =     0.0000
Log likelihood = -8840.1483                     Pseudo R2         =     0.0062

------------------------------------------------------------------------------
   marstat40 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Married      |  (base outcome)
-------------+----------------------------------------------------------------
Unmarried    |
         gpa |   .7239121   .0273391    -8.55   0.000     .6722636    .7795287
       _cons |   1.451043   .1770005     3.05   0.002     1.142482     1.84294
-------------+----------------------------------------------------------------
Divorced     |
         gpa |   .7156152     .03019    -7.93   0.000     .6588241    .7773017
       _cons |    1.09113   .1480386     0.64   0.520     .8363535    1.423517
-------------+----------------------------------------------------------------
Widowed      |
         gpa |   1.189674   .1938083     1.07   0.286     .8644893    1.637179
       _cons |   .0104002   .0057745    -8.22   0.000     .0035029    .0308786
------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

When we look at the results for gpa, we see that the relative risk ratio (RRR) is 0.72 for Unmarried, 0.72 for Divorced, and 1.19 for Widowed. This means that the higher the gpa, the lower the risk of being unmarried or divorced, but the higher the risk of being widowed, as compared to being married.

There are statistically significant differences between Married and Unmarried, and between Married and Divorced, according to gpa – as reflected in the p-values (0.000) and the 95% confidence intervals (0.67-0.78 and 0.66-0.78, respectively). The difference between Married and widowed is not statistically significant (p=0.29 and 95% CI=0.86-1.64).

Summary
At age 40, individuals who had higher grade point average at age 15 are less likely to be unmarried (RRR=0.72, 95% CI=0.67-0.78) or divorced (RRR=0.72, 95% CI=0.66-0.78), in comparison to being married. No significant differences in being widowed versus married according to grade point average, were found (RRR=1.19, 95% CI=0.86-1.64).

15.3.2 Simple multinomial regression with a binary x

Theoretical examples

Example 1

Suppose we want to examine the association between gender (x) and political views (y). Gender has the values 0=Man and 1=Woman, whereas political views has the values 1=Conservative, 2=Centre, and 3=Liberal. We choose Centre as the base outcome. Now, we get an RRR of 0.82 for Conservative, which means that women are less likely to be conservative than centre compared to men. The RRR for Liberal is 1.39, suggesting that women are more likely to be liberal than centre compared to men.

Example 2

Here we want to examine the association between having young children (x) and the type of pet owned (y). Having young children is measured as either 0=No young children and 1=Young children. Type of pet owned has the values 1=No pet, 2=Cat, 3=Dog, and 4=Other type of pet. The category No pet is chosen as the base outcome. Let us say that we get an RRR for Cat that is 1.50. This means that those who have young children are more likely to own a cat than no pet at all, compared to those who do not have young children. The RRR for Dog is 1.75, suggesting that those who have young children are more likely to own a dog than no pet at all, compared to those who do not have young children. Moreover, the RRR for Other type of pet is 1.96, which tells us that those who have young children are more likely to own another type of pet than no pet at all, compared to those who do not have young children.

Practical example

Dataset: StataData1.dta

Name	**Label**
marstat40	Marital status (Age 40, Year 2010)
sex	Sex

```
sum marstat40 sex if pop_multinom==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   marstat40 |      8,409     1.69378    .8148308          1          4
         sex |      8,409    .4960162    .5000139          0          1
```

```
mlogit marstat40 sex if pop_multinom==1, rrr b(1)
```

```
Multinomial logistic regression                 Number of obs     =      8,409
                                                LR chi2(3)        =      86.17
                                                Prob > chi2       =     0.0000
Log likelihood = -8851.8486                     Pseudo R2         =     0.0048

------------------------------------------------------------------------------
   marstat40 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Married      |  (base outcome)
-------------+----------------------------------------------------------------
Unmarried    |
         sex |    .762002   .0395645    -5.23   0.000     .6882722    .8436299
       _cons |   .5896851   .0206845   -15.06   0.000     .5505065     .631652
-------------+----------------------------------------------------------------
Divorced     |
         sex |   1.239931   .0718404     3.71   0.000     1.106827    1.389042
       _cons |   .3363761   .0143237   -25.59   0.000     .3094417    .3656549
-------------+----------------------------------------------------------------
Widowed      |
         sex |   3.484098    .937082     4.64   0.000     2.056608    5.902409
       _cons |   .0082154   .0019443   -20.29   0.000     .0051663    .0130642
------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

When we look at the results for sex, we see that the relative risk ratio (RRR) is 0.76 for Unmarried, 1.24 for Divorced, and 3.48 for Widowed. This means that women (who are coded as 1 and thus compared to men who are coded as 0), have a lower risk of being unmarried, but a higher risk of being a divorced or widowed, as compared to being married/having a registered partner.

There are statistically significant differences between Married and Unmarried, Married and Divorced, as well as Married and Widowed, according to sex – as reflected in the p-values (0.000) and the 95% confidence intervals (0.69-0.84, 1.11-1.39, and 2.06-5.90, respectively).

Summary
At age 40, women are less likely than men to be unmarried (RRR=0.76, 95% CI=0.69-0.84) but more likely to be divorced (RRR=1.24, 95% CI=1.11-1.39) or widowed (RRR=3.48, 95% CI=2.06-5.90), in comparison to being married.

15.3.3 Simple multinomial regression with a categorical (non-binary) x

Theoretical examples

Example 1

We want to investigate the association between educational attainment (x) and building type (y). Educational attainment has the values: 1=Compulsory, 2=Upper secondary, and 3=University. Building type has the values 1=Apartment, 2=Town house, and 3=Villa. We choose Compulsory as our reference category and Apartment as our base outcome. The RRR for Upper secondary in combination with Town house is 2.01, meaning that those with upper secondary education are more likely to live in a town house than an apartment, compared to those with compulsory education. The RRR for Upper secondary in combination with Villa is 1.32, meaning that those with upper secondary education are more likely to live in a villa than an apartment, compared to those with compulsory education. For University in combination with Town house, the RRR is 0.95, suggesting that those who have university education are less likely to live in a town house than an apartment compared to those with compulsory education. Finally, the RRR for University in combination with Villa is 3.44, meaning that those with university education are more likely to live in a villa than an apartment, compared to those with compulsory education.

Example 2

Suppose we are interested in the association between family type (x) and adolescent health behaviour (y). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. Adolescent health behaviour has the values 1=No smoking or alcohol consumption, 2=Smoking, 3=Alcohol consumption, and 4=Both smoking and alcohol consumption. We choose Two-parent household as our reference category, and No smoking or alcohol consumption as our base outcome. The RRR for Joint custody and Smoking is 1.20, meaning that adolescents living in joint custody are more likely to smoke than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the Single-parent household and Smoking is 1.49, meaning that adolescents living in single-parent household are more likely to smoke than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the Joint custody and Alcohol consumption is 1.00, meaning that adolescents living in joint custody are as likely to drink alcohol as not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the Single-parent household and Alcohol consumption is 2.02, meaning that adolescents living in single-parent household are more likely to drink alcohol than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for Joint custody and Both smoking and alcohol consumption is 1.55, meaning that adolescents living in joint custody are more likely to both smoke and drink alcohol than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the Single-parent household and Both smoking and alcohol consumption is 4.45, meaning that adolescents living in single-parent household are more likely to both smoke and drink alcohol than not to smoke or drink alcohol compared to those living in a two-parent household.

Practical example

Dataset: StataData1.dta

Name	**Label**
marstat40	Marital status (Age 40, Year 2010)
educ	Educational level (Age 40, Year 2010)

```
sum marstat40 educ if pop_multinom==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   marstat40 |      8,409     1.69378    .8148308          1          4
        educ |      8,409    2.181234    .7204204          1          3
```

```
mlogit marstat40 ib1.educ if pop_multinom==1, rrr b(1)
```

```
Multinomial logistic regression                 Number of obs     =      8,409
                                                LR chi2(6)        =     196.39
                                                Prob > chi2       =     0.0000
Log likelihood = -8796.7378                     Pseudo R2         =     0.0110

------------------------------------------------------------------------------------
       marstat40 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------------
Married          |  (base outcome)
-----------------+------------------------------------------------------------------
Unmarried        |
            educ |
Upper secondary  |   .7047889   .0495791    -4.97   0.000      .614017    .8089798
      University |   .4596606   .0338321   -10.56   0.000     .3979119    .5309917
                 |
           _cons |   .8080495   .0475563    -3.62   0.000     .7200159    .9068467
-----------------+------------------------------------------------------------------
Divorced         |
            educ |
Upper secondary  |   .7640576    .059078    -3.48   0.001     .6566138    .8890828
      University |   .4203944   .0349833   -10.41   0.000      .357128    .4948687
                 |
           _cons |   .5789474   .0376173    -8.41   0.000     .5097204    .6575763
-----------------+------------------------------------------------------------------
Widowed          |
            educ |
Upper secondary  |    .761382   .2323877    -0.89   0.372      .418601    1.384857
      University |    .640873   .2005333    -1.42   0.155     .3470768    1.183364
                 |
           _cons |   .0247678   .0062682   -14.61   0.000     .0150823    .0406731
------------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

When we look at the results for the dummies of educ, we see that the relative risk ratios (RRR) are 0.70 (Upper secondary) and 0.46 (University) for Unmarried, 0.76 (Upper secondary) and 0.42 (University) for Divorced, and 0.76 (Upper secondary) and 0.64 (University) for Widowed. In other words, individuals with higher levels of educational attainment are less likely to be unmarried, divorced, and widowed, as compared to being married.

All estimates are statistically significant except the ones for Widowed.

Test the overall effect

The output presented and interpreted above, is based on the relative risk ratios for the dummy variables of educ. But what about the overall statistical effect of educ on marstat40? We could assess it through contrast, which is a postestimation command. However, since the outcome has four categories, it quickly gets quite messy. We will therefore skip this here.

More information	help contrast

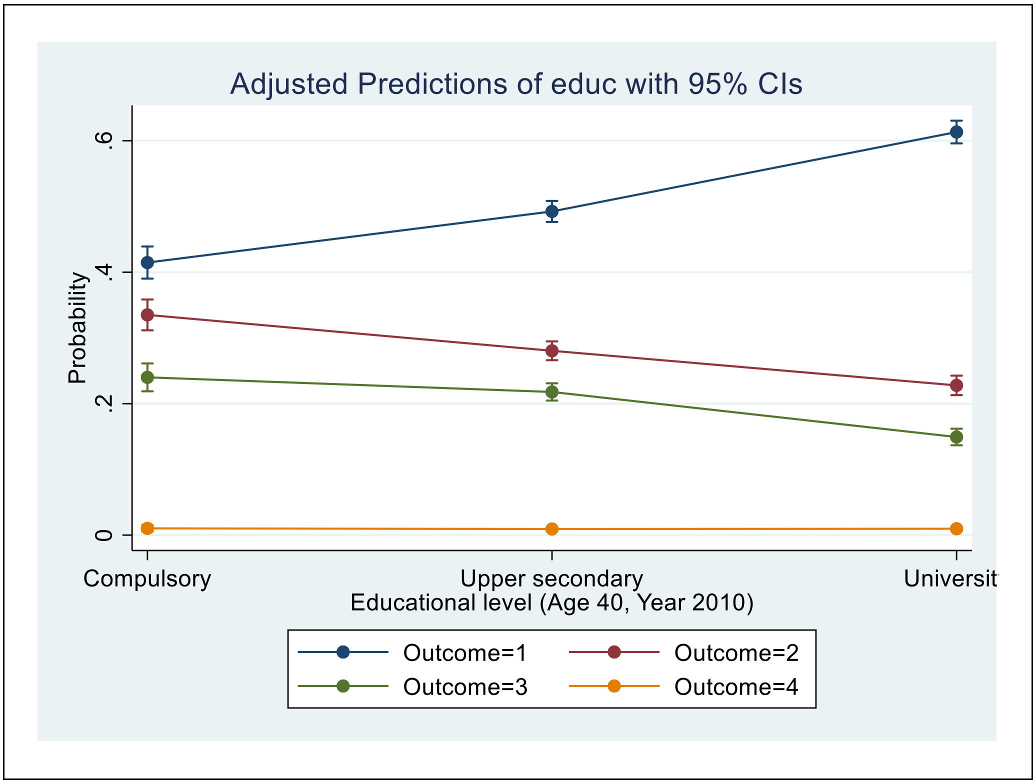
We will nonetheless produce a graph of the trend. First, however, we need to apply the post-estimation command margins.

Note This command can also be used for variables that are continuous or binary, but is particularly useful for categorical, non-binary (i.e. ordinal) variables.

```
margins educ
```

```
Adjusted predictions                              Number of obs     =      8,409
Model VCE    : OIM

1._predict   : Pr(marstat40==Married), predict(pr outcome(1))
2._predict   : Pr(marstat40==Unmarried), predict(pr outcome(2))
3._predict   : Pr(marstat40==Divorced), predict(pr outcome(3))
4._predict   : Pr(marstat40==Widowed), predict(pr outcome(4))

------------------------------------------------------------------------------------
                   |            Delta-method
                   |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
     _predict#educ |
      1#Compulsory |    .4146341   .0124814    33.22   0.000     .3901711    .4390972
1#Upper secondary  |    .4924383   .0081434    60.47   0.000     .4764775    .5083991
      1#University |    .6132382   .0087724    69.91   0.000     .5960445    .6304318
      2#Compulsory |    .3350449   .0119581    28.02   0.000     .3116074    .3584825
2#Upper secondary  |    .2804457   .0073172    38.33   0.000     .2661044    .2947871
      2#University |    .2277742   .0075545    30.15   0.000     .2129675    .2425808
      3#Compulsory |    .2400513   .0108208    22.18   0.000     .2188429    .2612598
3#Upper secondary  |    .2178297   .0067235    32.40   0.000     .2046518    .2310075
      3#University |    .1492537   .0064187    23.25   0.000     .1366733    .1618341
      4#Compulsory |    .0102696   .0025542     4.02   0.000     .0052635    .0152757
4#Upper secondary  |    .0092863   .0015624     5.94   0.000     .0062241    .0123485
      4#University |    .0097339   .0017685     5.50   0.000     .0062678    .0132001
------------------------------------------------------------------------------------
```

```
marginsplot
```



Note The y-axis shows predicted probabilities (i.e. not relative log odds or relative risk ratios). The different colours reflect the different categories of the y-variable.

More information	help marginsplot

Summary
Individuals with higher levels of educational attainment are less likely to be unmarried, divorced, and widowed, as compared to being married. Except for the educational differences in the risk of being widowed, the associations are statistically significant.

15.4 Multiple multinomial regression

Quick facts	
Number of variables	One dependent (y) At least two independent (x)
Scale of variable(s)	Dependent: nominal (with more than two categories) Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

Theoretical example

Example

Suppose we are interested to see if having young children (x), residential area (x), and income (x) are related to smoking (y). Having young children is measured as either 0=No young children and 1=Young children. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. We choose Metropolitan as our reference category. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK). Smoking has the values 1=Non-smoker, 2=Former smoker, and 3=Current smoker. We choose Non-smoker as our base outcome.

In the regression analysis, we get an RRR of 1.19 for Young children and Former smoker, suggesting that those who have young children are more likely to be former smokers than non-smokers compared to those who do not have young children. Then we get an RRR of 0.77 for Young children and Current smoker, which means that those who have young children are less likely to be current smokers than non-smokers compared to those who do not have young children. These results are adjusted for residential area and income.

The RRR for Smaller city and Former smoker is 2.09, which suggests that those who live in a smaller city are more likely to be former smokers than non-smokers compared to those who live in a metropolitan area. The RRR for Smaller city and Current smoker is 3.71, which suggests that those who live in a smaller city are more likely to be current smokers than non-smokers compared to those who live in a metropolitan area. The RRR for Rural and Former smoker is 3.59, which suggests that those who live in a rural area are more likely to be former smokers than non-smokers compared to those who live in a metropolitan area. The RRR for Rural and Current smoker is 5.01, which suggests that those who live in a rural area are more likely to be current smokers than non-smokers compared to those who live in a metropolitan area. These results are adjusted for having young children and income.

With regard to income, the RRR for Former smoker is 0.93, suggesting that for every unit increase in income, the risk of being a former smoker decreases. The RRR for Current smoker is 0.78, which means that for every unit increase in income, the risk of being a current smoker also decreases. These results are adjusted for having young children and residential area.

Practical example

Dataset: StataData1.dta

Name	**Label**
marstat40	Marital status (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)
sex	Sex
educ	Educational level (Age 40, Year 2010)

```
sum marstat40 gpa sex educ if pop_multinom==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   marstat40 |      8,409     1.69378    .8148308          1          4
         gpa |      8,409    3.184861    .6931467          1          5
         sex |      8,409    .4960162    .5000139          0          1
        educ |      8,409    2.181234    .7204204          1          3
```

```
mlogit marstat40 gpa sex ib1.educ if pop_multinom==1, rrr b(1)
```

```
Multinomial logistic regression                 Number of obs     =      8,409
                                                LR chi2(12)       =     300.63
                                                Prob > chi2       =     0.0000
Log likelihood = -8744.6189                     Pseudo R2         =     0.0169

------------------------------------------------------------------------------
       marstat40 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
Married          |  (base outcome)
-----------------+------------------------------------------------------------
Unmarried        |
             gpa |   .8715537   .0379514    -3.16   0.002      .800256    .9492036
             sex |   .8087497   .0428008    -4.01   0.000     .7290658    .8971428
                 |
            educ |
 Upper secondary |   .7527687   .0541059    -3.95   0.000      .653854    .8666472
      University |   .5317268   .0440554    -7.62   0.000     .4520262    .6254801
                 |
           _cons |   1.271671   .1666211     1.83   0.067     .9836622    1.644008
-----------------+------------------------------------------------------------
Divorced         |
             gpa |   .8495054   .0414234    -3.34   0.001      .772076    .9346999
             sex |   1.327299   .0786335     4.78   0.000     1.181791    1.490722
                 |
            educ |
 Upper secondary |   .7818215   .0618434    -3.11   0.002     .6695393    .9129334
      University |   .4696683   .0438977    -8.09   0.000      .391051     .564091
                 |
           _cons |   .7964132    .116721    -1.55   0.120     .5975686    1.061425
-----------------+------------------------------------------------------------
Widowed          |
             gpa |   1.281508   .2422937     1.31   0.190     .8846783    1.856339
             sex |   3.463111   .9400099     4.58   0.000     2.034325    5.895388
                 |
            educ |
 Upper secondary |   .6237379   .1947216    -1.51   0.131     .3382749    1.150097
      University |   .4522459   .1601109    -2.24   0.025     .2259537    .9051693
                 |
           _cons |   .0061186   .0037411    -8.34   0.000     .0018459    .0202817
------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

In this model, we have three x-variables: gpa, sex, and educ. When we put them together, their statistical effect on marstat40 is mutually adjusted.

When it comes to the relative risk ratios, they have changed in comparison to the simple regression models. For example, the relative risk ratios for gpa have decreased (become closer to 1). The relative risk ratios for sex have also decreased slightly – apart from the one for Divorced (which is a bit higher now). Concerning the dummies of educ, they are also closer to 1 now, except for the ones for Widowed.

Regarding statistical significance, the same results are in the single regression models are found here.

Note A specific relative risk ratio from a simple multinomial regression model can increase when other x-variables are included. Usually, it is just “noise”, i.e. not any large increases, and therefore not much to be concerned about. But it can also reflect that there is something going on that we need to explore further. There are many possible explanations for increases in multiple regression models: a) We actually adjust for a confounder and then “reveal” the “true” statistical effect. b) There are interactions among the x-variables in their effect on the y-variable. c) There is something called collider bias (which we will not address in this guide) which basically mean that both the x-variable and the y-variable causes another x-variable in the model. d) The simple regression models and the multiple regression model are based on different samples. e) It can be due to rescaling bias (see Chapter 18).

Summary
In the fully adjusted model, most differences are attenuated but the overall patterns remain the same.

Estimates table and coefficients plot

If we have multiple models, we can facilitate comparisons between the regression models by asking Stata to construct estimates tables and coefficients plots. What we do is to run the regression models one-by-one, save the estimates after each, and than use the commands estimates table and coefplot.

The coefplot option is not part of the standard Stata program, so unless you already have added this package, you need to install it:

```
ssc install coefplot
```

As an example, we can include the three simple regression models as well as the multiple regression model. The quietly option is included in the beginning of the regression commands to suppress the output.

Run and save the first simple regression model:

```
quietly mlogit marstat40 gpa if pop_multinom==1, rrr b(1)
```

```
estimates store model1
```

Run and save the second simple regression model:

```
quietly mlogit marstat40 sex if pop_multinom==1, rrr b(1)
```

```
estimates store model2
```

Run and save the third simple regression model:

```
quietly mlogit marstat40 ib1.educ if pop_multinom==1, rrr b(1)
```

```
estimates store model3
```

Run and save the multiple regression model:

```
quietly mlogit marstat40 gpa sex ib1.educ if pop_multinom==1, rrr b(1)
```

```
estimates store model4
```

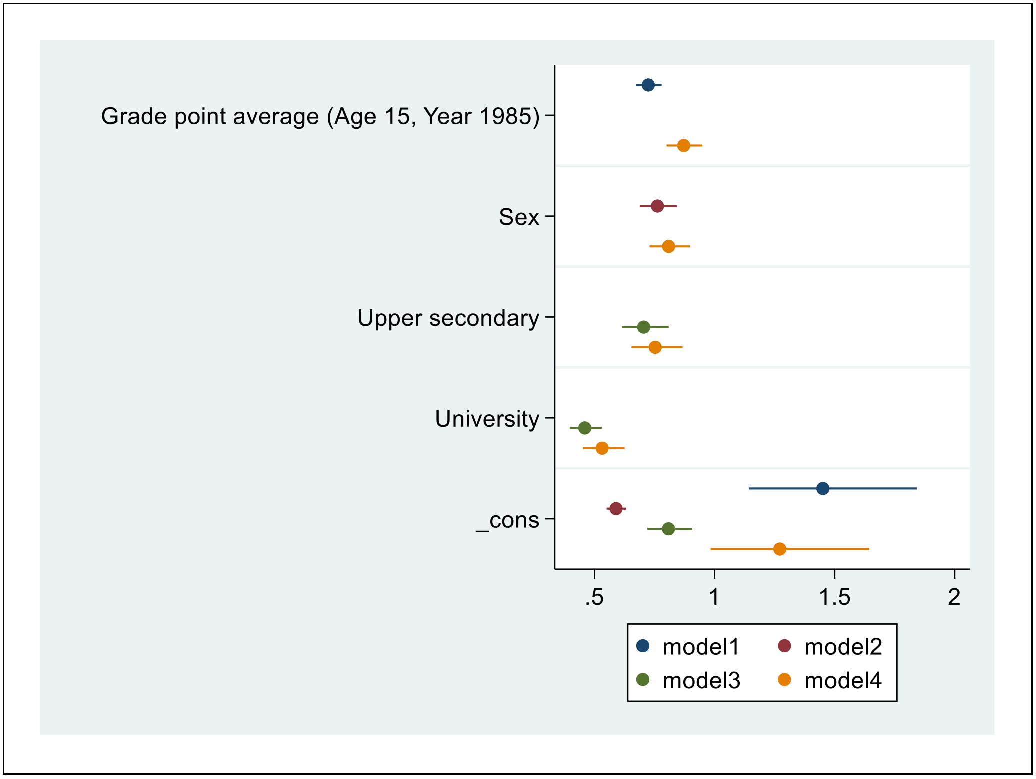
Produce the estimates table (include the option eform to show relative risk ratios):

```
estimates table model1 model2 model3 model4, eform
```

```
-------------------------------------------------------------------
    Variable |   model1       model2       model3       model4
-------------+-----------------------------------------------------
Married      |
         gpa |   (omitted)                              (omitted)
         sex |                (omitted)                 (omitted)
             |
        educ |
Upper sec..  |                             (omitted)    (omitted)
 University  |                             (omitted)    (omitted)
             |
       _cons |   (omitted)    (omitted)    (omitted)    (omitted)
-------------+-----------------------------------------------------
Unmarried    |
         gpa |   .72391212                              .87155368
         sex |                  .762002                 .80874972
             |
        educ |
Upper sec..  |                             .70478886     .7527687
 University  |                             .45966065    .53172679
             |
       _cons |   1.4510425    .58968508    .80804954    1.2716714
-------------+-----------------------------------------------------
Divorced     |
         gpa |   .71561519                              .84950536
         sex |                 1.239931                 1.3272988
             |
        educ |
Upper sec..  |                              .7640576    .78182146
 University  |                             .42039442    .46966833
             |
       _cons |   1.0911296    .33637608    .57894737    .79641316
-------------+-----------------------------------------------------
Widowed      |
         gpa |   1.1896736                              1.2815081
         sex |                3.4840981                  3.463111
             |
        educ |
Upper sec..  |                             .76138199    .62373791
 University  |                               .640873    .45224591
             |
       _cons |   .01040019    .00821543     .0247678    .00611861
-------------------------------------------------------------------
```

Produce the coefficients plot (include the option eform to show relative risk ratios):

```
coefplot model1 model2 model3 model4, eform
```



Note You can improve the graph by using the Graph Editor to delete "_cons" as well as to adjust the category and label names.

15.4.1 Alternative base outcomes

Regardless of whether you are performing a simple or multiple multinomial regression analysis, you always need to choose a base outcome for the y-variable. Our results will only show us to differences between the base outcome and the other categories of the outcome, and not contrast the remaining combinations. Of course, you can repeat your analysis and alternate between the base outcomes – or you can use the listcoef command. This command requires that you install a user-written package first. So, if you have not installed it already, type:

```
search spost13_ado
```

Click on the first link in the list, and then choose Click here to install.

Let us go back to the multiple regression analysis that we conducted earlier. There, we chose Married as the base outcome. The quietly option is included in the beginning of the command to suppress the output.

```
quietly mlogit marstat40 gpa sex ib1.educ if pop_multinom==1, rrr b(1)
```

We move on to the listcoef command (which is a postestimation command). Beware that this often produces comprehensive output.

listcoef

```
mlogit (N=8409): Factor change in the odds of marstat40

Variable: gpa (sd=0.693)
-------------------------------------------------------------------------------
                             |          b        z    P>|z|       e^b   e^bStdX
-----------------------------+-------------------------------------------------
Married      vs Unmarried    |     0.1375    3.157    0.002     1.147     1.100
Married      vs Divorced     |     0.1631    3.345    0.001     1.177     1.120
Married      vs Widowed      |    -0.2480   -1.312    0.190     0.780     0.842
Unmarried    vs Married      |    -0.1375   -3.157    0.002     0.872     0.909
Unmarried    vs Divorced     |     0.0256    0.472    0.637     1.026     1.018
Unmarried    vs Widowed      |    -0.3855   -2.021    0.043     0.680     0.766
Divorced     vs Married      |    -0.1631   -3.345    0.001     0.850     0.893
Divorced     vs Unmarried    |    -0.0256   -0.472    0.637     0.975     0.982
Divorced     vs Widowed      |    -0.4111   -2.144    0.032     0.663     0.752
Widowed      vs Married      |     0.2480    1.312    0.190     1.282     1.188
Widowed      vs Unmarried    |     0.3855    2.021    0.043     1.470     1.306
Widowed      vs Divorced     |     0.4111    2.144    0.032     1.509     1.330
-------------------------------------------------------------------------------

Variable: sex (sd=0.500)
-------------------------------------------------------------------------------
                             |          b        z    P>|z|       e^b   e^bStdX
-----------------------------+-------------------------------------------------
Married      vs Unmarried    |     0.2123    4.011    0.000     1.236     1.112
Married      vs Divorced     |    -0.2831   -4.779    0.000     0.753     0.868
Married      vs Widowed      |    -1.2422   -4.576    0.000     0.289     0.537
Unmarried    vs Married      |    -0.2123   -4.011    0.000     0.809     0.899
Unmarried    vs Divorced     |    -0.4954   -7.514    0.000     0.609     0.781
Unmarried    vs Widowed      |    -1.4544   -5.327    0.000     0.234     0.483
Divorced     vs Married      |     0.2831    4.779    0.000     1.327     1.152
Divorced     vs Unmarried    |     0.4954    7.514    0.000     1.641     1.281
Divorced     vs Widowed      |    -0.9590   -3.496    0.000     0.383     0.619
Widowed      vs Married      |     1.2422    4.576    0.000     3.463     1.861
Widowed      vs Unmarried    |     1.4544    5.327    0.000     4.282     2.069
Widowed      vs Divorced     |     0.9590    3.496    0.000     2.609     1.615
-------------------------------------------------------------------------------

Variable: 2.educ (sd=0.497)
-------------------------------------------------------------------------------
                             |          b        z    P>|z|       e^b   e^bStdX
-----------------------------+-------------------------------------------------
Married      vs Unmarried    |     0.2840    3.951    0.000     1.328     1.152
Married      vs Divorced     |     0.2461    3.112    0.002     1.279     1.130
Married      vs Widowed      |     0.4720    1.512    0.131     1.603     1.265
Unmarried    vs Married      |    -0.2840   -3.951    0.000     0.753     0.868
Unmarried    vs Divorced     |    -0.0379   -0.449    0.654     0.963     0.981
Unmarried    vs Widowed      |     0.1880    0.599    0.549     1.207     1.098
Divorced     vs Married      |    -0.2461   -3.112    0.002     0.782     0.885
Divorced     vs Unmarried    |     0.0379    0.449    0.654     1.039     1.019
Divorced     vs Widowed      |     0.2259    0.717    0.473     1.253     1.119
Widowed      vs Married      |    -0.4720   -1.512    0.131     0.624     0.791
Widowed      vs Unmarried    |    -0.1880   -0.599    0.549     0.829     0.911
Widowed      vs Divorced     |    -0.2259   -0.717    0.473     0.798     0.894
-------------------------------------------------------------------------------

Variable: 3.educ (sd=0.482)
-------------------------------------------------------------------------------
                             |          b        z    P>|z|       e^b   e^bStdX
-----------------------------+-------------------------------------------------
Married      vs Unmarried    |     0.6316    7.623    0.000     1.881     1.356
Married      vs Divorced     |     0.7557    8.086    0.000     2.129     1.439
Married      vs Widowed      |     0.7935    2.241    0.025     2.211     1.466
Unmarried    vs Married      |    -0.6316   -7.623    0.000     0.532     0.738
Unmarried    vs Divorced     |     0.1241    1.212    0.226     1.132     1.062
Unmarried    vs Widowed      |     0.1619    0.454    0.650     1.176     1.081
Divorced     vs Married      |    -0.7557   -8.086    0.000     0.470     0.695
Divorced     vs Unmarried    |    -0.1241   -1.212    0.226     0.883     0.942
Divorced     vs Widowed      |     0.0378    0.105    0.916     1.039     1.018
Widowed      vs Married      |    -0.7935   -2.241    0.025     0.452     0.682
Widowed      vs Unmarried    |    -0.1619   -0.454    0.650     0.851     0.925
Widowed      vs Divorced     |    -0.0378   -0.105    0.916     0.963     0.982
-------------------------------------------------------------------------------
```

In the table above, you can see the differences between all base outcomes, stratified by the x-variables included in the model.

Note The coefficients (in the column called b) are relative log odds, not relative risk ratios.

It is possible to make the output a bit more compact, for example by specifying the pvalue option:

```
listcoef, pvalue(05)
```

With the above-specified option, only variables/categories with a p-value below 0.05 will be displayed in the table.

More information	help listcoef

You can get some further assistance in interpreting the coefficients in the table by the mlogitplot command. This command requires that you install a user-written package first. So, if you have not installed it already, type:

```
search spost13_ado
```

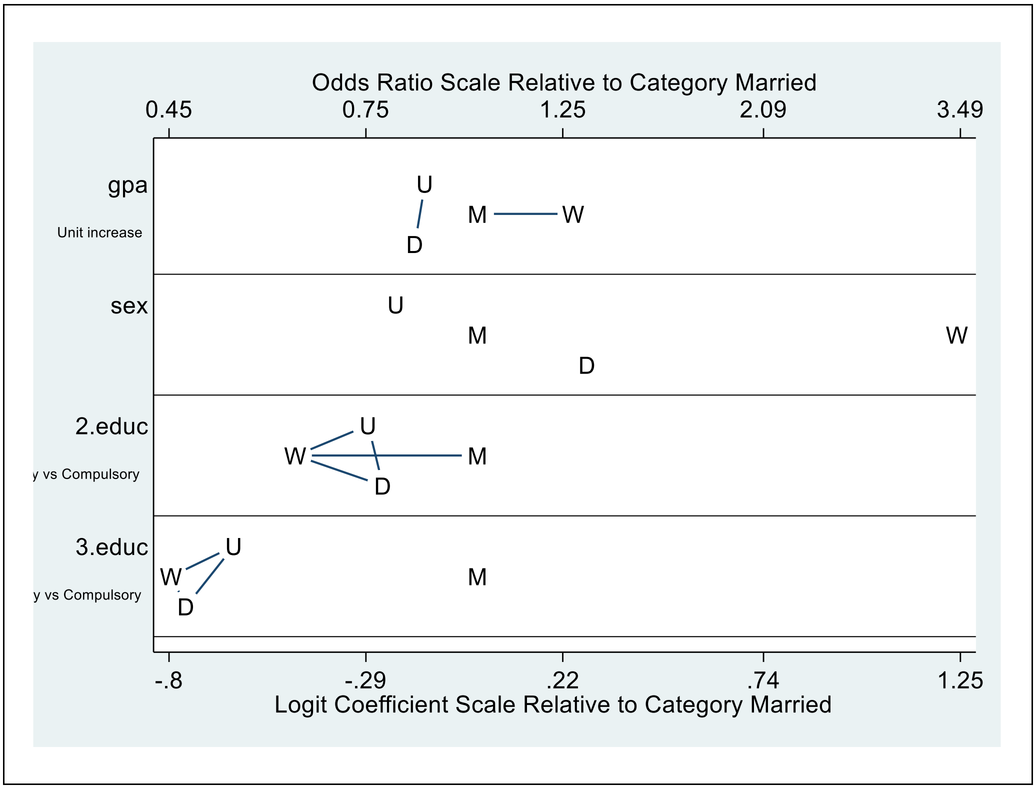
Click on the first link in the list, and then choose Click here to install.

Let us go back to the multiple regression analysis that we conducted earlier. The quietly option is included in the beginning of the command to suppress the output.

```
quietly mlogit marstat40 gpa sex ib1.educ if pop_multinom==1, rrr b(1)
```

The mlogitplot command creates an odds ratios plot for mlogit. We should specify the amount of change when we order the mlogitplot. In our multiple regression model, we have three variables. The first is gpa: this is a continuous variable which can be described using unit change (amount option: one). The second is sex, which is a binary variable that can be described using binary change (amount option: bin). The third is educ, which is included as a factor variable – this automatically plots changes from 0 to 1.

```
mlogitplot gpa sex ib1.educ, amount(one bin)
```



The letters denote the different categories of marstat40. On the y-axis, we have the variables/categories included in the model. The x-axis shows the odds ratios (as well as the relative log odds) in relation to the base outcome (which is Married, or M, in this case).

Note Between some of the letters, there are lines. This means that the difference is not statistically significant.

More information	help mlogitplot

15.5 Model diagnostics

The assumptions behind multinomial regression are similar to the ones for logistic regression. Since the y-variable has multiple categories, model diagnostics are nonetheless slightly more complicated.

More information	help mlogit postestimation

Checklist	
Categorical outcome	The y-variable should be categorical (and non-binary). Check whether it is possible to group similar categories (cf. the Blue bus/Red bus problem). Although it makes most sense to use multinomial regression analysis if the y-variable is nominal with more than two categories, it is possible to use a binary outcome – however, then you could just as well go with a plain logistic regression (unless you want to obtain some of the test available for multinomial regression analysis). It is also possible to have an ordinal y-variable (e.g., if the assumptions for ordinal regression were violated, you can try a multinomial regression instead; the latter does not assume parallel lines).
Independence of errors	Data should be independent, i.e. not derived from any dependent samples design, e.g. before-after measurements/paired samples.
No multicollinearity	Multicollinearity may occur when two or more x-variables that are included simultaneously in the model are strongly correlated with each another. Actually, this does not violate the assumptions, but is does create greater standard errors which makes it harder to reject the null hypothesis.

Types of model diagnostics	
Fit statistics	Assess model fit
Correlation matrix	Check for multicollinearity

15.5.1 Assess model fit

With the command fitstat, we can produce various types of model fit statistics. This command requires that you install a user-written package first. So, if you have not installed it already, type:

```
ssc install spost13_ado
```

Click on the first link in the list, and then choose Click here to install.

More information	`help fitstat`

Practical example

We perform this test for the full model, so let us go back to the example from the multiple regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly mlogit marstat40 gpa sex ib1.educ if pop_multinom==1, rrr b(1)
```

And then we produce the statistics:

```
fitstat
```

```
                         |      mlogit
-------------------------+-------------
Log-likelihood           |
                   Model |   -8744.619
          Intercept-only |   -8894.934
-------------------------+-------------
Chi-square               |
       Deviance(df=8394) |   17489.238
               LR(df=12) |     300.629
                 p-value |       0.000
-------------------------+-------------
R2                       |
                McFadden |       0.017
      McFadden(adjusted) |       0.015
            Cox-Snell/ML |       0.035
  Cragg-Uhler/Nagelkerke |       0.040
                   Count |       0.523
         Count(adjusted) |       0.002
-------------------------+-------------
IC                       |
                     AIC |   17519.238
        AIC divided by N |       2.083
              BIC(df=15) |   17624.794
```

This reports the log-likelihoods of the full (Model) and empty (Intercept-only) models, the deviance, the likelihood ratio test, Akaike's Information Criterion (AIC), AIC/N, and the Bayesian Information Criterion (BIC). In addition, we obtain different types of R2 estimates (which are seldom used).

One very practical thing is that we can use these statistics to compare models. For example, we might want to see whether model fit improves if we include or exclude one or more x-variables, or if we make any transformations of the included x-variables.

Let us assume that we want to see here if our multiple regression model has a better fit if we exclude the variable gpa.

First, we run the original model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly mlogit marstat40 gpa sex ib1.educ if pop_multinom==1, rrr b(1)
```

And then save the statistics:

```
fitstat, save
```

Then we run the alternative model (output suppressed here as well):

```
quietly mlogit marstat40 sex ib1.educ if pop_multinom==1, rrr b(1)
```

And then compare the statistics:

```
fitstat, dif
```

```
                         |     Current        Saved   Difference
-------------------------+---------------------------------------
Log-likelihood           |
                   Model |   -8754.137    -8744.619       -9.518
          Intercept-only |   -8894.934    -8894.934        0.000
-------------------------+---------------------------------------
Chi-square               |
       D(df=8397/8394/3) |   17508.274    17489.238       19.036
          LR(df=9/12/-3) |     281.593      300.629      -19.036
                 p-value |       0.000        0.000        0.000
-------------------------+---------------------------------------
R2                       |
                McFadden |       0.016        0.017       -0.001
      McFadden(adjusted) |       0.014        0.015       -0.001
            Cox-Snell/ML |       0.033        0.035       -0.002
  Cragg-Uhler/Nagelkerke |       0.037        0.040       -0.002
                   Count |       0.522        0.523       -0.001
         Count(adjusted) |       0.000        0.002       -0.002
-------------------------+---------------------------------------
IC                       |
                     AIC |   17532.274    17519.238       13.036
        AIC divided by N |       2.085        2.083        0.002
        BIC(df=12/15/-3) |   17616.718    17624.794       -8.075

Note: Likelihood-ratio test assumes current model nested in saved model.

Difference of    8.075 in BIC provides strong support for current model.
```

It seems as both the chi-square (as indicated by the significant LR test), and the AIC favours the original (saved) model with gpa, whereas the BIC favours the model without gpa. This is not surprising given that BIC tends to be lower for more parsimonious (simpler) models. Should we change our model based on these statistics by, in this case, excluding gpa? That is a difficult question that needs to be considered carefully (by reflecting upon theory, previous research, and other alternatives for estimation).

Note When we compare BIC and/or AIC values, we prefer the model with the lowest values.

15.5.2 Correlation matrix

As the x-variables become more strongly correlated, it becomes more difficult to determine which of the variables are actually producing the statistical effect on the y-variable. This is the problem with multicollinearity.

One way of assessing multicollinearity is using the estat vce command, with the corr (short for correlation) option.

More information	help estat vce

Practical example

The first step is re-run the multiple multinomial regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly mlogit marstat40 gpa sex ib1.educ if pop_multinom==1, rrr b(1)
```

Next, we try the estat vce command. By adding the corr (=correlation) option, we will get a correlation matrix instead of a covariance matrix.

```
estat vce, corr
```

The table is too extensive to be pasted here. But per usual, we go through the coefficients and see if there are any strong correlations between the variables/categories (see Chapter 7.2).

16. POISSON REGRESSION

Outline

Content

This chapter starts with an introduction to Poisson regression and then presents the function in Stata. After this, we offer some practical examples of how to perform simple and multiple Poisson regression, as well as how to generate and interpret model diagnostics.

16.1 Introduction

Poisson regression is used when y is based on count data.

Some examples of count data

- Number of visits to the hospital in a year
- Number of days in unemployment in a month
- Number of text messages sent per day
- Number of goals scored in a game

With a linear regression, one assumes that the value of y can be predicted based on the values of one or more x-variables – as well as their residuals. A requirement is that the residuals are normally distributed. A Poisson regression is a type of generalised linear model which instead assumes a Poisson distribution. A Poisson distribution describes the probability that a number of events will take place within a given interval of time (or space), conditioned upon the fact that the events occur with a constant speed and that every event is independent from any preceding events. When performing a Poisson regression, one uses a link function that allows for a linear combination of the x-variables to predict the logarithm of y.

Poisson regression is used to predict the rate that y changes given the values of the independent variables. As for any regression analysis, we get a coefficient – here, it is called log incidence rate – that shows the effect of x on y. Usually, we focus on something called the incidence rate ratio (IRR). We can calculate the incidence-rate ratio by taking the exponent of the coefficient.

Overdispersion or zero-inflation?

Poisson regression assumes that the mean is equivalent to the variance. If the variance is greater than the mean, we get something called overdispersion. In this case we can apply a so-called negative binomial regression. Another common problem is that we have a lot of observations with the value 0. In this case, it might be necessary to apply a zero-inflated Poisson regression model. These specifications will be briefly explored in this guide as well.

Other names for Poisson regression

Poisson regression is sometimes called, e.g., log-linear regression.

16.1.1 Poisson regression in short

If you have only one x, it is called simple regression, and if you have more than one x, it is called multiple regression.

Regardless of whether you are doing a simple or a multiple regression, x-variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from Poisson regression		
Effect		
Incidence rate ratio (IRR)	The exponent of log incidence rate	
	Log incidence rate	The logarithm of incidence rate
	Incidence rate	The rate at which events occur
Direction		
Negative	IRR below 1	
Positive	IRR above 1	
Statistical significance		
P-value	p<0.05 Statistically significant at the 5% level p<0.01 Statistically significant at the 1 % level p<0.001 Statistically significant at the 0.1% level	
95% Confidence intervals	Interval does not include 1: Statistically significant at the 5% level Interval includes 1: Statistically non-significant at the 5% level	

Incidence-rate ratio (IRR)

In Poisson regression analysis, the effect that x has on y is reflected by an incidence rate ratio (IRR):

IRR below 1	For every unit increase in x, the incidence rate of y decreases.
IRR above 1	For every unit increase in x, the incidence rate of y increases.

Exactly how one interprets the IRR in plain writing depends on the measurement scale of the x-variable. That is why we will present examples later for continuous, binary, and categorical (non-binary) x-variables.

Note Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for Poisson regression is 1.

Note An IRR can never be negative – it can range between 0 and infinity.

How to *not* interpret incidence rate ratios

The incidence rate ratios produced with Poisson regression analysis are not the same as risk ratios (see Section 4.7.6). IRRs tend to be inflated when they are above 1 and understated when they are below 1. This becomes more problematic the more common the outcome is (i.e. the more "non-zeros" we have). However, the rarer the outcome is (<10% is usually considered a reasonable cut-off here), the closer incidence rate ratios and risks ratios become.

Many would find it compelling to interpret IRRs in terms of percentages. For example, an IRR of 1.20 might lead to the interpretation that the incidence rate of the outcome increases by 20%. If the IRR is 0.80, some would then suggest that the incidence rate decreases by 20%. We would to urge you to carefully reflect upon the latter kind of interpretation since incidence rate ratios are not symmetrical: it can take any value above 1 but cannot be below 0. Thus, the choice of reference category might lead to quite misleading conclusions about effect size. The former kind of interpretation is usually considered reasonable when IRRs are below 2. If they are above 2, it is better to refer to "times", i.e. an IRR of 4.07 could be interpreted as "more than four times the odds of…".

Take home message
It is completely fine to discuss the results more generally in terms of higher or lower incidence rates/risks. However, if you want to give exact numbers to exemplify, you need to consider the asymmetry of incidence rate ratios as well as the size of the IRR.

P-values and confidence intervals

In Poisson regression analysis you can get information about statistical significance, in terms of both p-values and confidence intervals (also see Section 5.2).

Note The p-values and the confidence intervals will give you partly different information, but they are not contradictory. If the p-value is below 0.05, the 95% confidence interval will not include 1 and, if the p-value is above 0.05, the 95% confidence interval will include 1.

When you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5% level, the 1% level, or the 0.1% level).

When it comes to confidence intervals, Stata will by default choose 95% level confidence intervals. It is however possible to change the confidence level for the intervals. For example, you may instruct Stata to show 99% confidence intervals instead.

R-Squared

R-Squared (or R2) does not work very well due to the assumptions behind Poisson regression. Stata produces a pseudo R2, but due to inherent bias this is seldom used.

Simple versus multiple regression models

The difference between simple and multiple regression models, is that in a multiple regression each x-variable's effect on y is estimated while accounting for the other x-variables' effects on y. We then say that these other x-variables are "held constant", or "adjusted for", or "controlled for". Because of this, multiple regression analysis is a way of dealing with the issue of confounding variables, and to some extent also mediating variables (see Section 9.3).

It is highly advisable to run a simple regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted coefficients with (i.e. what happened to the coefficients when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that they become weaker – which would of course be expected if the x-variables overlapped in their effect on y.

A note

Remember that a regression analysis should follow from theory as well as a comprehensive set of descriptive statistics and knowledge about the data. In the following sections, we will – for the sake of simplicity – not form any elaborate analytical strategy where we distinguish between x-variables and z-variables (see Section 9.3). However, we will define an analytical sample and use a so-called pop variable (see Section 11.5).

16.2 Function

Basic command	poisson depvar indepvars	
Useful options	poisson depvar indepvars, irr	
Explanations	depvar indepvars irr	Insert the name of the y-variable. Insert the name of the x-variable(s) that you want to use. Produces incidence rate ratios.
More information	help poisson	

Note The Poisson command produces log incidence rates, unless otherwise specified.

A walk-through of the output

When we perform a Poisson regression in Stata, the table looks like this:

```
Poisson regression                              Number of obs     =      8,874
                                                LR chi2(2)        =     235.80
                                                Prob > chi2       =     0.0000
Log likelihood =  -14056.49                     Pseudo R2         =     0.0083

------------------------------------------------------------------------------
        yvar |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xvar1 |   1.048746   .0168106     2.97   0.003      1.01631    1.082217
       xvar2 |   .9824951   .0011421   -15.19   0.000     .9802593    .9847361
       _cons |   2.674361   .0824411    31.91   0.000     2.517564    2.840924
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

In this example, yvar is a count variable ranging between 0 and 17, whereas xvar1 is a binary (0/1) variable and xvar2 is a continuous variable ranging between 1 and 40.

The upper part of the table shows a model summary. This is what the different rows mean:

Row	**Explanation**
Log likelihood	This value does not mean anything in itself, but can be used if we would like compare nested models.
Number of obs	The number of observations included in the model.
LR chi2(x)	The likelihood ratio (LR) chi-square test. The number within the brackets shows the degrees of freedom (one per variable).
Prob >chi2	Shows the probability of obtaining the chi-square statistic given that there is no statistical effect of the x-variables on y. If the p-value is below 0.05, we can conclude that the overall model is statistically significant.
Pseudo R2	A type of R-squared value. Seldom used.

The lower part of the table presents the parameter estimates from the analysis.

Column	**Explanation**
	The first column lists the y-variable on top, followed by our x-variable(s). The last row represents the constant (intercept).
IRR	These are the incidence rate ratios.
Std. Err.	The standard errors associated with the coefficient.
Z	Z-value (coefficient divided by the standard error of the coefficient).
P>|z|	P-value.
[95% Conf. Interval]	95% confidence intervals (lower limit and upper limit).

The analytical sample used for the examples

In the subsequent sections, we will use the following variables:

Dataset: StataData1.dta

Name	**Label**
children	Number of children (Age 40, Year 2010)
siblings	Number of siblings (Age 15, Year 1985)
sex	Sex
educ	Educational level (Age 40, Year 2010)

```
sum children siblings sex educ
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    children |      9,053    1.737214    1.552791          0         10
    siblings |      9,977    1.780395     1.33614          0         10
         sex |     10,000       .4892    .4999083          0          1
        educ |      9,183    2.173691    .7263263          1          3
```

We define our analytical sample through the following command:

```
gen pop_poisson=1 if children!=. & siblings!=. & sex!=. & educ!=.
```

This means that new the variable pop_poisson gets the value 1 if the four variables do not have missing information. In this case, we have 9,014 individuals that are included in our analytical sample.

```
tab pop_poisson
```

```
pop_poisson |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      9,014      100.00      100.00
------------+-----------------------------------
      Total |      9,014      100.00
```

16.3 Simple Poisson regression

Quick facts	
Number of variables	One dependent (y) One independent (x)
Scale of variable(s)	Dependent: count Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

16.3.1 Simple Poisson regression with a continuous x

Theoretical examples

Example 1

The association between number of children (x) and the number of sick days during a year (y) is examined. The number of children ranges between 0 and 10, whereas the number of sick days in a year ranges between 0 and 365. The IRR we get is 1.07, suggesting that the higher the number of children, the higher the rate of sick days.

Example 2

In this example, we use a sample of soccer player to analyse the association between body mass index (x) and the number of goals scored during a soccer season (y). Body mass index ranges between 15 and 35, whereas the number of goals ranges from 0 to 60. We find that the IRR is 0.90. In other words, the higher the body mass index, the lower the rate of goals scored in a season.

Practical example

Dataset: StataData1.dta

Name	**Label**
children	Number of children (Age 40, Year 2010)
siblings	Number of siblings(Age 15, Year 1995)

```
sum children siblings if pop_poisson==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    children |      9,014    1.740515    1.552944          0         10
    siblings |      9,014    1.784114    1.327277          0         10
```

```
poisson children siblings if pop_poisson==1, irr
```

```
Poisson regression                              Number of obs     =      9,014
                                                LR chi2(1)        =       3.26
                                                Prob > chi2       =     0.0711
Log likelihood = -16126.481                     Pseudo R2         =     0.0001

------------------------------------------------------------------------------
    children |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    siblings |   1.010844   .0060207     1.81   0.070     .9991126    1.022714
       _cons |   1.707165   .0228468    39.96   0.000     1.662969    1.752537
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

When we look at the results for siblings, we see that the incidence rate ratio (IRR) is 1.01. Thus, for each additional sibling, the rate of children is 1.01 times higher. That is not much.

The association between siblings and children is not statistically significant, as reflected in the p-value (0.070) and the 95% confidence intervals (1.00-1.02).

Summary

There is a positive association between number of siblings and number of children at age 40 (IRR=1.01). The association is however not statistically significant (95% CI=1.00-1.02).

16.3.2 Simple Poisson regression with a binary x

Theoretical examples

Example 1

We examine the association between gender (x) and the number of online healthcare visits per year (y) by means of a simple Poisson regression analysis. Gender has the values 0=Man and 1=Woman, whereas the number of online healthcare visits ranges between 0 and 25. The IRR we get is 1.72. This would mean that women have a higher rate of online healthcare visits per year in comparison to men.

Example 2

In this study, the association between employment status (x) and the number of coffee cups consumed per day (y) is examined. Employment status is coded as 0=Unemployed and 1=Employed. The number of coffee cups consumed per day ranges between 0 and 15. We get an IRR of 0.67. In other words, employed individuals have a lower rate of coffee cups consumed per days as compared to unemployed individuals.

Practical example

Dataset: StataData1.dta

Name	Label
children	Number of children (Age 40, Year 2010)
sex	Sex

```
sum children sex if pop_poisson==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    children |      9,014    1.740515    1.552944          0         10
         sex |      9,014    .4905702    .4999388          0          1
```

```
poisson children sex if pop_poisson==1, irr
```

```
Poisson regression                              Number of obs     =      9,014
                                                LR chi2(1)        =     305.39
                                                Prob > chi2       =     0.0000
Log likelihood = -15975.413                     Pseudo R2         =     0.0095

------------------------------------------------------------------------------
    children |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.323079   .0212812    17.41   0.000      1.28202    1.365454
       _cons |   1.502395    .018088    33.81   0.000     1.467359    1.538269
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

When we look at the results for sex, we see that the incidence rate ratio (IRR) is 1.32. Thus, one unit increase in sex is associated with a higher rate of children. This means that women have a rate of children that is 1.32 times higher compared to that of men.

The association between sex and children is statistically significant, as reflected in the p-value (0.000) and the 95% confidence intervals (1.28-1.37).

Summary

Women have a statistically significantly higher rate of children, compared to men (IRR=1.32; 95% CI=1.28-1.37).

16.3.3 Simple Poisson regression with a categorical (non-binary) x

Theoretical examples

Example 1

We conduct a study among people who subscribe to a fishing magazine, focusing on the association between experience of fishing (x) and the number of fishes caught during the individual's last fishing expedition (y). Experience of fishing has three categories: 1=Low level, 2=Medium level, and 3=High level. Low level is chosen as the reference category. The number of catches ranges between 0 and 30. We find that the IRR is 1.50 for Medium level and 2.03 for High level. This means that individuals with more experience have a higher rate of catches.

Example 2

In this example, we examine the association between temperament (x) and the number of cigarettes smoked per week (y). Temperament is categorised as: 1=Sanguine, 2=Choleric, 3=Melancholic, and 4=Phlegmatic. Phlegmatic is chosen as the reference category. The number of cigarettes ranges from 0 to 150. We find that the IRR is 0.81 for Melancholic, 1.29 for Choleric, and 3.73 for Sanguine. In other words, individuals with melancholic temperament have a lower rate of cigarette smoking compared to the phlegmatic, whereas the opposite is true for individuals whose temperament is characterised as choleric or sanguine.

Practical example

Dataset: StataData1.dta

Name	**Label**
children	Number of children (Age 40, Year 2010)
educ	Educational level (Age 40, Year 2010)

```
sum children educ if pop_poisson==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    children |      9,014    1.740515    1.552944          0         10
        educ |      9,014    2.174728     .725944          1          3
```

The variable educ has three categories: 1=Compulsory, 2=Upper secondary, and 3=University. Here, we (with ib1) specify that the first category (Compulsory) will be the reference category.

```
poisson children ib1.educ if pop_poisson==1, irr
```

```
Poisson regression                              Number of obs     =      9,014
                                                LR chi2(2)        =     147.29
                                                Prob > chi2       =     0.0000
Log likelihood = -16054.463                     Pseudo R2         =     0.0046

---------------------------------------------------------------------------------
       children |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
           educ |
Upper secondary |   1.171898   .0273307     6.80   0.000     1.119537    1.226709
     University |   1.318967   .0310659    11.75   0.000     1.259463    1.381283
                |
          _cons |    1.45913   .0290839    18.96   0.000     1.403226    1.517262
---------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

When we look at the results for the dummies for educ, we see that the incidence rate ratios are 1.17 for Upper secondary and 1.32 for University. Thus, having a higher level of educational attainment is associated with a higher rate of children.

Both dummies for educ are significantly different from the reference category, as reflected in the p-values and the 95% confidence intervals.

Test the overall effect

The output presented and interpreted above, is based on the relative rate ratios for the dummy variables of educ. Let us also assess the overall statistical effect of educ on children? We can assess it through contrast, which is a postestimation command.

```
contrast p.educ, noeffects
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
        educ |
   (linear)  |          1      138.16     0.0000
(quadratic)  |          1        1.43     0.2317
      Joint  |          2      144.22     0.0000
------------------------------------------------
```

Here, we focus on the row for linear, which shows a p-value (P>chi2) below 0.05. This suggests that we have a statistically significant trend in children according to educ.

More information	help contrast

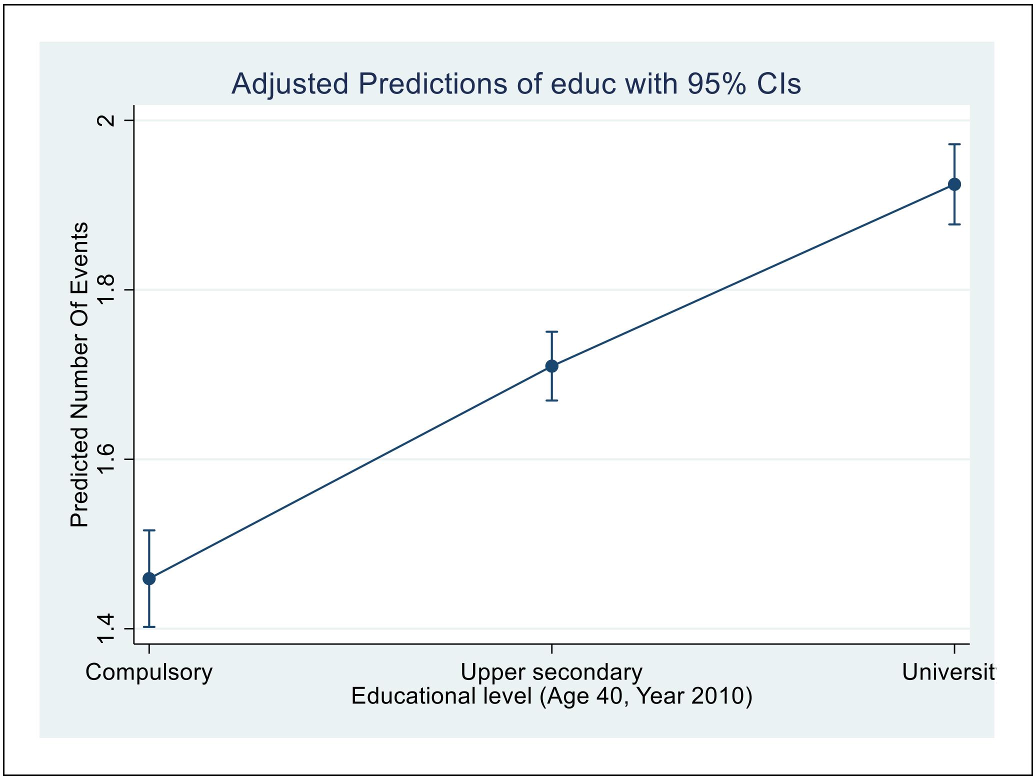
We will also produce a graph of the trend. First, however, we need to apply the post-estimation command margins.

Note This command can also be used for variables that are continuous or binary, but is particularly useful for categorical, non-binary (i.e. ordinal) variables.

```
margins educ
```

```
Adjusted predictions                            Number of obs     =      9,014
Model VCE    : OIM

Expression   : Predicted number of events, predict()

----------------------------------------------------------------------------------
                 |            Delta-method
                 |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
            educ |
     Compulsory  |    1.45913   .0290839    50.17   0.000     1.402127    1.516134
Upper secondary  |   1.709952   .0207043    82.59   0.000     1.669373    1.750532
     University  |   1.924545   .0241494    79.69   0.000     1.877213    1.971877
----------------------------------------------------------------------------------
```

```
marginsplot
```



Note The y-axis shows predicted number of events (i.e. not log incidence rates or incidence rate ratios).

This graph quite clearly shows that the higher the level of educational attainment, the higher the number of children.

More information	help marginsplot

Summary
At age 40, there is a clear, and statistically significant, trend in the rate of children according to educational level: higher levels of education are associated with a higher rate of children.

16.4 Multiple Poisson regression

Quick facts	
Number of variables	One dependent (y) At least two independent (x)
Scale of variable(s)	Dependent: count Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

Theoretical example

Example

Suppose we are interested to see if having young children (x), residential area (x), and income (x) are related to the number of pets owned (y). Having young children is measured as either 0=No young children and 1=Young children. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. We choose Metropolitan as our reference category. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK). The number of pets owned ranges between 0 and 50.

We get an IRR for Young children that is 1.23. That means that those who have young children are have a higher rate of pets, compared to those who do not have young children. This association is adjusted for residential area and income.

With regards to residential area, we get an IRR for Smaller city of 1.30, whereas the IRR for Rural is 7.44. This suggests that those who live in a smaller city have a higher rate of pets compared to those living in metropolitan areas, and so are those living in rural areas. These results are adjusted for having young children and income.

Finally, the IRR for income is 0.98. This suggests that for every unit increase in income (i.e. for every additional one thousand SEK), the rate of pets decreases. This association is adjusted for having young children as well as residential area.

Practical example

Dataset: StataData1.dta

Name	**Label**
children	Number of children (Age 40, Year 2010)
siblings	Number of siblings (Age 15, Year 1985)
sex	Sex
educ	Educational level (Age 40, Year 2010)

```
sum children siblings sex educ if pop_poisson==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    children |      9,014    1.740515    1.552944          0         10
    siblings |      9,014    1.784114    1.327277          0         10
         sex |      9,014    .4905702    .4999388          0          1
        educ |      9,014    2.174728     .725944          1          3
```

```
poisson children siblings sex ib1.educ if pop_poisson==1, irr
```

```
Poisson regression                              Number of obs     =      9,014
                                                LR chi2(4)        =     431.93
                                                Prob > chi2       =     0.0000
Log likelihood = -15912.144                     Pseudo R2         =     0.0134

----------------------------------------------------------------------------------
        children |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
        siblings |   1.016487   .0061187     2.72   0.007     1.004565     1.02855
             sex |   1.304617   .0210647    16.47   0.000     1.263977    1.346563
                 |
            educ |
 Upper secondary |   1.147301    .026924     5.86   0.000     1.095726    1.201303
      University |   1.291208   .0307602    10.73   0.000     1.232305    1.352926
                 |
           _cons |   1.253923   .0311061     9.12   0.000     1.194415    1.316397
----------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

In this model, we have three x-variables: siblings, sex, and educ. When we put them together, their statistical effect on educ is mutually adjusted.

When it comes to the incidence rate ratios, they have changed in comparison to the simple regression models. For example, the odds ratio for siblings has increased marginally 1.01 to 1.02. The incidence rate ratio for sex has become slightly closer to 1 (from 1.32 to 1.30). This is also the case for the dummies of educ: the incidence risk ratio for Upper secondary has changed from 1.17 to 1.15 and the one for University has changed from 1.32 to 1.29.

The association between the siblings and children has become statistically significant ($p<0.05$) after mutual adjustment. However, it was very close to being significant also in the simple model ($p=0.07$). The associations between sex and children on the one hand, and between educ and children on the other hand, are still statistically significant.

Note A specific incidence risk ratio from a simple Poisson regression model can increase when other x-variables are included. Usually, it is just “noise”, i.e. not any large increases, and therefore not much to be concerned about. But it can also reflect that there is something going on that we need to explore further. There are many possible explanations for increases in multiple regression models: a) We actually adjust for a confounder and then “reveal” the “true” statistical effect. b) There are interactions among the x-variables in their effect on the y-variable. c) There is something called collider bias (which we will not address in this guide) which basically mean that both the x-variable and the y-variable causes another x-variable in the model. d) The simple regression models and the multiple regression model are based on different samples. e) It can be due to rescaling bias (see Chapter 18).

Summary

In the fully adjusted model, it can be observed that the association between the number of siblings and the number of children at age 40 now reaches a statistically significant level (IRR=1.02; 95% CI=1.00-1.03). The associations between sex and number of children as well as between educational level and number of children have become somewhat attenuated, but remain statistically significant.

Estimates table and coefficients plot

If we have multiple models, we can facilitate comparisons between the regression models by asking Stata to construct estimates tables and coefficients plots. What we do is to run the regression models one-by-one, save the estimates after each, and than use the commands estimates table and coefplot.

The coefplot option is not part of the standard Stata program, so unless you already have added this package, you need to install it:

```
ssc install coefplot
```

As an example, we can include the three simple regression models as well as the multiple regression model. The quietly option is included in the beginning of the regression commands to suppress the output.

Run and save the first simple regression model:

```
quietly poisson children siblings if pop_poisson==1, irr
```

```
estimates store model1
```

Run and save the second simple regression model:

```
quietly poisson children sex if pop_poisson==1, irr
```

```
estimates store model2
```

Run and save the third simple regression model:

```
quietly poisson children ib1.educ if pop_poisson==1, irr
```

```
estimates store model3
```

Run and save the multiple regression model:

```
quietly poisson children siblings sex ib1.educ if pop_poisson==1, irr
```

```
estimates store model4
```

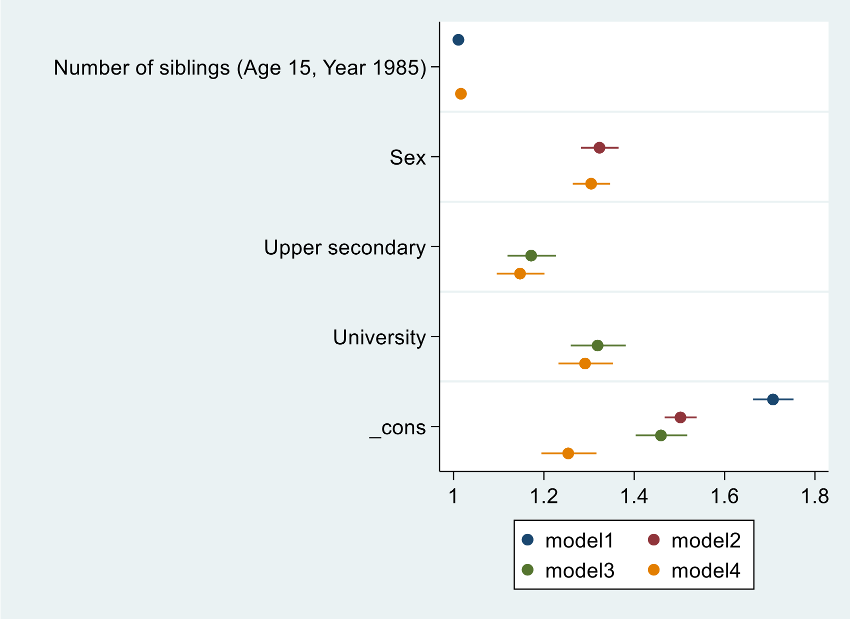
Produce the estimates table (include the option eform to show the incidence rate ratios):

```
estimates table model1 model2 model3 model4, eform
```

```
------------------------------------------------------------------
    Variable |   model1       model2       model3       model4
-------------+----------------------------------------------------
    siblings |  1.0108444                              1.0164867
         sex |               1.3230793                 1.3046169
             |
        educ |
Upper sec..  |                            1.1718982    1.1473006
 University  |                            1.3189674    1.2912078
             |
       _cons |  1.7071654    1.5023955    1.4591304    1.2539235
------------------------------------------------------------------
```

Produce the coefficients plot (include the option eform to show the incidence rate ratios):

```
coefplot model1 model2 model3 model4, eform
```



Note You can improve the graph by using the Graph Editor to delete "_cons" as well as to adjust the category and label names.

16.5 Model diagnostics

The assumptions behind Poisson regression are similar to the ones we have for other types of generalised linear models. In addition, we also assume that there is no overdispersion or zero inflation.

More information	help poisson postestimation

Checklist	
Count outcome	The y-variable has to be a count.
Independence of errors	Data should be independent, i.e. not derived from any dependent samples design, e.g. before-after measurements/paired samples.
Correct model specification	Your model should be correctly specified. This means that the x-variables that are included should be meaningful and contribute to the model. No important (confounding) variables should be omitted (often referred to as omitted variable bias).
No multicollinearity	Multicollinearity may occur when two or more x-variables that are included simultaneously in the model are strongly correlated with each another. Actually, this does not violate the assumptions, but is does create greater standard errors which makes it harder to reject the null hypothesis.
No overdispersion	The mean should be equivalent to the variance.
No zero inflation	The difference between observed zeros and predicted zeros is small.

Types of model diagnostics	
Link test	Assess model specification
Correlation matrix	Check for multicollinearity
Deviance goodness-of-fit test and Pearson goodness-of-fit test	Assess model fit (no overdispersion or zero inflation)

16.5.1 Link test

With the command linktest, we can assess whether our model is correctly specified. This test uses the linear predicted value (called _hat) and the linear predicted value squared (_hatsq) to rebuild the model. We expect _hat to be statistically significant, and _hatsq to be statistically non-significant. If one or both of these expectations are not met, the model is mis-specified.

However, do not rely too much on this test – remember that you should also use theory and common sense to guide your decisions. It is very seldom relevant to focus on this test if our ambition is to investigate associations (and not to make the best possible prediction of the outcome).

More information	help linktest

Practical example

We perform this test for the full model, so let us go back to the example from the multiple regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly poisson children siblings sex ib1.educ if pop_poisson==1, irr
```

And then we run the test:

```
linktest
```

```
Poisson regression                              Number of obs     =      9,014
                                                LR chi2(2)        =     438.94
                                                Prob > chi2       =     0.0000
Log likelihood =  -15908.64                     Pseudo R2         =     0.0136

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   1.893997   .3425963     5.53   0.000     1.222521    2.565474
      _hatsq |  -.8193392   .3105512    -2.64   0.008    -1.428008   -.2106701
       _cons |  -.2214068   .0889658    -2.49   0.013    -.3957765   -.0470371
------------------------------------------------------------------------------
```

Although the p-value for the variable _hat is below 0.05, the p-value for _hatsq is also below 0.05, which means that our model is mis-specified.

We could try to amend this by transforming any of the included variables (e.g. through categorisation, or log transformation), excluding any of the included variables, or adding more variables to the model (other x-variables or e.g. interactions between the included variables).

Of course, this should be explored before we continue to assess model fit – but for the sake of simplicity, we will ignore this problem in the following sections.

16.5.2 Correlation matrix

As the x-variables become more strongly correlated, it becomes more difficult to determine which of the variables are actually producing the statistical effect on the y-variable. This is the problem with multicollinearity.

One way of assessing multicollinearity is using the estat vce command, with the corr (short for correlation) option.

More information	help estat vce

Practical example

The first step is re-run the multiple Poisson regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly poisson children siblings sex ib1.educ if pop_poisson==1, irr
```

Next, we try the estat vce command. By adding the corr (=correlation) option, we will get a correlation matrix instead of a covariance matrix.

```
estat vce, corr
```

```
Correlation matrix of coefficients of poisson model

             | children
             |                               2.         3.
        e(V) | siblings        sex       educ       educ      _cons
-------------+-------------------------------------------------------
children     |
    siblings |   1.0000
         sex |  -0.0399     1.0000
      2.educ |   0.0878    -0.0717     1.0000
      3.educ |   0.1304    -0.0794     0.7272     1.0000
       _cons |  -0.5098    -0.2868    -0.7062    -0.7159     1.0000
```

The table shows the correlations between the different variables/categories. In line with the earlier sections on correlation analysis (see Chapter 7.2), we can conclude that the coefficients suggest (very) weak correlations here. The only exceptions are two of the dummies for educ, which is not a huge problem since they reflect the same underlying variable.

16.5.3 Deviance goodness-of-fit test and Pearson goodness-of-fit test

There are two critical assumptions that we have to test. First, that there is no problem with overdispersion (or underdispersion, for that matter), which means that the assumption of mean=variance is violated. Second, that there is no problem with zero inflation (i.e. excess zeros).

To test model fit we can use the estat gof command, which relies on postestimation.

More information	help estat gof

Practical example

The first step is re-run the multiple Poisson regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly poisson children siblings sex ib1.educ if pop_poisson==1, irr
```

Then we use the estat gof command, which produces quite some output:

```
estat gof
```

```
         Deviance goodness-of-fit =  14908.76
         Prob > chi2(9009)        =    0.0000

         Pearson goodness-of-fit  =  12364.43
         Prob > chi2(9009)        =    0.0000
```

The fact that the p-values are below 0.05 means that the Poisson model does not fit our data, and we should explore other alternatives, such as negative binomial regression or zero-inflated Poisson regression. These will be presented in the following section.

16.6 Alternatives to Poisson regression

We will explore two alternatives to Poisson regression:

- Negative binomial regression (nbreg command).
- Zero-inflated Poisson regression (zip command).

These will subsequently be compared using the countfit command.

16.6.1 Negative binomial regression model

The negative binomial regression model (nbreg command) is similar to a Poisson regression, only that the variance is allowed to be greater than what is assumed in a Poisson model. This extra variance is the overdispersion. If not accounted for, overdispersion leads to deflated standard errors which in turn may lead to errenous inference.

More information	help nbreg

Practical example

Let us first do a simple check to see what the situation looks like regarding overdispersion for our outcome children. Of course, this will not take any x-variable into consideration.

```
sum children, detail
```

```
           Number of children (Age 40, Year 2010)
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs               9,053
25%            0              0       Sum of Wgt.       9,053

50%            2                      Mean           1.737214
                        Largest       Std. Dev.      1.552791
75%            3              9
90%            4              9       Variance       2.411161
95%            4              9       Skewness       .5696771
99%            6             10       Kurtosis       2.742044
```

The variance is considerably higher than the mean, which suggests that overdispersion might be an issue. Accordingly, it is a good idea to try out a negative binomial regression model.

Thus, we will re-run the multiple regression model that we specified for Poisson regression earlier, but now with the nbreg command:

```
nbreg children siblings sex ib1.educ if pop_poisson==1, irr
```

```
Negative binomial regression                    Number of obs     =      9,014
                                                LR chi2(4)        =     287.03
Dispersion     = mean                           Prob > chi2       =     0.0000
Log likelihood = -15641.938                     Pseudo R2         =     0.0091

------------------------------------------------------------------------------------
        children |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------------
        siblings |   1.017336   .0075522     2.32   0.021     1.002641    1.032247
             sex |   1.306172    .025613    13.62   0.000     1.256924    1.357349
                 |
            educ |
Upper secondary  |   1.150272   .0322482     4.99   0.000     1.088772    1.215246
     University  |   1.294558   .0370545     9.02   0.000     1.223932    1.369259
                 |
           _cons |   1.248594    .037102     7.47   0.000     1.177953    1.323472
-----------------+------------------------------------------------------------------
        /lnalpha |  -1.284739   .0564131                     -1.395307   -1.174171
-----------------+------------------------------------------------------------------
           alpha |   .2767228   .0156108                       .247757     .309075
------------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline incidence rate.
LR test of alpha=0: chibar2(01) = 540.41               Prob >= chibar2 = 0.000
```

The output is very similar to the one we got for the Poisson regression. Additionally, we are presented with the results from the log-transformed overdispersion parameter (/lnalpha), as well as the untransformed estimate (alpha).

Note that we also get a LR test presented below the table, which compares this model to a Poisson model. The fact that the p-value (Prob >= chibar2) is below 0.05 (0.000) suggests that this model fits the data better than the traditional Poisson.

16.6.2 Zero-inflated Poisson regression

The zero-inflated Poisson regression models the data in two steps. The first step assumes that the excess zero counts come from a logit model (this is default), whereas the remaining counts come from a Poisson model.

More information	help zip

Practical example

As the first step, we need to generate a variable that specifies whether the outcome is a zero (value 1) or not (value 0), despite that it might seem a bit backwards.

```
gen nochildren=children
```

```
recode nochildren (0=1) (1/10=0)
```

Then we will re-run the multiple regression model that we specified for Poisson regression earlier, but now with the zip command. Here, we must also specify the inflate option, where we include the variable – nochildren – that we just generated.

```
zip children siblings sex ib1.educ if pop_poisson==1, irr inflate(nochildren)
```

```
Zero-inflated Poisson regression                Number of obs     =      9,014
                                                Nonzero obs       =      6,215
                                                Zero obs          =      2,799

Inflation model = logit                         LR chi2(4)        =      31.48
Log likelihood  = -10278.99                     Prob > chi2       =     0.0000

----------------------------------------------------------------------------------
        children |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
children         |
        siblings |   1.004593    .006135     0.75   0.453     .9926401    1.016689
             sex |   1.034321   .0166694     2.09   0.036      1.00216    1.067514
                 |
            educ |
Upper secondary  |   1.068627   .0250581     2.83   0.005     1.020625    1.118886
     University  |   1.124861   .0267493     4.95   0.000     1.073636     1.17853
                 |
           _cons |   2.278097   .0570867    32.86   0.000     2.168913    2.392778
-----------------+----------------------------------------------------------------
inflate          |
      nochildren |   52.99146   17928.82     0.00   0.998    -35086.85    35192.83
           _cons |  -25.62297    4647.69    -0.01   0.996    -9134.928    9083.682
----------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline incidence rate.
```

The output is very similar to the one we got for the Poisson regression. There is a part of the table called “inflate”; this refers to the estimate for the variable predicting the excess zeros. The estimate of 52.99 suggests that for each unit increase in nochildren (i.e. being a zero vs not being a zero), there is a large increase in IRR.

16.6.3 Compare fit of alternative count models

The command countfit will be used to compare the fit of the Poisson regression, negative binomial regression, and zero-inflated Poisson regression. This produces a number of different estimates and tests, as well as a graph.

More information	help countfit

Practical example

We will adapt the multiple regression model that we have worked with earlier:

```
countfit children siblings sex ib1.educ if pop_poisson==1, prm nbreg zip
```

Note prm=Poisson, nbreg=negative binomial, and zip=zero-inflated.

This will produce a serious amount of output. We will include one portion at a time here.

```
--------------------------------------------------------------------
                      Variable |    PRM        NBRM         ZIP
-------------------------------+------------------------------------
children                       |
Number of siblings (Age 15, ~r |     1.016       1.017       1.005
                               |      2.72        2.32        0.73
                           Sex |     1.305       1.306       1.047
                               |     16.47       13.62        2.43
                               |
Educational level (Age 40, Y~  |
              Upper secondary  |     1.147       1.150       1.091
                               |      5.86        4.99        3.15
                   University  |     1.291       1.295       1.168
                               |     10.73        9.02        5.59
                      Constant |     1.254       1.249       1.975
                               |      9.12        7.47       22.90
-------------------------------+------------------------------------
lnalpha                        |
                      Constant |                 0.277
                               |                -22.77
-------------------------------+------------------------------------
inflate                        |
Number of siblings (Age 15, ~r |                             0.949
                               |                             -1.90
                           Sex |                             0.360
                               |                            -13.77
                               |
Educational level (Age 40, Y~  |
              Upper secondary  |                             0.821
                               |                             -2.24
                   University  |                             0.644
                               |                             -4.74
                      Constant |                             0.636
                               |                             -4.92
-------------------------------+------------------------------------
Statistics                     |
                         alpha |                 0.277
                             N |      9014        9014        9014
                            ll | -1.59e+04   -1.56e+04   -1.51e+04
                           bic | 31869.821   31338.515   30275.166
                           aic | 31834.288   31295.876   30204.101
--------------------------------------------------------------------
                                                        legend: b/t
```

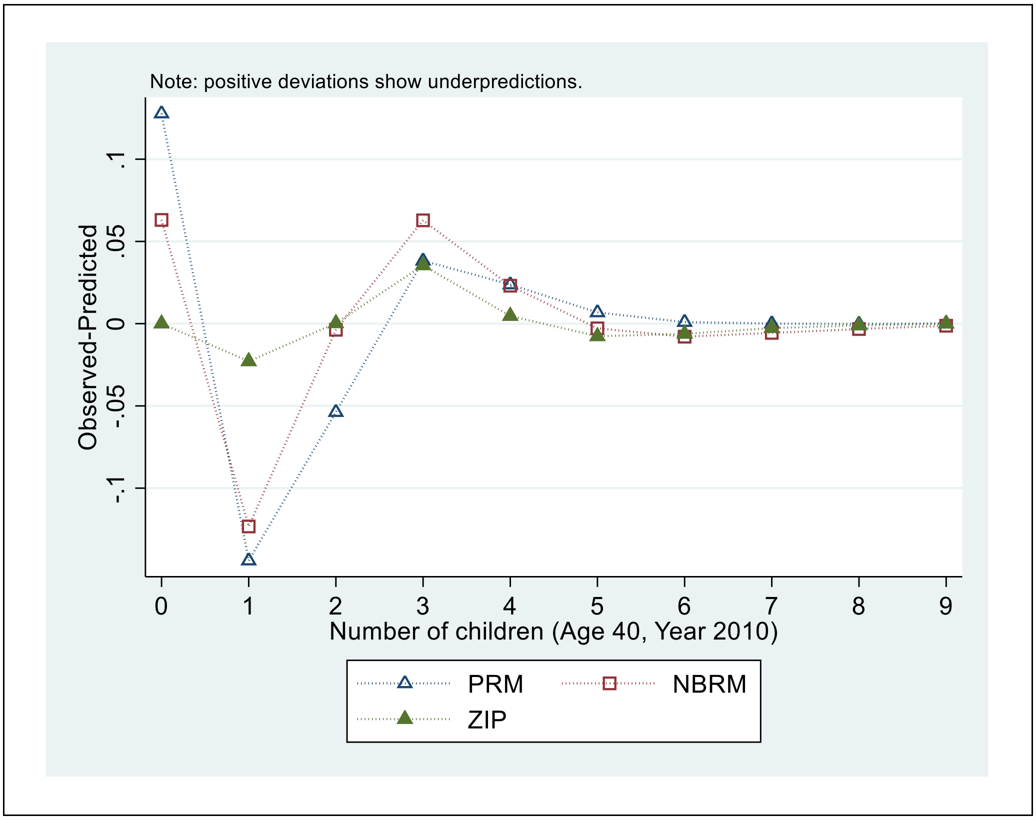
The most interesting here is Statistics part, where we get the values for the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). These can be seen as relative measure of model fit: the lower the values, the better. We note that the BIC and AIC values are lowest for the zero-inflated Poisson model, indicating that this model has the best fit.

```
Tests and Fit Statistics

PRM            BIC= 31869.821  AIC= 31834.288  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs NBRM      BIC= 31338.515  dif=   531.305  NBRM    PRM   Very strong
               AIC= 31295.876  dif=   538.412  NBRM    PRM
               LRX2=  540.412  prob=    0.000  NBRM    PRM   p=0.000
-------------------------------------------------------------------------
  vs ZIP       BIC= 30275.166  dif=  1594.655  ZIP     PRM   Very strong
               AIC= 30204.101  dif=  1630.187  ZIP     PRM
               Vuong=       .  prob=        .  ZIP     PRM   p=.
-------------------------------------------------------------------------
NBRM           BIC= 31338.515  AIC= 31295.876  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs ZIP       BIC= 30275.166  dif=  1063.349  ZIP     NBRM  Very strong
               AIC= 30204.101  dif=  1091.776  ZIP     NBRM
-------------------------------------------------------------------------
ZIP            BIC= 30275.166  AIC= 30204.101  Prefer  Over  Evidence

Vuong test is not appropriate for testing zero-inflation. To force the
the computation of the test, use option -forcevuong-.
```

This part of the output ties back to the BIC and AIC statistics. The results here support what we already concluded: the zero-inflated Poisson model is preferable over the Poisson model and negative binomial model.

```
Comparison of Mean Observed and Predicted Count

            Maximum       At      Mean
Model      Difference    Value    |Diff|
---------------------------------------------
PRM         -0.144         1      0.040
NBRM        -0.123         1      0.030
ZIP          0.036         3      0.008

PRM: Predicted and actual probabilities

Count    Actual     Predicted     |Diff|    Pearson
------------------------------------------------
0         0.311       0.183       0.128    805.387
1         0.159       0.303       0.144    616.498
2         0.205       0.259       0.054    100.694
3         0.189       0.151       0.038     87.188
4         0.092       0.068       0.024     75.072
5         0.032       0.025       0.007     16.819
6         0.009       0.008       0.001      1.099
7         0.002       0.002       0.000      0.010
8         0.000       0.001       0.000      0.670
9         0.000       0.000       0.000      3.555
------------------------------------------------
Sum       1.000       1.000       0.396   1706.993

NBRM: Predicted and actual probabilities

Count    Actual     Predicted     |Diff|    Pearson
------------------------------------------------
0         0.311       0.247       0.063    144.851
1         0.159       0.283       0.123    484.864
2         0.205       0.209       0.004      0.593
3         0.189       0.127       0.063    280.898
4         0.092       0.069       0.023     69.911
5         0.032       0.035       0.003      2.095
6         0.009       0.017       0.008     33.611
7         0.002       0.008       0.006     36.249
8         0.000       0.004       0.003     26.398
9         0.000       0.002       0.001      8.972
------------------------------------------------
Sum       1.000       0.999       0.297   1088.441

ZIP: Predicted and actual probabilities

Count    Actual     Predicted     |Diff|    Pearson
------------------------------------------------
0         0.311       0.310       0.000      0.000
1         0.159       0.182       0.023     25.899
2         0.205       0.205       0.000      0.003
3         0.189       0.154       0.036     73.775
4         0.092       0.087       0.005      2.286
5         0.032       0.040       0.008     13.320
6         0.009       0.015       0.006     22.800
7         0.002       0.005       0.003     13.326
8         0.000       0.001       0.001      7.409
9         0.000       0.000       0.000      0.019
------------------------------------------------
Sum       1.000       1.000       0.081    158.836
```

The first part shows a comparison between the mean observed and predicted counts, followed by the predicted and actual probabilities for each value of children, separate for the three models. The maximum difference is smallest for the zero-inflated Poisson model, which supports this as the preferable model. It is also interesting to note the difference between the actual and predicted probabilities specifically for the value 0, for the three models. Again, the difference is smallest for the ZIP model.



This graph plots the observed (actual) *minus* predicted probabilities seen in the table. Positive deviations reflect underprediction and negative deviations reflect overprediction. The negative binomial (NBRM) and Poisson (PRM) models tend to underpredict the value 0 and overpredict the value 1, while the zero-inflated Poisson model does a better job across the whole scale.

16.7 Hurdle regression

Finally, we would like to make you aware that a viable alternative to zero-inflated Poisson regression is hurdle regression. Like zero-inflated Poisson regression, hurdle regression will model the outcome in two steps. But where the first step in the zero-inflated Poisson regression predicts whether the outcome is 0, the first step in the hurdle regression predicts whether the outcome is 1.

Assume that we are interested in predicting the number of months an individual has received means-tested social assistance. Means-tested social assistance is a relatively rare outcome (at least at population level), so the vast majority of individuals have not received any social assistance. Since data include both recipients and non-recipients, the first model (typically a logistic regression model) determines whether one has received social assistance, and the second model (typically a model for count data) determines the number of months in receipt of social assistance given that one has received benefits (i.e. when the 'hurdle' has been crossed). Such an approach thus allows for testing hypotheses about whether there are different processes governing the occurrence and the continuation of the outcome.

Hurdle regression comes in many versions, of which Poisson-logit and negative binomial-logit are two examples. With the commands hplogit and hnblogit, we can produce the Poisson-logit and negative binomial-logit versions of the hurdle model. These commands require separate installations. We will not go through these here, but if you are interested, we suggest that you install hplogit and hnblogit and then review the help files.

17. COX REGRESSION

Outline

Content

This chapter starts with an introduction to Cox regression and then presents the function in Stata. After this, we offer some practical examples of how to perform simple and multiple Cox regression, as well as how to generate and interpret model diagnostics.

17.1 Introduction

Cox regression is used when the outcome is time-to-event. Accordingly, the outcome can be said to consist of two components of information: whether an event has occurred or not, and the time at risk (i.e. the time up until the event has occurred). The event itself can be of any sort, such as death, hospitalization, job loss, or childbirth.

Some examples

- Time from birth to death.
- Time from marriage to divorce.
- Time from cancer diagnosis to death from cancer.
- Time from admission to hospital to discharge from hospital.
- Time from start of the game to the first goal.

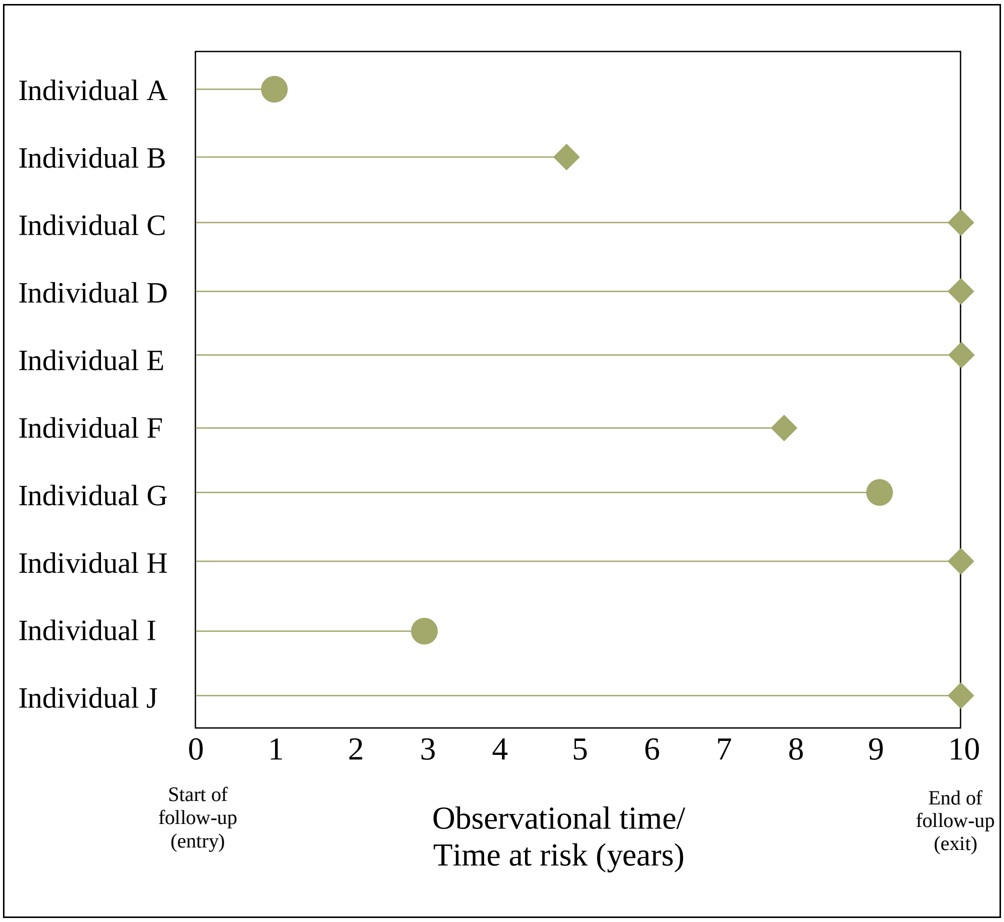
Cox regression analysis is a type of survival analysis. This term makes it sound like it is all about life and death – but survival analysis can actually be applied to any type of time-to-event data. It should also be mentioned that survival analysis goes by different names depending on discipline and research field (life table analysis, hazard analysis, duration analysis, transition analysis, event history analysis, etc). We personally prefer the term time-to-event analysis.

17.1.1 Observational time and censoring

A key ingredient in survival analysis is observational time (or time at risk), i.e. the period during which an individual (or any other type of observation) is actually observed. The observational period (also called follow-up period) starts when the individual enters the study and ends at: a) the occurrence of the event, b) by the end of follow-up, c) at loss to follow-up/dropout, or d) in the case of death. In relation to these points, we need to discuss the concept of censoring.

Censoring	
Left-censoring	The term left-censoring applies to situations when we know that the event occurred prior to the start of the observation period, but we cannot be sure about the exact time the event occurred.
Right-censoring	In cases b-d, as specified above this table, the term right-censoring would apply. It means that we are only able to know anything about the development of the individual up until that specific point (we do not know if or when the event will happen afterwards). Put differently, right-censoring refers to a situation where an individual can no longer be observed and the event has not occurred during the observational/follow-up period.
Interval-censoring	Interval-censoring refers to instances when we know an event occurred between two time points, but we cannot be sure when the event occurred within that interval.

Below is an illustration of what (right-)censoring might look like. Lines that end with a circle denote that the event has taken place (individuals A, G, and I), whereas the lines that end with a diamond denote that the individual has been censored. In the case of individuals C, D, E, H, and J, they are censored at the end of follow-up. Individuals B and F have been censored earlier than that, perhaps because of death or emigration.



One assumption that we need to make is that the censoring is non-informative. In other words, we assume that those that are censored are not being censored because they have a lower or higher risk of the event itself. Such violations are difficult to formally test but are probably quite common. For example, it is likely that individuals that are censored because of death would be more likely to have experienced the event at some point if they had not died (at least if the event is related to health in some way).

To summarise, it is necessary to have precise information about when the follow-up starts, when it ends, and when the event occurred. There must also be an unambiguous definition of the time scale. Moreover, we need to consider censoring (most survival analyses have right-censoring), since our results might be biased otherwise.

Note There is also a second concept in survival analysis, often confused with censoring, called truncation. Truncation refers to situations when the observation period for certain individuals is smaller or larger than your study's observation period. As you cannot observe these individuals (and researchers are therefore not aware of their existence) they are not included. This may also introduce bias into your results. Left and right truncation is quite common in health sciences, and special methods are required to deal with it, but these will not be covered in this guide.

17.1.2 Survival function

In survival analysis, the survival function plays a central role. This can be defined as the probability that an individual survives beyond time t. While t can range from 0 to infinity (∞), the survival function is restricted to vary between 0 and 1. Moreover, the probability of survival at t=0 is 1, whereas the probability of survival goes to 0 when t=∞. In theory, the survival function is smooth, but we usually observe events on a discrete time scale (e.g. years, months, days).

17.1.3 Hazard function

We also have something called the hazard function (or failure function): the instantaneous rate of failure at time t, conditional upon the fact that the event has not yet occurred. A related concept is the cumulative hazard, which describes the accumulated risk up until time t. Neither the hazard function nor the cumulative hazard function is a measure of probability but can be thought of as a measure of risk: the greater the value, the greater the risk of failure.

17.1.4 Tied failure times

Tied failure times, or "ties," refers to instances when two (or more) individuals experience the health event, or are censored, at the same time. The number of ties in your data may depend on, e.g., how detailed your time variables are (exact times have a lower likelihood of ties) and what process you are modelling (relative to others, some events may have a higher likelihood of ties). Since the outcome is often measured on a continuous scale when using Cox regression, we assume that ties are relatively rare. Tied failure times will be discussed in more detail in Section 17.2. Survival models also have different approaches for dealing with ties; these methods are outlined in Section 17.8.6.

17.1.5 Non-parametric, parametric, and semi-parametric models

It is easy to estimate and graph the survival function and hazard function: we can use non-parametric methods such as the Kaplan-Meier product-limit estimator (see Section 17.4.1).

Alternatively, we can estimate the survival distribution based on parametric regression models – in this context often referred to accelerated failure time models (or location-scale models). Within this framework, there are many different types which all assume different shapes of the distribution (e.g. exponential, Weibull, log-normal, log-logistic, Gompertz, and generalised gamma). While these will not be covered in this guide, you can explore them in Stata if you want to:

More information	help stintreg

Then we have the proportional hazards model – or simply Cox regression – which is a semi-parametric type of model. Unlike non-parametric methods, proportional hazards models can account for more of the detail of the data. Additionally, they are more flexible compared to parametric models since there are fewer assumptions.

Note For many (if not most) variables that capture events (i.e. case vs non-case), observational time/time at risk is relevant to consider. Yet, this information is not always available. Even when it is available, many researchers do not make any use of it and instead perform analyses suitable for binary outcomes, e.g. logistic regression.

Note Although time-to-event is a continuous variable, it is seldom a feasible alternative to apply a linear regression. This is primarily due to the incapability of linear regression models to account for censoring, but also because time-to-event variables often have a skewed distribution.

Note In some instances, Poisson regression is a viable alternative to Cox regression. This is for example the case when data are grouped (i.e. aggregated).

17.2 The Cox regression model

The Cox regression model is used to measure the effects of one or more risk factors (or exposures; x-variables) on the hazard rate. The hazard rate is the effect measure in Cox models: the risk of the occurrence of a health event, given the individual's survival until that timepoint. The basic Cox model notation states that the hazard at time *t* is equal to the baseline hazard at time *t* multiplied by the exponentiated product of the vector of regression coefficients and the vector of covariates. Let's dig in!

As we mentioned in Section 17.1.5, health events modelled using parametric models (e.g., exponential, Weibull, Gompertz, and Poisson) are each assumed to have a distinct distribution that is described by one or several parameters. In other words, the baseline hazard for each of these models has a distinct shape and varies in a specific way. Using parametric survival models require that you understand the assumption(s) underlying the shape of the respective distribution and have an idea that the baseline hazard in your data approximately follows this shape.

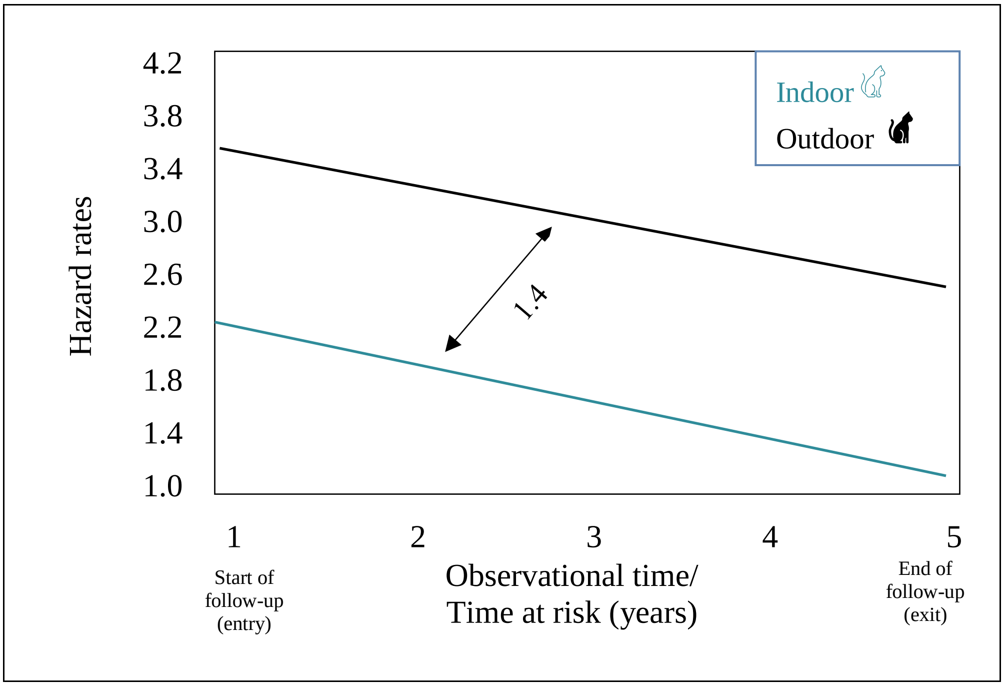
Cox regression is a semi-parametric approach; the model does contain a parametric component, but also a non-parametric component. In Cox models, the baseline hazard function is non-parametric: it can wander freely with no parameters. This means that the Cox model does not make any assumptions about the shape of the baseline hazard or the distribution of survival times. In fact, estimating the baseline hazard is not needed to make inferences about the relative hazard rates. Unlike parametric models where we need to be very careful about which model we specify, we do not need to specify a distribution for the Cox model. Importantly though, *the baseline hazard*, no matter the shape, *is assumed to be the same for every unit of observation*. Summary so far: the baseline hazard at time *t* is the value of the hazard when all covariates are equal to zero, the baseline hazard does not have a specific shape but is the same for all units of observation, and it is not assumed that survival times follow a specific distribution.

In Cox models, the covariate vector (or covariate function; group of one or more covariates, x-variables) is modelled parametrically. Covariates influence the baseline hazard in a specific way. The hazard function is multiplied by the covariate vector to obtain the effect of the covariates. The covariate vector induces a multiplicative and proportional shift in the baseline hazard, but does not change the shape of the baseline hazard. Furthermore, the multiplicative effect of the covariates is not time dependent; it is the same at any time *t* during the follow-up period. Please note that this is true for fixed covariates, or covariates that do not depend on time. Cox regression can also be used to model time-dependent covariates, which may vary over time, but we will not discuss time-dependent covariates in this guide. The effect of the covariates underlies a very important assumption in Cox regression: the proportional hazards assumption.

Proportional hazards

Now we know that the effect of any (group of) covariate(s) is the same at any point in time during follow-up. Therefore, the relationship between the covariates and the event (outcome; dependent variable; y-variable) is constant. This means that, for any two units of observation (for example, any two individuals), the ratio of the hazard functions is constant and dependent on the covariate values. In other words, the hazard functions are *proportional* to one another, at any point in time. The estimated hazard ratio (which compares the hazard functions of one individual to another) does not vary over time, *even if* the *size* of the hazard changes (e.g., increases, decreases, increases then decreases, etc.) or remains constant. Therefore, if the proportional hazards assumption holds, the hazard curves for any two individuals should be proportional to one another over time. When graphed, these curves should be parallel to each other, and definitely should not cross. Since this is a very important assumption when using Cox regression, you should always formally test that the proportional hazards assumption is valid – but more on this later.

Below is a simplified illustration of what we mean by proportional hazards. We want to estimate the risk of dying among indoor versus outdoor cats diagnosed with feline leukemia virus over a five-year period. If the risk of dying (the hazard) among outdoor cats is 1.4 times higher than the risk of dying among indoor cats at the beginning of the observation period, the proportional hazards assumption implies that the risk of dying among outdoor cats remains 1.4 times higher at all later time points. In other words, the difference between the two hazards should remain 1.4 across the five-year period, no matter the shape of the distribution.



Tied failure times

As noted, tied failure times (ties) occur when two (or more) individuals have the same time to event, or they experience the event at the same time. For example, in the Olympic games, a tie for first place can be problematic: instead of one person each being ranked first, second, and third, suddenly two people are ranked first, no one is ranked second, and one person is ranked third. What does this have to do with Cox models? Cox regression is based on the partial likelihood function: the product of the conditional probabilities. Stay with us.

Imagine that we have a group of individuals who are at risk of the event (failure) at time t. For each event (failure time), we can calculate what is called the conditional probability of the event occurring (failure). To calculate the likelihood function, the numerator should contain only the individual who experiences the event at time t; the denominator contains all the other individuals in the group (risk set) for whom the event has not yet occurred. As such, calculating the likelihood function depends on the *order* of the failure times, not when the failures occur. Therefore, in theory, only one individual can fail at each failure time. If more than one individual fails at a single failure time, this is a tie, or a tied failure. Suddenly we have more than one individual in the numerator, the ordering of the failures is unclear, and we have no idea who won the gold medal. In conclusion, we want to minimize the number of tied failures.

Other names for Cox regression

Cox regression is sometimes called, e.g., proportional hazards regression.

17.2.1 Cox regression in short

If you have only one x, it is called simple regression, and if you have more than one x, it is called multiple regression. Regardless of whether you are doing a simple or a multiple regression, the x-variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from Cox regression		
Effect		
Hazard ratio (HR)	The exponent of hazard rate	
	Hazard rate	The probability that if the event has not yet occurred, it will occur in the next time interval, divided by the length of that interval.
Direction		
Negative	HR below 1	
Positive	HR above 1	
Statistical significance		
P-value	p<0.05 Statistically significant at the 5% level p<0.01 Statistically significant at the 1% level p<0.001 Statistically significant at the 0.1% level	
95% Confidence intervals	Interval does not include 1: Statistically significant at the 5% level Interval includes 1: Statistically non-significant at the 5% level	

Hazard ratio (HR)

In Cox regression analysis, the effect that x has on y is reflected by a hazard ratio (HR):

HR below 1	For every unit increase in x, the hazard rate of y decreases.
HR above 1	For every unit increase in x, the hazard rate of y increases.

Exactly how one interprets the HR in plain writing depends on the measurement scale of the x-variable. That is why we will present examples later for continuous, binary, and categorical (non-binary) x-variables.

Note Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for Cox regression is 1.

Note A HR can never be negative – it can range between 0 and infinity.

How to *not* interpret hazard ratios

The hazard ratios produced with Cox regression analysis are not the same as risk ratios (see Section 4.7.6). HRs tend to be inflated when they are above 1 and understated when they are below 1. This becomes more problematic the more common the outcome is (i.e. the more "cases" we have). However, the rarer the outcome is (<10% is usually considered a reasonable cut-off here), the closer hazard ratios and risks ratios become.

Many would find it compelling to interpret HRs in terms of percentages. For example, an HR of 1.20 might lead to the interpretation that the hazard rate of the outcome increases by 20%. If the HR is 0.80, some would then suggest that the hazard rate decreases by 20%. We would to urge you to carefully reflect upon the latter kind of interpretation since hazard ratios are not symmetrical: it can take any value above 1 but cannot be below 0. Thus, the choice of reference category might lead to quite misleading conclusions about effect size. The former kind of interpretation is usually considered reasonable when HRs are below 2. If they are above 2, it is better to refer to "times", i.e. an HR of 4.07 could be interpreted as "more than four times the hazard rate of…".

Take home messages
Do not interpret incidence hazard ratios as risk ratios, unless the outcome is very rare (<10%, but even then, be careful).
It is completely fine to discuss the results more generally in terms of higher or lower hazard rates/risks. However, if you want to give exact numbers to exemplify, you need to consider the asymmetry of hazard ratios as well as the size of the HR.

P-values and confidence intervals

In Cox regression analysis you can get information about statistical significance, in terms of both p-values and confidence intervals (also see Section 5.2).

Note The p-values and the confidence intervals will give you partly different information, but they are not contradictory. If the p-value is below 0.05, the 95% confidence interval will not include 1 and, if the p-value is above 0.05, the 95% confidence interval will include 1.

When you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5% level, the 1% level, or the 0.1% level).

When it comes to confidence intervals, Stata will by default choose 95% level confidence intervals. It is however possible to change the confidence level for the intervals. For example, you may instruct Stata to show 99% confidence intervals instead.

R-Squared

R-Squared (or R2) does not work very well due to the assumptions behind Cox regression. Stata produces a pseudo R2, but due to inherent bias this is seldom used.

Simple versus multiple regression models

The difference between simple and multiple regression models, is that in a multiple regression each x-variable's effect on y is estimated while accounting for the other x-variables' effects on y. We then say that these other x-variables are "held constant", or "adjusted for", or "controlled for". Because of this, multiple regression analysis is a way of dealing with the issue of confounding variables, and to some extent also mediating variables (see Section 9.3).

It is highly advisable to run a simple regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted coefficients with (i.e. what happened to the coefficients when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that they become weaker – which would of course be expected if the x-variables overlapped in their effect on y.

A note

Remember that a regression analysis should follow from theory as well as a comprehensive set of descriptive statistics and knowledge about the data. In the following sections, we will – for the sake of simplicity – not form any elaborate analytical strategy where we distinguish between x-variables and z-variables (see Section 9.3). However, we will define an analytical sample and use a so-called pop variable (see Section 11.5).

17.3 Declare that the data are time-to-event data

Before we can analyse time-to-event data (a.k.a. survival-time data), we need to declare this for Stata. This is a bit complicated, but take a deep breath!

Of course, we need one key variable, namely the variable that reflects the event. Moreover, we need to create some time variables. They should be in date format. It is also good practice to include an identification variable.

Key variables	
failure	Indicates whether the individual has experienced the event or not. For example: event is coded as 1 and no event as 0.
event	The date that the individual experiences the event.
origin	The date that defines when the time is zero. This is optional (but we think it makes sense to give the same date here as enter). However, if we want to attain age as the timescale, origin is specified as the date of birth.
enter	The date that the individual becomes at risk, i.e. enters the observational period. For example, if you have a follow-up period, enter is specified as the start date of that period.
exit	The date that the individual exits the study, i.e. the latest time that the individual is at risk. This is optional (the default is that the individual is removed after the event has occurred). However, it is useful if you want to specify the end date of the follow-up period.
id	Identification (id) number.

Note We do not actually need to name the variables failure, event, origin, enter, exit, or id. This is our choice.

The next step is to use stset to declare that the data is time-to-event data.

We will take our point-of-departure in the following command structure:

```
stset event, failure(failure==1) enter(time enter) exit(time exit) origin (time origin)
scale(365.25) id(id)
```

Note The scale option transforms the observational time from days (default) to years.

More information	help stset

Practical example

Dataset: StataData1.dta

Name	Label
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
cvd_year_str	Year of out-patient care due to CVD (Ages 41-50, Years 2011-2020)
cvd_month_str	Month of out-patient care due to CVD (Ages 41-50, Years 2011-2020)
cvd_day_str	Day of out-patient care due to CVD (Ages 41-50, Years 2011-2020)

Failure

In this example, we will focus on the variable cvd, which measures the occurrence of out-patient care due to cardiovascular disease (CVD). It looks like this:

```
tab cvd
```

```
Out-patient |
care due to |
  CVD (Ages |
     41-50, |
      Years |
 2011-2020) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |      9,482       94.82       94.82
        Yes |        518        5.18      100.00
------------+-----------------------------------
      Total |     10,000      100.00
```

Event

Connected to this variable, we have a variable (cvd_date_str) reflecting year, month, and day of the individual's first out-patient care event due to CVD. This variable is currently a string variable which we need to transform into a time variable through a series of steps, earlier described in this guide. We will just quickly repeat these steps here (see Sections 2.4.6-2.4.7 for more details).

Note If your dataset already contains this time variable (i.e. if you are using a saved dataset where you already performed the practical exercises in Sections 2.4.6-2.4.7), you should *not* perform these commands again.

```
gen cvd_year_str= substr(cvd_date_str,1,4)
```

```
gen cvd_month_str= substr(cvd_date_str,5,2)
```

```
gen cvd_day_str= substr(cvd_date_str,7,2)
```

```
gen cvd_year=real(cvd_year_str)
```

```
gen cvd_month=real(cvd_month_str)
```

```
gen cvd_day=real(cvd_day_str)
```

```
gen cvd_date=mdy(cvd_month,cvd_day,cvd_year)
```

```
format %d cvd_date
```

This is what the cvd_date variable looks like in the 100 first individuals (sorted by id).

```
tab cvd_date in 1/100
```

```
   cvd_date |      Freq.     Percent        Cum.
------------+-----------------------------------
  03dec2011 |          1        7.14        7.14
  17jul2012 |          1        7.14       14.29
  05dec2012 |          1        7.14       21.43
  27mar2013 |          1        7.14       28.57
  02apr2013 |          1        7.14       35.71
  27may2013 |          1        7.14       42.86
  17dec2013 |          1        7.14       50.00
  26jan2014 |          1        7.14       57.14
  17apr2016 |          1        7.14       64.29
  15sep2016 |          1        7.14       71.43
  13sep2017 |          1        7.14       78.57
  24jan2018 |          1        7.14       85.71
  08sep2018 |          1        7.14       92.86
  11mar2019 |          1        7.14      100.00
------------+-----------------------------------
      Total |         14      100.00
```

But we need to do one more step (should be performed even if you carried out the previous steps in Sections 2.4.6-2.4.7): impose the censoring date for the individuals that do not have an event. As will be explained later, we will censor the individuals at the end of follow-up (December 31, 2020). We will create a new variable for this purpose. This is actually the one that we will use in the analysis.

```
gen cvd_faildate=cvd_date
```

```
replace cvd_faildate=mdy(12,31,2020) if cvd_faildate==.
```

```
format %d cvd_faildate
```

Origin

In this dataset, the individuals are born in 1970. We do not have any detailed information about birth date, so we will specify the date as being in the middle of the year (June 30, 1970). After this, we format cvd_origin to be a date variable.

```
gen cvd_origin=mdy(6,30,1970)
```

```
format %d cvd_origin
```

Enter

The follow-up of out-patient care due to CVD starts on January 1, 2011:

```
gen cvd_enter=mdy(1,1,2011)
```

```
format %d cvd_enter
```

Exit

The follow-up period ends on December 31, 2020. This will be the exit date for the individuals that do not experience the event.

```
gen cvd_exit=mdy(12,31,2020)
```

```
format %d cvd_exit
```

For the individuals that do experience the event, the exit date will be replace with the failure date (i.e. the date that they experience the event):

```
replace cvd_exit=cvd_faildate if cvd==1
```

Note In the current example, we will keep it simple and just assume that all individuals stayed alive and did not drop out during the follow-up period.

Stset

Now, we have what we need to stset the data:

```
stset cvd_faildate, failure(cvd==1) enter(time cvd_enter) exit(time cvd_exit) origin
(time cvd_origin) scale(365.25) id(id)
```

There are four variables that are created when we use stset.

Variables created by stset	
_t0	Analysis time when observational period starts
_t	Analysis time when observational period ends
_d	Indicator of event (=1 if event has occurred)
_st	Indicator of whether the individual is included in the stset

Let us have a look at the variables we used for stset, including the new ones created. We will display this only for the 10 first individuals in the dataset.

```
list cvd_faildate cvd_origin cvd_enter cvd_exit _st _d _t _t0 in 1/10
```

```
     +---------------------------------------------------------------------------------------+
     | cvd_fai~e   cvd_ori~n   cvd_enter    cvd_exit   _st   _d          _t         _t0 |
     |---------------------------------------------------------------------------------------|
  1. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
  2. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
  3. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
  4. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
  5. | 11mar2019   30jun1970   01jan2011   11mar2019     1    1   48.695414   40.506502 |
     |---------------------------------------------------------------------------------------|
  6. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
  7. | 17jul2012   30jun1970   01jan2011   17jul2012     1    1   42.047912   40.506502 |
  8. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
  9. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
 10. | 31dec2020   30jun1970   01jan2011   31dec2020     1    0   50.505133   40.506502 |
     +---------------------------------------------------------------------------------------+
```

All individuals enter on the same date (cvd_enter=01jan2011) and they have the same origin date (cvd_origin=30jun1970). This means that the age at the beginning of the follow-up period is the same of everyone (_t0=40.506502).

We can see that the 5th and 7th individual has the value 1 for the variable _d. In other words, they have experienced the event (out-patient care due to CVD). The corresponding date is shown in cvd_failure (11mar2019 and 17jul2012, respectively), and the corresponding age at the event is shown in _t (48.695414 and 42.047912, respectively). For the remaining individuals, cvd_failure is set to 31dec2020 and _t is estimated to 50.505133.

This is the output we got when we executed the stset command:

```
                id:  id
     failure event:  cvd == 1
obs. time interval:  (cvd_faildate[_n-1], cvd_faildate]
 enter on or after:  time cvd_enter
 exit on or before:  time cvd_exit
    t for analysis:  (time-origin)/365.25
            origin:  time cvd_origin

------------------------------------------------------------------------------
     10,000  total observations
          0  exclusions
------------------------------------------------------------------------------
     10,000  observations remaining, representing
     10,000  subjects
        518  failures in single-failure-per-subject data
 97,394.305  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =   40.5065
                                          last observed exit t =  50.50513
```

Here, we can see that we have 10,000 individuals, of which 518 have experienced the event (out-patient care due to CVD). We also have a total of 97,394 years at risk and under observation. The “earliest observed entry t” is 40.5, reflecting age at enter. The “last observed exit t” is 50.5, which reflects age at exit.

Want to unset the data?

To remove the st markers from the dataset, just type:

```
stset, clear
```

Want to do a different stset?

It is not uncommon that we apply a number of Cox regressions with different outcomes, using the same dataset. In that case, you should create a set of time variables for each outcome (some variables can often be reused, e.g. origin and enter). Just make sure that you have the right stset active before you carry out the analysis. To check the current status, you can write:

```
st
```

Note You do not have to unset the data before doing another stset.

17.4 Descriptive analysis

Before we move on to Cox regression analysis, let us explore the time-to-event data properly first.

We can start with a simple description of how the data are arranged:

```
stdescribe, noshow
```

```
                                   |-------------- per subject --------------|
Category                   total        mean         min     median        max
------------------------------------------------------------------------------
no. of subjects            10000
no. of records             10000           1           1          1          1

(first) entry time                   40.5065     40.5065    40.5065    40.5065
(final) exit time                   50.24593    40.53388   50.50513   50.50513

subjects with gap              0
time on gap if gap             0           .           .          .          .
time at risk           97394.305    9.739431    .0273785   9.998631   9.998631

failures                     518       .0518           0          0          1
------------------------------------------------------------------------------
```

This shows, among other things, that we have 10,000 individuals in the analytical sample, of which 518 (5.18%) have experienced the outcome (cvd).

We also get some descriptive statistics for entry time, exit time, and time at risk. Since we have specified cvd_origin as date of birth, the values for mean/min/median/max entry time and exit time reflect age.

In the output above, mean age at entry is 40.51. This is actually the same for all individuals since they have the same date for cvd_origin and the same date for cvd_enter, which also explains why the same value is presented for min, median, and max.

The mean age at exit is 50.25 (min: 40.53, median: 50.50, max: 50.50). The reason why the same value is given for median and max is because a great majority of the individuals in the sample have not experienced the event are thus are censored at the end of follow-up (which equals age 50.50).

Time at risk is here presented as years. We can see that the mean is 9.74 (min=0.03, median=10.00, max=10.00). Again, the median and max values are the same since most individuals are censored at the end of follow-up (i.e. after 10 years).

More information	help stdescribe

Restricted mean survival time

Another way of obtaining the mean time at risk is through stci, which produces the restricted mean survival time (same as mean time at risk) along with information on standard errors and confidence intervals.

Estimating the restricted mean (or average) survival time is determined by calculating the area under the survival curve, restricting the estimation to the longest follow-up time. Below, we also include the noshow option.

```
stci, rmean noshow
```

```
             |    no. of   restricted
             |  subjects         mean      Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |     10000     9.739431(*)    .012822        9.7143     9.76456

(*) largest observed analysis time is censored, mean is underestimated
```

The restricted mean survival time in this example is 9.74 (years). Because of the way that our model is specified, the restricted mean survival time is the same as the mean time at risk.

However, the estimate has been flagged by Stata since the observation with the longest follow-up time is censored, which leads to the survivor function not reaching zero. As a consequence, the mean is underestimated.

Extended mean survival time

An alternative to the restricted mean survival time is to look at the extended mean survival time instead. This extends the survivor function from the last observed time to zero by using an exponential function.

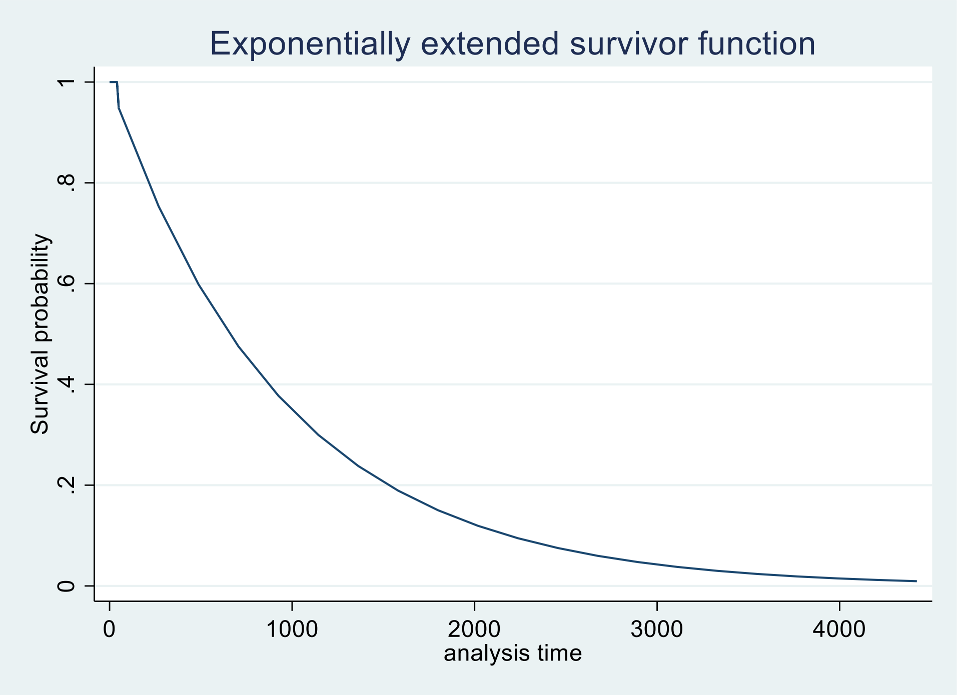
```
stci, emean noshow
```

```
             |     no. of     extended
             |   subjects         mean
-------------+----------------------
       total |      10000     910.0801
```

The extended mean survival time is 910 (years), which of course is a completely absurd estimate. This shows that the extended mean survival time should be used very cautiously.

We can produce a graph to have a closer look at the issue:

```
stci, emean graph
```



The curve shows the survival probability (i.e. the probability of not experiencing out-patient care due to CVD) across analysis time (years). The area under the curve is the proportion of individuals not experiencing the event. In sum, it takes more than 4,500 years for the survival function to reach 0 – although life expectancy is indeed increasing globally, we can probably conclude that estimating the extended mean survival time is not a reasonable alternative in the context of this example. This is perhaps not completely surprising since we do not expect that the entire sample at some point will be experiencing out-patient care due to CVD.

More information	help stci

Summary statistics

It is also possible to produce some summary statistics:

```
stsum, noshow
```

```
         |               incidence       no. of    |------ Survival time -----|
         | time at risk     rate        subjects        25%       50%       75%
---------+---------------------------------------------------------------------
   total |  97394.30527   .0053186         10000          .         .         .
```

This shows time at risk, the incidence rate, and number of subjects, as well as survival time at the 25th, 50th, and 75th percentile.

Note Survival time at the 50th percentile is the same as median survival time.

More information	help stsum

Median survival time

Another way of obtaining the median survival time through the following command:

```
stci, median noshow
```

In the current example, we do not get any values for survival time at the 25th, 50th, or 75th percentile since the percentage of failure (i.e. proportion of cases with cvd) is very low (less than 5%). It would nevertheless be possible to estimate survival time at the 1st to 4th percentile. Let us try out the first and last of these:

```
stci, p(1) noshow
```

```
stci, p(4) noshow
```

```
             |    no. of
             |  subjects         1%      Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------------------------
       total |     10000    42.63929             .        42.2122     42.8501


             |    no. of
             |  subjects         4%      Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------------------------
       total |     10000    48.18891             .         47.551     48.7064
```

Since we specified origin as date of birth (well, not the exact date) when we applied stset to the data, we actually get the median survival age for the two percentiles. For the 1st percentile, median survival age is 42.64 (95% CI: 42.21-42.85), whereas it is 48.19 (95% CI: 47.55-48.71) for the 4th percentile.

More information	help stci

17.4.1 Kaplan-Meier curves

A Kaplan-Meier curve is a descriptive (non-parametric) method that visualizes the survival function (or hazard function). The visual representation is based on the Kaplan-Meier estimator (also called product-limit estimator). On the y-axis, we get conditional probabilities, whereas time bands (time intervals) are displayed on the x-axis. The time bands can be based on any time unit, such as, hours, days, months, or years.

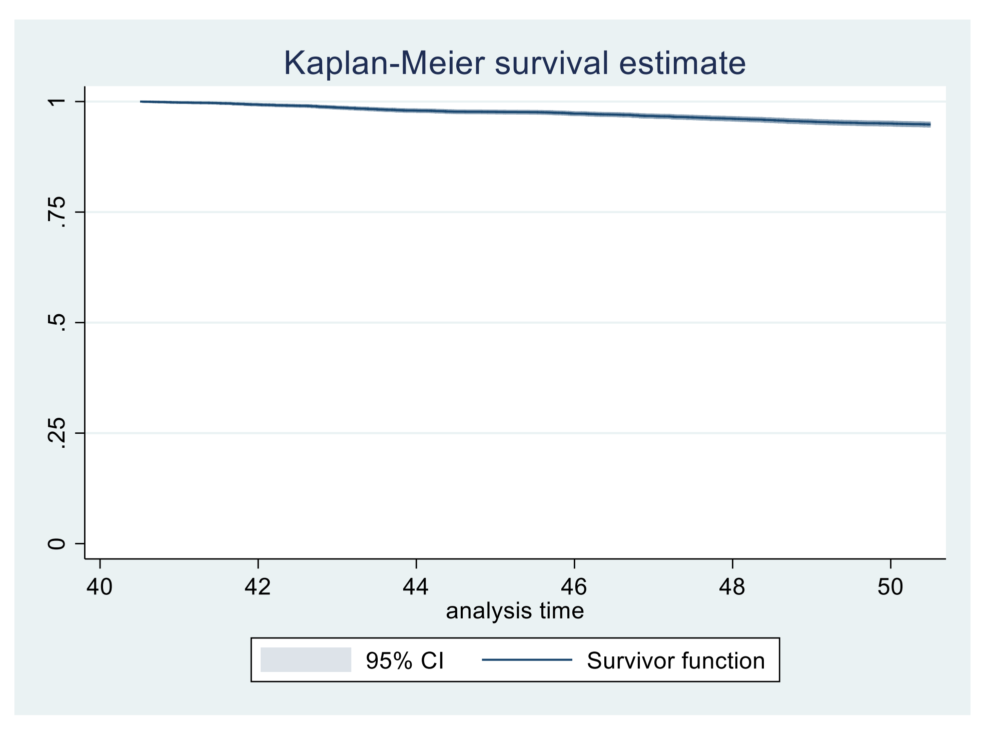
Note Kaplan-Meier estimation is very similar to life-table methods. Primary differences lie in the methods for choosing time bands and handling ties. Similar to Cox regression, Kaplan-Meier estimation assumes that ties are rare, since time-to-event is measured on a continuous scale.

Kaplan-Meier survivor function

We can generate a graph showing the Kaplan-Meier survivor function. This shows the probability of survival at each time band, calculated as the number of individuals surviving divided by the number of individuals at risk.

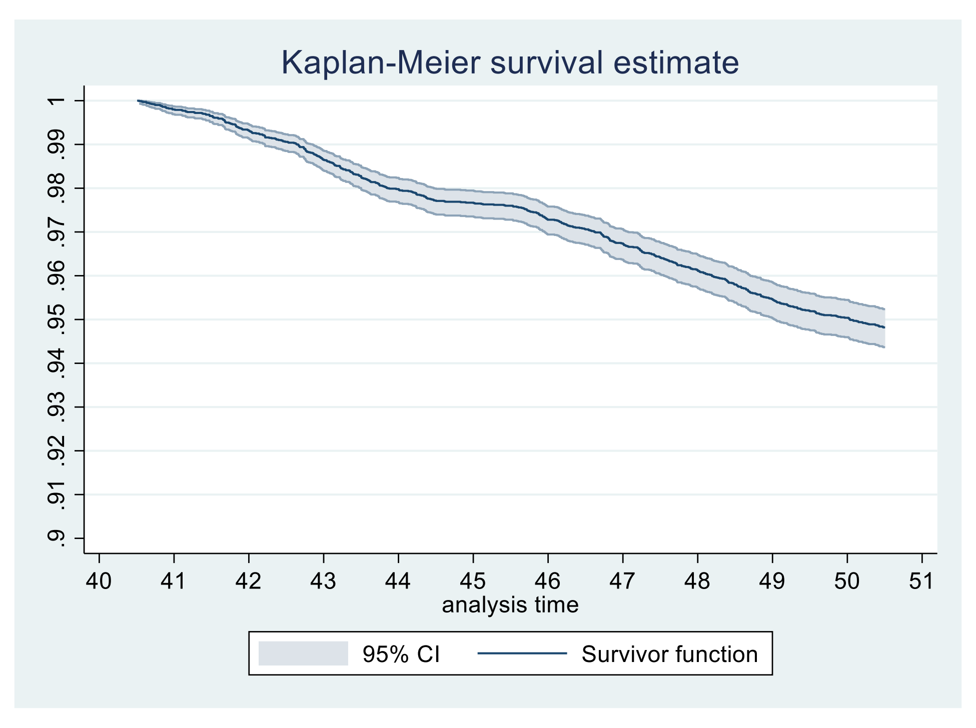
In the graph, we will include confidence intervals (ci option) and use the noorigin option to exclude the time before follow-up from the graph.

```
sts graph, survival ci noorigin
```



The scale is a bit too wide – we do not really see the line properly. This can be fixed by specifying the y-axis (with the ylabel option). While we are at it, we can adjust the x-axis (with the xlabel option) as well.

```
sts graph, survival ci noorigin ylabel(.90(0.01)1) xlabel(40(1)51)
```



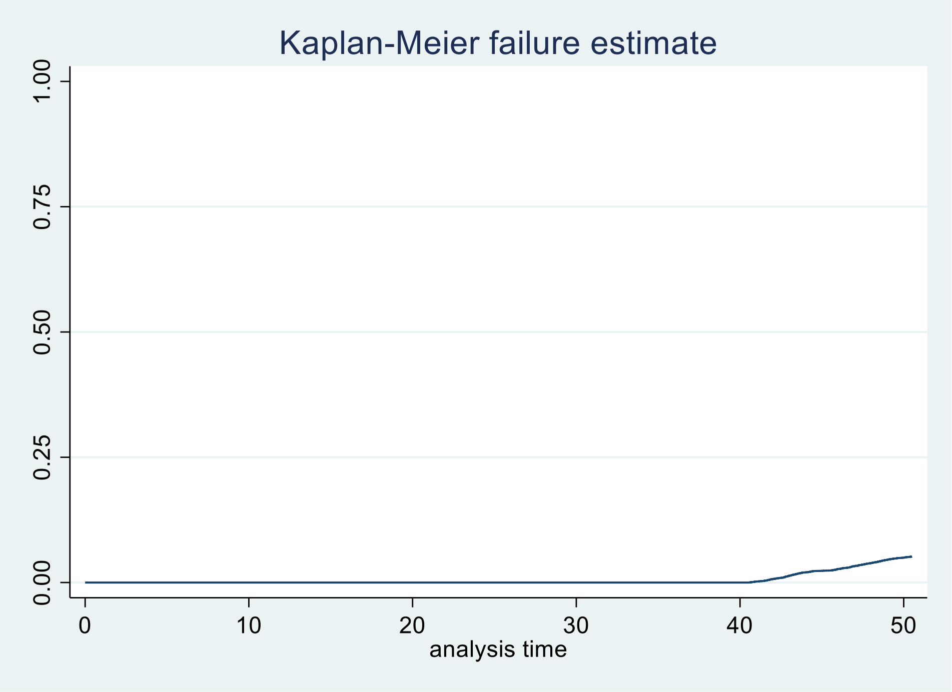
Above, we have specified that the y-axis should range from 0.90 to 1, with one tick per 0.01 unit. And the x-axis ranges from 40 to 51, with one tick per 1 unit.

More information	help sts graph

Kaplan-Meier failure function

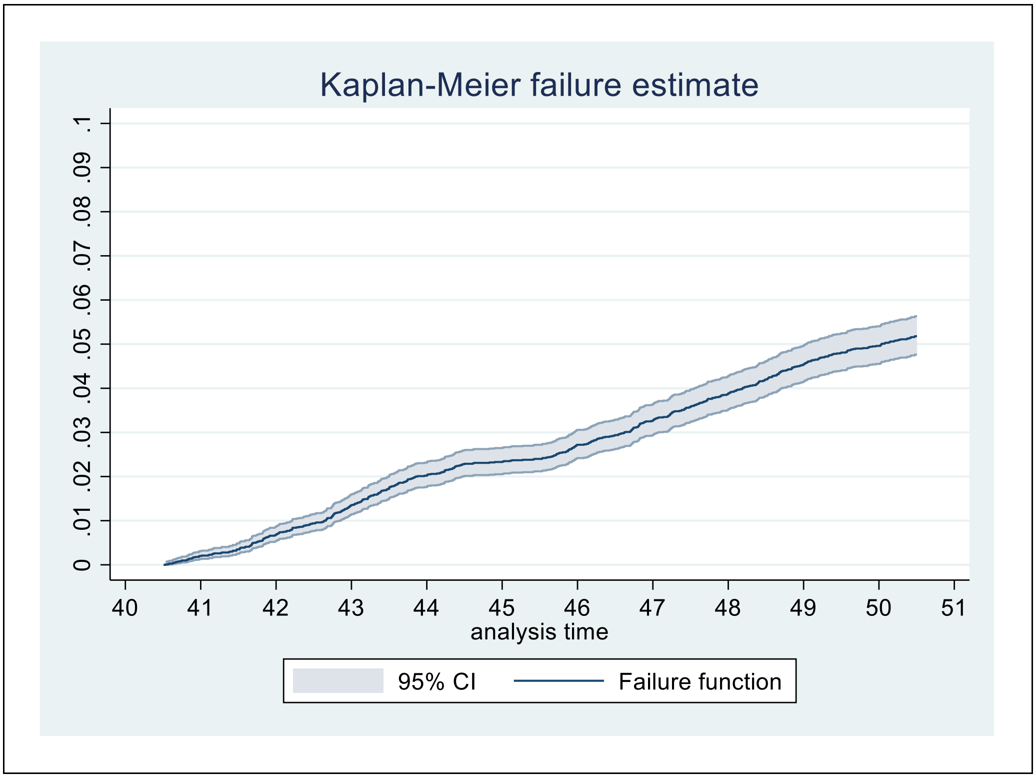
We could also produce the opposite, namely a graph of the failure function:

```
sts graph, failure
```



Again, we can change the specification to adapt the scale:

```
sts graph, failure ci noorigin ylabel(.10(0.01)0) xlabel(40(1)51)
```

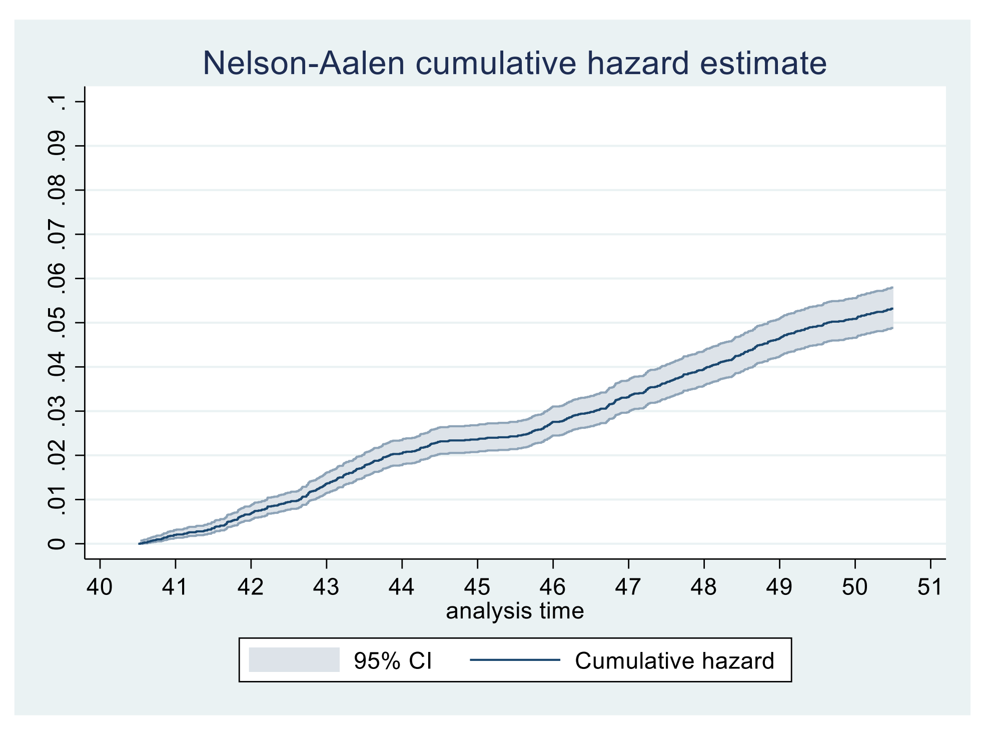


More information	help sts graph

17.4.2 Nelson-Aalen cumulative hazard function

An alternative to the Kaplan-Meier curves is to graph the Nelson-Aalen estimate of cumulative hazard function. We add the ci option and the noorigin option, as well as change the specification of the scales:

```
sts graph, cumhaz ci noorigin ylabel(.10(0.01)0) xlabel(40(1)51)
```



More information	help sts graph

17.5 Function

Note It is crucial to remember that before stcox can be performed, we need to stset our data. See Section 17.3 for more information and to stset the dataset so it is ready for the upcoming example.

Basic command	stcox indepvars	
Useful options	stcox indepvars, noshow	
Explanations	indepvars	Insert the name of the x-variable(s) that you want to use.
	irr	Produces incidence rate ratios.
More information	help stcox	

Note The dependent variable (y-variable) is already specified through stset.

A walk-through of the output

When we perform a Cox regression in Stata, the table looks like this:

```
Cox regression -- Breslow method for ties

No. of subjects =        8,126                  Number of obs    =       8,126
No. of failures =          147
Time at risk    =   55903.7399
                                                LR chi2(2)       =       46.84
Log likelihood  =   -1292.4291                  Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xvar1 |   .4281216   .0766169    -4.74   0.000     .3014641    .6079932
       xvar2 |   .9450562   .0112373    -4.75   0.000     .9232861    .9673396
------------------------------------------------------------------------------
```

In this example, xvar1 is a binary (0/1) variable and xvar2 is a continuous variable ranging between 1 and 40.

The upper part of the table shows a model summary. This is what the different rows mean:

Row	**Explanation**
No. of subjects	Number of subjects included in the model.
No. of failures	Number of cases (i.e. subjects that have experienced the event).
Time at risk	Total observational time, according to the specified unit (e.g. year).
Log likelihood	This value does not mean anything in itself, but can be used if we would like compare nested models.
Number of obs	The number of observations included in the model.
LR chi2(x)	The likelihood ratio (LR) chi-square test. The number within the brackets shows the degrees of freedom (one per variable).
Prob >chi2	Shows the probability of obtaining the chi-square statistic given that there is no statistical effect of the x-variables on y. If the p-value is below 0.05, we can conclude that the overall model is statistically significant.

The lower part of the table presents the parameter estimates from the analysis.

Column	**Explanation**
	The first column lists the predicted value of the y-variable on top (_t), followed by our x-variable(s).
Haz. Ratio	These are the hazard ratios.
Std. Err.	The standard errors associated with the coefficient.
Z	Z-value (coefficient divided by the standard error of the coefficient).
P>|z|	P-value.
[95% Conf. Interval]	95% confidence intervals (lower limit and upper limit).

The analytical sample used for the examples

In the subsequent sections, we will use the following variables:

Dataset: StataData1.dta

Name	**Label**
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
gpa	Grade point average (Age 15, Year 1985)
sex	Sex
marstat40	Marital status (Age 40, Year 2010)

```
sum cvd gpa sex marstat40
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         cvd |     10,000       .0518    .2216341          0          1
         gpa |      9,380    3.178614    .6996298          1          5
         sex |     10,000       .4892    .4999083          0          1
   marstat40 |      8,950     1.69933    .8147083          1          4
```

We define our analytical sample through the following command:

```
gen pop_cox=1 if cvd!=. & gpa!=. & sex!=. & marstat40!=.
```

This means that new the variable pop_cox gets the value 1 if the four variables do not have missing information. In this case, we have 8,464 individuals that are included in our analytical sample.

```
tab pop_cox
```

```
    pop_cox |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      8,464      100.00      100.00
------------+-----------------------------------
      Total |      8,464      100.00
```

17.6 Simple Cox regression

Quick facts	
Number of variables	One dependent (y) One independent (x)
Scale of variable(s)	Dependent: time-to-event Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

17.6.1 Simple Cox regression with a continuous x

Theoretical examples

Example 1

We want to estimate the effect of age (x) on all-cause mortality (y) among a group of individuals ages 65 and older within a ten-year follow-up period. The failure event is death (0=No event, 1=Event). Age is measured in years, with values ranging from 65 to 100. The HR for age in years is 1.13, which suggests that the expected hazard is 1.13 times higher for an individual who is one year older than another individual.

Example 2

In this example, we estimate the association between weight (x) and hospitalization attributable to cardiovascular disease (y) on a population of 50-year-old women, who are followed for five years. The failure event is hospitalization for cardiovascular disease (0=No event, 1=Event). Weight at age fifty is measured in kilograms, ranging from 48 to 114. We find that the HR is 1.03. This suggests that a one-kilogram increase in weight is associated with a 3% increase in the expected hazard for hospitalization.

Practical example

Dataset: StataData1.dta

Name	**Label**
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
gpa	Grade point average (Age 15, Year 1985)

```
sum cvd gpa if pop_cox==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         cvd |      8,464     .048913     .215699          0          1
         gpa |      8,464    3.184664    .6935797          1          5
```

```
stcox gpa if pop_cox==1, noshow
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,464                  Number of obs    =      8,464
No. of failures =          414
Time at risk    =  82729.84805
                                                LR chi2(1)       =     101.33
Log likelihood  =   -3683.1099                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   .4834111    .035387    -9.93   0.000     .4187998    .5579905
------------------------------------------------------------------------------
```

When we look at the results for gpa, we see that the hazard ratio (HR) is 0.48. Thus, for each unit increase in grade point average, the hazard of out-patient care due to CVD decreases.

The association is statistically significant, as reflected in the p-value (0.000) and the 95% confidence intervals (0.42-0.56).

Summary

The higher the grade point average at age 15, the lower the risk of having experienced out-patient care due to CVD in ages 41-50 (HR=0.48). The association is statistically significant (95% CI=0.42-0.56).

17.6.2 Simple Cox regression with a binary x

Theoretical examples

Example 1

Suppose we are interested in the estimated effect of smoking (x) on mortality due to lung cancer (y) within a ten-year follow-up period. The failure event is death due to lung cancer (0=No event, 1=Event). Smoking status is coded as 0=Never smoked and 1=Ever smoked. The HR for ever-smokers is 2.73. This result suggests that, compared to never-smokers (the reference category), ever-smokers have a 2.73 times higher risk of dying.

Example 2

We want to estimate the effect of sex (x) on all-cause mortality risk (y) among important characters during the first seven seasons of HBO's popular television series, *Game of Thrones*. The failure event is death (0=No event, 1=Event). Sex is coded as 0=Male and 1=Female. The HR for females is 0.80, which indicates that, compared to males, important female characters have a 20% lower risk of dying during the first seven seasons.

Note This example is based on a published article. The curious reader can read more here: Lystad, R. P., & Brown, B. T. (2018). "Death is certain, the time is not": mortality and survival in Game of Thrones. *Injury Epidemiology*, 5:44.

Practical example

Dataset: StataData1.dta

Name	**Label**
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
sex	Sex

```
sum cvd sex if pop_cox==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         cvd |      8,464     .048913     .215699          0          1
         sex |      8,464    .4956285    .5000104          0          1
```

```
stcox sex if pop_cox==1, noshow
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,464                  Number of obs    =      8,464
No. of failures =          414
Time at risk    =  82729.84805
                                                LR chi2(1)       =      41.13
Log likelihood  =   -3713.2076                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5236135   .0542108    -6.25   0.000     .4274487    .6414127
------------------------------------------------------------------------------
```

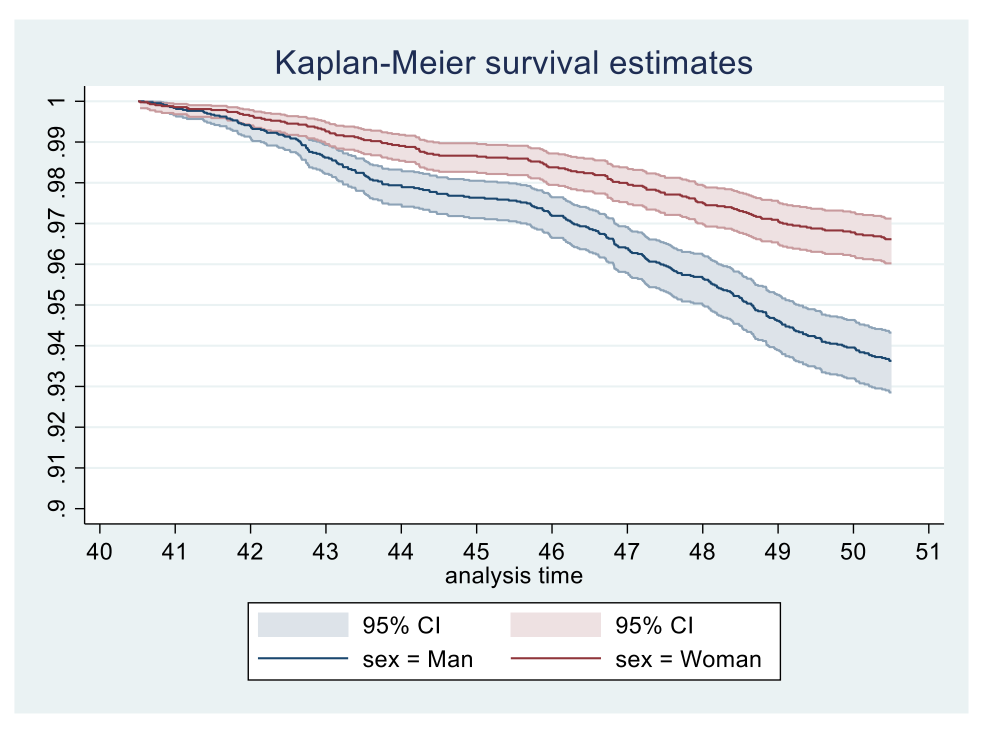
When we look at the results for sex, we see that the hazard ratio (HR) is 0.52. Thus, one unit increase in sex is associated with a lower hazard of out-patient care due to CVD. This means that women have a lower risk of experiencing the outcome compared to men.

The association is statistically significant, as reflected in the p-value (0.000) and the 95% confidence intervals (0.43-0.64).

Kaplan-Meier curves

It is possible to illustrate survival curves (or failure curves) separately for men and women, by means of the Kaplan-Meier estimator.

```
sts graph if pop_cox==1, survival ci noorigin ylabel(.90(0.01)1) xlabel(40(1)51) by(sex)
```



More information	help sts graph

Although we already know this from the Cox regression analysis, it is possible to use a log-rank test to assess whether the differences between the curves of men and women are statistically significant. One can think of this as the censored data equivalent to a non-parametric ANOVA. This tests the hypothesis that there is no difference between the groups. If the test is statistically significant (p<0.05), we reject this hypothesis.

```
sts test sex if pop_cox==1, noshow
```

```
Log-rank test for equality of survivor functions

      |   Events         Events
sex   |  observed       expected
------+-------------------------
Man   |       272         207.31
Woman |       142         206.69
------+-------------------------
Total |       414         414.00

            chi2(1) =      40.44
            Pr>chi2 =     0.0000
```

In this case, the p-value (Pr>chi2) is below 0.05 (0.000), suggesting that there is a statistically significant difference between men and women in the probability of out-patient care due to CVD.

More information	help sts test

Summary
Women have a lower risk of out-patient care due to CVD in ages 41-50, compared to men (HR=0.52; 95% CI=0.43-0.64).

17.6.3 Simple Cox regression with a categorical (non-binary) x

Theoretical examples

Example 1

In this example, we estimate the association between age (x) and hospitalization for attempted suicide (y) among individuals ages 18 to 45 over a five-year period. The failure event is hospitalization for attempted suicide (0=No event, 1=Event). Age is coded into three categories: 1=Ages 18-25, 2=Ages 26-35, and 3=Ages 36-45. Ages 36-45 is selected as the reference category (HR=1.00). For Ages 18-25, HR=3.77; for Ages 26-35, HR=1.08. This means that, compared to individuals ages 36-45, individuals ages 18-25 have a 3.77 times higher risk for hospitalization for attempted suicide, whereas individuals ages 26-35 have a 1.08 times higher risk for hospitalization compared to the reference category.

Example 2

We are interested in the relationship between marital status (x) and death attributable to Covid-19 over a six-month period. The failure event is death due to Covid-19 (y) (0=No event, 1=Event). Marital status is coded in four categories: 1=Married (reference category), 2=Divorced, 3=Widowed, and 4=Never married. The HRs for divorced, widowed, and never married individuals are 1.56, 1.98, and 2.47, respectively. Relative to those who are married, divorced, widowed, and never-married individuals have a higher hazard of mortality from Covid-19.

Note This example is largely based on the following publication: Drefahl, S., et al. (2020). Socio-demographic risk factors of COVID-19 deaths in Sweden: A nationwide register study. *Stockholm Research Reports in Demography*.

Practical example

Dataset: StataData1.dta

Name	**Label**
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
marstat40	Marital status (Age 40, Year 2010)

```
sum cvd marstat40 if pop_cox==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         cvd |      8,464     .048913     .215699          0          1
   marstat40 |      8,464    1.694353    .8147792          1          4
```

The variable marstat40 has four categories: 1=Married, 2=Unmarried, 3=Divorced, and 4=Widowed. Here, we (with ib1) specify that the first category (Married) will be the reference category.

```
stcox ib1.marstat40 if pop_cox==1, noshow
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,464                  Number of obs    =      8,464
No. of failures =          414
Time at risk    =  82729.84805
                                                LR chi2(3)       =     110.55
Log likelihood  =      -3678.5                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   marstat40 |
  Unmarried  |   2.864836    .348911     8.64   0.000     2.256478     3.63721
   Divorced  |   3.113382   .3994967     8.85   0.000     2.421083    4.003641
    Widowed  |   2.456653   1.122484     1.97   0.049     1.003266    6.015496
------------------------------------------------------------------------------
```

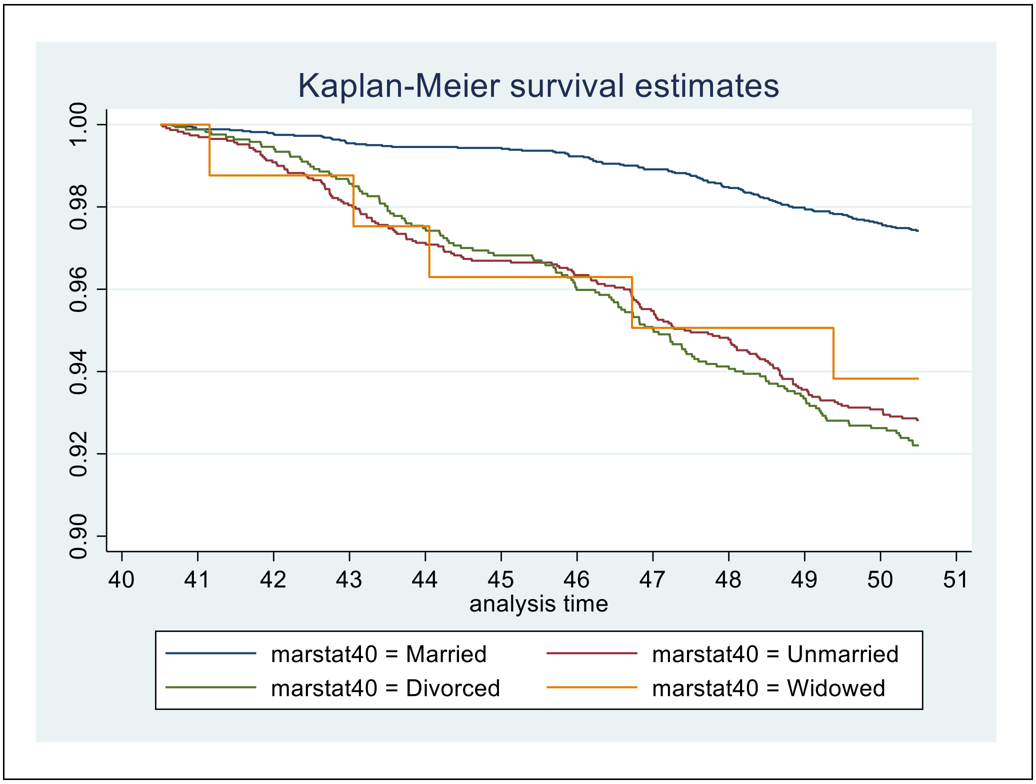
When we look at the results for the dummies for marstat40, we see that the hazard ratio is 2.87 for Unmarried, 3.11 for Divorced, and 2.46 for Widowed. Thus, all three groups have much higher hazards of out-patient care due to CVD compared to those who are married.

All three dummies for marstat40 are significantly different from the reference category, as reflected in the p-values and the 95% confidence intervals.

Kaplan-Meier curves

It is possible to illustrate survival curves (or failure curves) separately for the categories of marstat40, by means of the Kaplan-Meier estimator. Due to the few cases occurring in the category Widowed, however, the confidence intervals go bananas. Therefore, we will not include the ci option below.

```
sts graph if pop_cox==1, survival noorigin ylabel(.90(0.02)1) xlabel(40(1)51)
by(marstat40)
```



More information	help sts graph

While we already know this from the Cox regression analysis, it is possible to use a log-rank test to assess whether the differences between the curves of married, unmarried, divorced, and widowed individuals are statistically significant. This tests the hypothesis that there is no difference between the groups. If the test is statistically significant (p<0.05), we reject this hypothesis.

```
sts test marstat40 if pop_cox==1, noshow
```

```
Log-rank test for equality of survivor functions

          |   Events         Events
marstat40 |  observed       expected
----------+-------------------------
Married   |       114         219.12
Unmarried |       165         110.71
Divorced  |       130          80.26
Widowed   |         5           3.91
----------+-------------------------
Total     |       414         414.00

                chi2(3) =     108.18
                Pr>chi2 =     0.0000
```

In this case, the p-value (Pr>chi2) is below 0.05 (0.000), suggesting that there is a statistically significant difference between the categories of marital status in the probability of out-patient care due to CVD.

More information	help sts test

Note Since marstat40 is not an ordinal variable, there is no point of using marginsplot to plot any trend.

Summary
At age 40, being unmarried, divorced, or widowed is associated with significantly higher risks of out-patient care due to CVD, as compared to being married.

17.7 Multiple Cox regression

Quick facts	
Number of variables	One dependent (y) At least two independent (x)
Scale of variable(s)	Dependent: time-to-event Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

Theoretical example

Example

In a post-apocalyptic universe, we would like to study the effects of sex, marital status, and urbanicity on cause-specific mortality during a year-long, localized epidemic. In this example, the failure event is death due to an engineered zombie virus (0=No event, 1=Event). Sex is coded as 0=Male and 1=Female, whereas marital status is coded in four categories: 1=Married, 2=Divorced, 3=Widowed, and 4=Never married. Urbanicity is coded into three categories: 0=Rural, 1=Suburban, and 2=Urban. Our reference categories are male, married, and rural, respectively.

The HR for females is 0.92. Relative to males, and holding marital status and urbanicity constant, females have an estimated 8% lower mortality attributable to zombie virus.

Estimating the effects of marital status on survival, we get an HR of 1.34 for divorced individuals, an HR of 0.96 for widowed individuals, and an HR of 2.28 for never-married individuals. Holding the other covariates constant, divorced and never-married individuals have a higher estimated hazard for mortality compared to married individuals. Conversely, these results indicate that widowed individuals have an estimated 4% lower risk of mortality relative to those who are married.

For our measure of urbancity, the HR is 1.45 for suburban areas and 12.78 for urban areas. After controlling for sex and marital status, we estimate that individuals living in the suburbs have a 1.45 times higher risk of death relative to individuals living in rural areas. Individuals living in urban areas have a 12.78 times higher risk of mortality due to the zombie virus, relative to individuals in rural areas and holding the other covariates constant. Yikes.

Practical example

Dataset: StataData1.dta

Name	Label
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
gpa	Grade point average (Age 15, Year 1985)
sex	Sex
marstat40	Marital status (Age 40, Year 2010)

```
sum cvd gpa sex marstat40 if pop_cox==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         cvd |      8,464     .048913     .215699          0          1
         gpa |      8,464    3.184664    .6935797          1          5
         sex |      8,464    .4956285    .5000104          0          1
   marstat40 |      8,464    1.694353    .8147792          1          4
```

```
stcox gpa sex ib1.marstat40 if pop_cox==1, noshow
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,464                  Number of obs    =       8,464
No. of failures =          414
Time at risk    =  82729.84805
                                                LR chi2(5)       =      218.62
Log likelihood  =   -3624.4625                  Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   .5486937   .0408619    -8.06   0.000     .4741764    .6349214
         sex |   .5900992   .0623588    -4.99   0.000      .479705    .7258985
             |
   marstat40 |
  Unmarried  |   2.519405   .3085353     7.55   0.000     1.981786    3.202868
   Divorced  |    2.95507   .3807907     8.41   0.000     2.295524    3.804116
    Widowed  |   2.967418    1.35938     2.37   0.018     1.209042    7.283094
------------------------------------------------------------------------------
```

In this model, we have three x-variables: gpa, sex, and marstat40. When we put them together, their statistical effect on cvd is mutually adjusted.

When it comes to the hazard ratios, they have changed in comparison to the simple regression models. For example, the hazard ratios have become closer to 1 for both gpa and sex: they changed from 0.48 to 0.55 for gpa, and from 0.52 to 0.59 for sex.

This is largely also the case for the dummies of marstat40: the hazard ratio for Unmarried has changed from 2.86 to 2.52 and the one for Divorced has changed from 3.11 to 2.96. The hazard ratio for Widowed has nevertheless increased: from 2.46 to 2.97.

All x-variables still demonstrate statistically significant associations with cvd.

Note A specific hazard ratio from a simple Cox regression model can increase when other x-variables are included. Usually, it is just “noise”, i.e. not any large increases, and therefore not much to be concerned about. But it can also reflect that there is something going on that we need to explore further. There are many possible explanations for increases in multiple regression models: a) We actually adjust for a confounder and then “reveal” the “true” statistical effect. b) There are interactions among the x-variables in their effect on the y-variable. c) There is something called collider bias (which we will not address in this guide) which basically mean that both the x-variable and the y-variable causes another x-variable in the model. d) The simple regression models and the multiple regression model are based on different samples. e) It can be due to rescaling bias (see Chapter 18).

Summary
In the fully adjusted model, it can be observed that while most associations are slightly attenuated in strength, they remain largely the same as in the simple models.

Estimates table and coefficients plot

If we have multiple models, we can facilitate comparisons between the regression models by asking Stata to construct estimates tables and coefficients plots. What we do is to run the regression models one-by-one, save the estimates after each, and than use the commands estimates table and coefplot.

The coefplot option is not part of the standard Stata program, so unless you already have added this package, you need to install it:

```
ssc install coefplot
```

As an example, we can include the three simple regression models as well as the multiple regression model. The quietly option is included in the beginning of the regression commands to suppress the output.

Run and save the first simple regression model:

```
quietly stcox gpa if pop_cox==1, noshow
```

```
estimates store model1
```

Run and save the second simple regression model:

```
quietly stcox sex if pop_cox==1, noshow
```

```
estimates store model2
```

Run and save the third simple regression model:

```
quietly stcox ib1.marstat40 if pop_cox==1, noshow
```

```
estimates store model3
```

Run and save the multiple regression model:

```
quietly stcox gpa sex ib1.marstat40 if pop_cox==1, noshow
```

```
estimates store model4
```

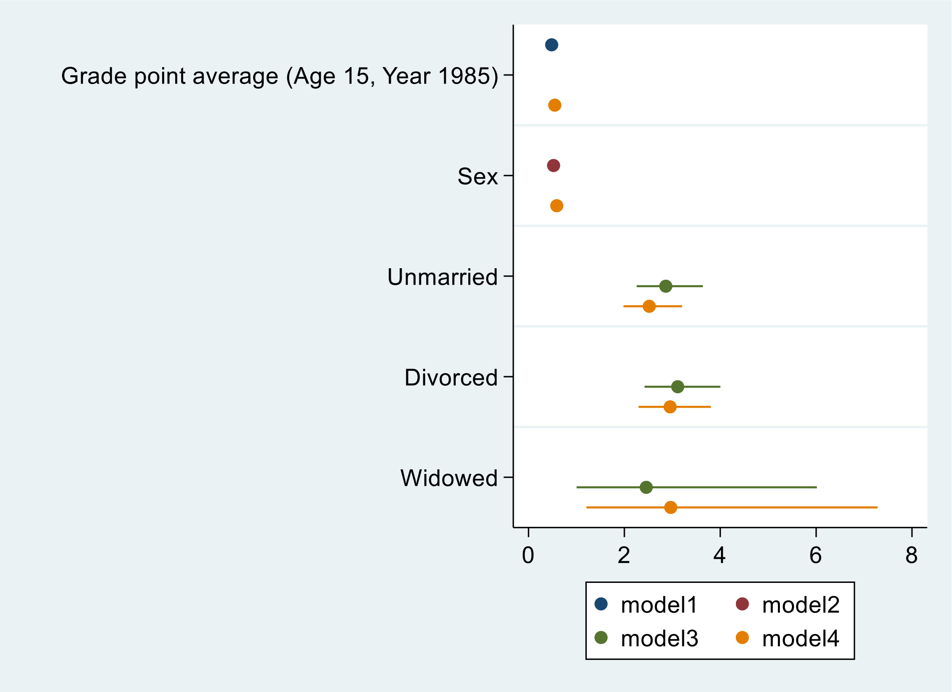
Produce the estimates table (include the option eform to show hazard ratios):

```
estimates table model1 model2 model3 model4, eform
```

```
------------------------------------------------------------------
    Variable |   model1       model2       model3       model4
-------------+----------------------------------------------------
         gpa |  .48341112                               .54869369
         sex |               .52361346                  .59009923
             |
   marstat40 |
  Unmarried  |                            2.8648356    2.5194046
   Divorced  |                            3.1133821      2.95507
    Widowed  |                            2.4566531    2.9674176
------------------------------------------------------------------
```

Produce the coefficients plot (include the option eform to show hazard ratios):

```
coefplot model1 model2 model3 model4, eform
```



Note You can improve the graph by using the Graph Editor to adjust the category and label names.

17.8 Model diagnostics

The assumptions behind Cox regression are similar to other types of generalized linear models. Nevertheless, there are some additional assumptions that need to be tested, such as the hazards being proportional and the failure times not being tied.

More information	help stcox postestimation

Checklist	
Time-to-event outcome	The y-variable has to reflect time-to-event.
Independence of errors	Data should be independent, i.e. not derived from any dependent samples design, e.g. before-after measurements/paired samples.
Correct model specification	Your model should be correctly specified. This means that the x-variables that are included should be meaningful and contribute to the model. No important (confounding) variables should be omitted (often referred to as omitted variable bias).
No multicollinearity	Multicollinearity may occur when two or more x-variables that are included simultaneously in the model are strongly correlated with each another. Actually, this does not violate the assumptions, but is does create greater standard errors which makes it harder to reject the null hypothesis.
Proportional hazards	The ratio of the hazards is constant over time.
Failure times not tied	The number of ties in your data is minimal.

Types of model diagnostics	
Link test	Assess model specification
Correlation matrix	Check for multicollinearity
Log-log plot of survival	Check proportional hazards assumption
Kaplan-Meier and predicted survival plot	Check proportional hazards assumption
Schoenfeld residuals	Check proportional hazards assumption
Tied failure times	Use one of four approaches

17.8.1 Link test

With the command linktest, we can assess whether our model is correctly specified. This test uses the linear predicted value (called _hat) and the linear predicted value squared (_hatsq) to rebuild the model. We expect _hat to be statistically significant, and _hatsq to be statistically non-significant. If one or both of these expectations are not met, the model is mis-specified.

However, do not rely too much on this test – remember that you should also use theory and common sense to guide your decisions. It is very seldom relevant to focus on this test if our ambition is to investigate associations (and not to make the best possible prediction of the outcome).

More information	help linktest

Practical example

We perform this test for the full model, so let us go back to the example from the multiple regression analysis. The quietly option is included in the beginning of the command to suppress the output.

```
quietly stcox gpa sex ib1.marstat40 if pop_cox==1, noshow
```

And then we run the test:

```
linktest
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,464                  Number of obs    =      8,464
No. of failures =          414
Time at risk    =  82729.84805
                                                LR chi2(2)       =     219.26
Log likelihood  =   -3624.1461                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   1.172454   .2256356     5.20   0.000      .730216    1.614692
      _hatsq |   .0648255   .0805195     0.81   0.421    -.0929899    .2226408
------------------------------------------------------------------------------
```

Since the p-value for the variable _hat is below 0.05 and the p-value for _hatsq is above 0.05, it means that our model correctly specified.

17.8.2 Correlation matrix

As the x-variables become more strongly correlated, it becomes more difficult to determine which of the variables are actually producing the statistical effect on the y-variable. This is the problem with multicollinearity.

One way of assessing multicollinearity is using the estat vce command, with the corr (short for correlation) option.

More information	help estat vce

Practical example

The first step is re-run the multiple Cox regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly stcox gpa sex ib1.marstat40 if pop_cox==1, noshow
```

Next, we try the estat vce command. By adding the corr (=correlation) option, we will get a correlation matrix instead of a covariance matrix.

```
estat vce, corr
```

```
Correlation matrix of coefficients of cox model

             |                            2.        3.        4.
        e(V) |      gpa       sex  marst~40  marst~40  marst~40
-------------+--------------------------------------------------
         gpa |   1.0000
         sex |  -0.1506    1.0000
 2.marstat40 |   0.0855    0.0466    1.0000
 3.marstat40 |   0.0807   -0.0550    0.5602    1.0000
 4.marstat40 |   0.0054   -0.0717    0.1525    0.1521    1.0000
```

The table shows the correlations between the different variables/categories. In line with the earlier sections on correlation analysis (see Chapter 7.2), we can conclude that the coefficients suggest (very) weak correlations here. The only exceptions are two of the dummies for marstat40, which is not a huge problem since they reflect the same underlying variable.

17.8.3 Log-log plot of survival

The log-log plot of survival plots the curve for each category of any categorical x-variable versus ln(analysis time). If the curves are parallel, the proportional hazards assumption is not violated.

Note It is possible to adjust for additional x-variables. However, for the sake of simplicity, we will just run simple regression models for our categorical variables.

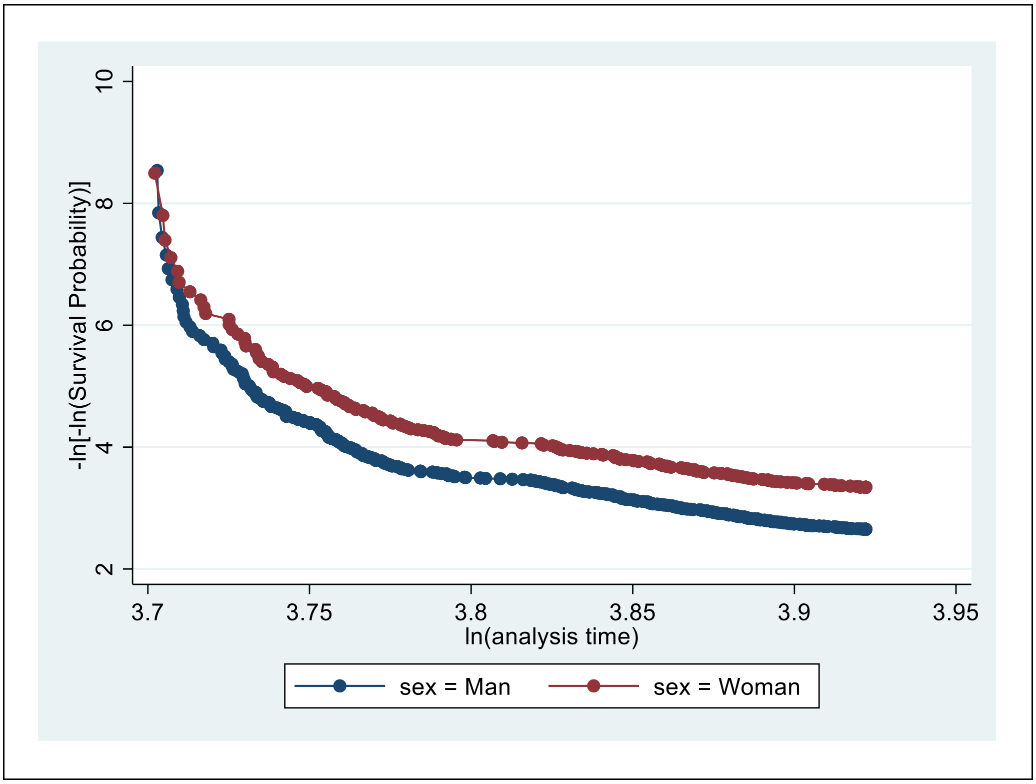
More information	help stphplot

Practical example

The first step is re-run a Cox regression model. We will start with the simple one that we did for sex. The quietly option is included in the beginning of the command to suppress the output. After this, we use the stphplot command.

```
quietly stcox sex if pop_cox==1, noshow
```

```
stphplot, by(sex)
```



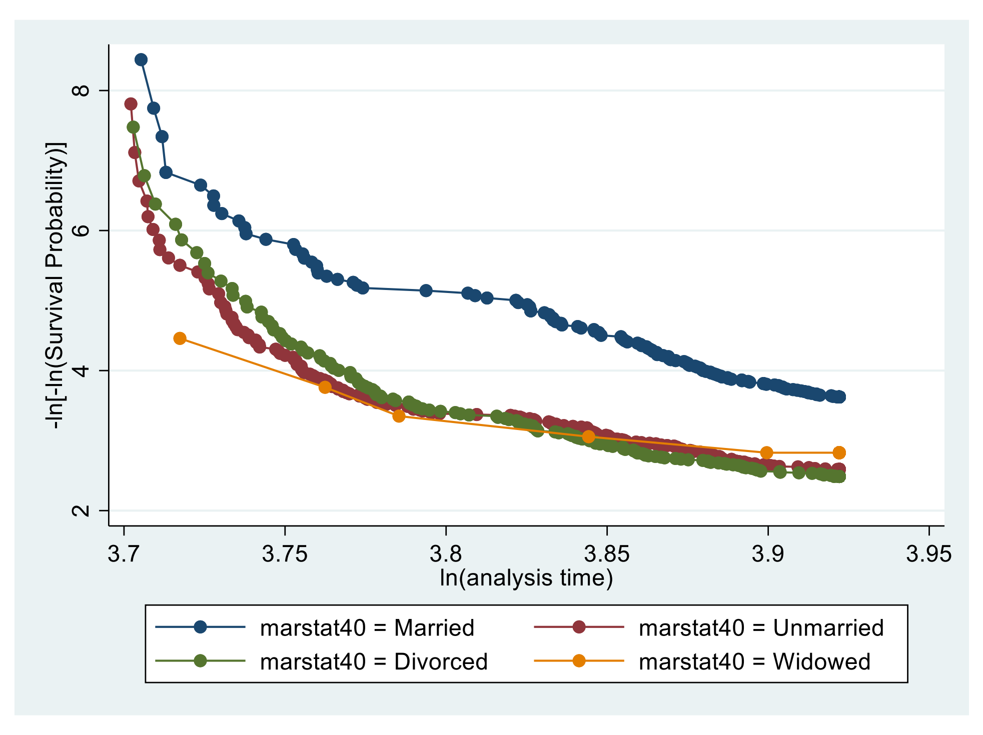
Apart from the very beginning of the curves, they look quite parallel.

We can produce the curves also for our variable marstat40. The quietly option is included in the beginning of the command to suppress the output.

```
quietly stcox ib1.marstat40 if pop_cox==1, noshow
```

Then we use the stphplot command.

```
stphplot, by(marstat40)
```



These curves look quite messy. Apart from the very beginning of the curves, the ones for married and unmarried look parallel, as do the ones for married and divorced. The widowed look worse – most likely due to the small size of the group (with few cases).

17.8.4 Kaplan-Meier and predicted survival plot

Here, we will plot the Kaplan-Meier observed survival curves and compare them to the Cox predicted curves for the same x-variable. If the observed values are close to the predicted values, it is less likely that there is a violation of the proportional hazards assumption.

Note Again, this command does not work very well for continuous x-variables, so we will stick to the categorical ones.

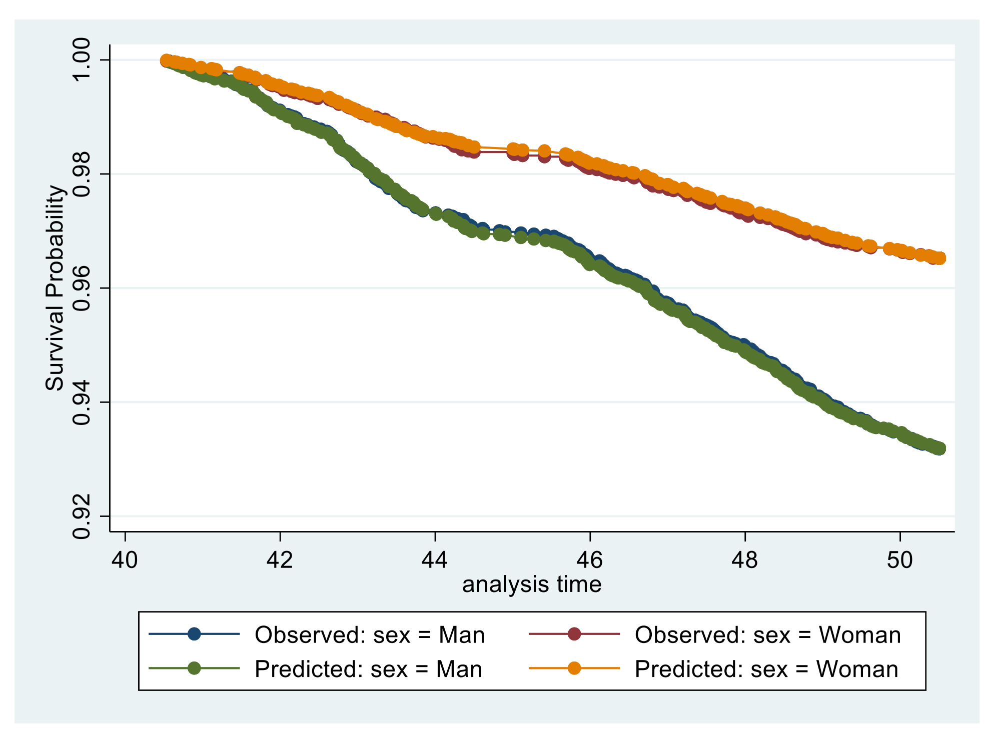
More information	help stcoxkm

Practical example

The first step is re-run a Cox regression model. We will start with the simple one that we did for sex. The quietly option is included in the beginning of the command to suppress the output. After this, we use the stcoxkm command.

```
quietly stcox sex if pop_cox==1, noshow
```

```
stcoxkm, by(sex)
```



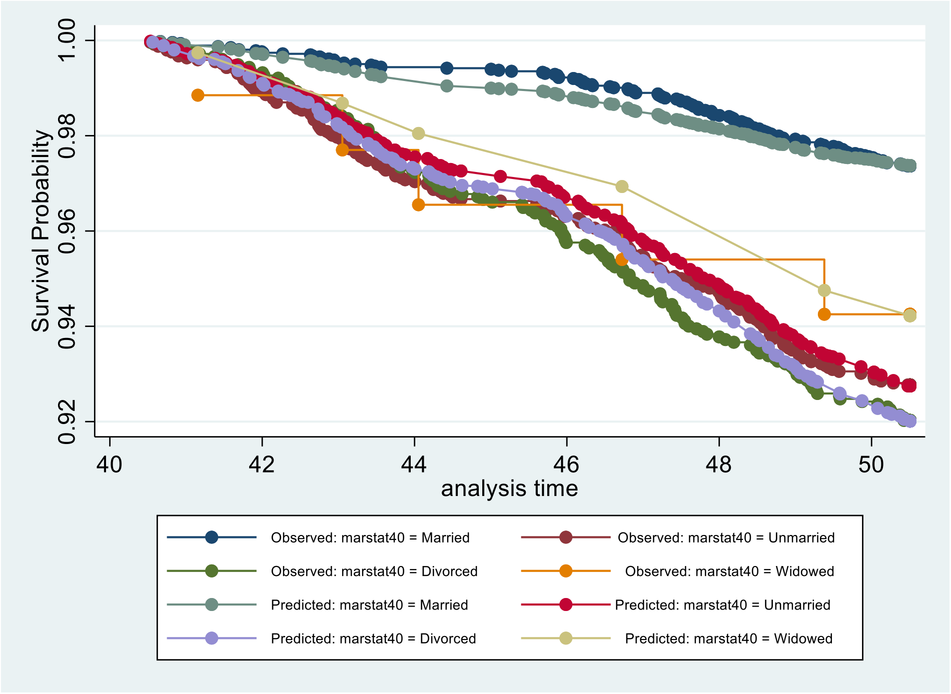
These curves overlap rather nicely in terms of the observed and predicted values.

We can produce the curves also for our variable marstat40. The quietly option is included in the beginning of the command to suppress the output.

```
quietly stcox ib1.marstat40 if pop_cox==1, noshow
```

Then we use the stcoxkm command.

```
stcoxkm, by(marstat40)
```



Overall, the observed and predicted values overlap rather OK. There are some exceptions, especially when it comes to the widowed.

17.8.5 Schoenfeld residuals

Schoenfeld residuals can be used to test the proportionality of the model as a whole. There is also an option to test the proportionality of each x-variable.

More information	help estat phtest

Practical example

The first step is re-run the multiple Cox regression model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly stcox gpa sex ib1.marstat40 if pop_cox==1, noshow
```

Next, we try the estat phtest command.

```
estat phtest, detail
```

```
Test of proportional-hazards assumption

Time:  Time
----------------------------------------------------------------
            |       rho            chi2       df       Prob>chi2
------------+---------------------------------------------------
gpa         |      0.06514         1.71        1          0.1911
sex         |     -0.05301         1.16        1          0.2811
1b.marstat40|            .            .        1               .
2.marstat40 |     -0.17683        12.93        1          0.0003
3.marstat40 |     -0.11754         5.72        1          0.0168
4.marstat40 |     -0.06899         1.97        1          0.1609
------------+---------------------------------------------------
global test |                     17.10        5          0.0043
----------------------------------------------------------------
```

If the p-value (Prob>chi2) is below 0.05, it means that we should reject the proportionality assumption. The global test suggests this model violates the proportionality assumption. As indicated by the p-values for the x-variables, our problem seems to be marstat40. We will not explore this further, but a solution might be to transform marstat40 into a binary variable instead (e.g. Married vs. not Married).

17.8.6 Tied failure times

Though the Cox model assumes that the hazard function is continuous, and therefore are no tied failure times, ties nonetheless occur. Stata provides four options for dealing with tied failures in your data when calculating the partial likelihood. A brief explanation for these methods given a hypothetical tie between two individuals is outlined below.

Method	**Explanation**
Exact marginal calculation	In this method, we assume that time is continuous, the two individuals did not really fail at the same time, but our measurements are imprecise. So, we do not know the order in which they failed. The likelihood calculation is based on the probability that the two individuals fail in any order, which is the sum of the probability that individual 1 fails first + the probability that individual 2 fails first.
Exact partial calculation	In this method, we assume that time is discrete and that the two individuals really did fail at the same time. So, we treat it as a multinomial problem where the conditional probability is derived from a set of possibilities. This method can take a long time to calculate and may produce questionable results if the risk sets are large and with many ties.
Breslow approximation	Approximates the exact marginal calculation by using a common denominator for all failure events. In other words, the risk sets for the second – nth failure events are not adjusted for previous failures. This is the fastest method, and works better if the number of failures is small relative to the size of the risk set.
Efron approximation	Also approximates the exact marginal calculation, but the risk set for the second – nth failure events are adjusted using probability weights. The Efron approximation is more accurate than the Breslow approximation but is relatively slower to calculate.

In Stata, the Breslow method is the default method, and does not need to be specified. You may remember seeing “Cox regression -- Breslow method for ties” in the output from your practical examples earlier in this section. If one of the other three methods is more suitable, you can specify the method after the comma per the below examples:

```
stcox var1…varx, efron
```

```
stcox var1…varx, exactm
```

```
stcox var1…varx, exactp
```

Note If there are no ties in your data, you will obtain identical results, no matter which method you select. Having a few ties in your data will also not yield wildly different results.

You can also check the number of ties in your data in Stata.

First save your data! Then, after your data have been stset, keep only the failures:

```
keep if _d
```

Sort by time:

```
sort _t
```

Generate a count of the instances of time:

```
by _t : gen number = _n
```

Keep one observation representing time:

```
by _t : keep if _n==1
```

Check the average number of failures per time:

```
summarize number
```

Check the frequency of the number of failures:

```
tab number
```

Note You can use the preserve command before dropping observations, and the restore command at the end to return your data to its original state.

17.9 Laplace regression

Finally, we would like to make you aware that a viable alternative (or complement) to Cox regression is Laplace regression. Laplace regression can be used to estimate the effect of exposures (or treatments) on survival percentiles and thereby it allows for direct interpretation of the exposure–outcome association in terms of time gained or lost.

We will not go through Laplace regression in this version of the guide, but if you are interested, we suggest that you install the laplace regression package and then review the help file.

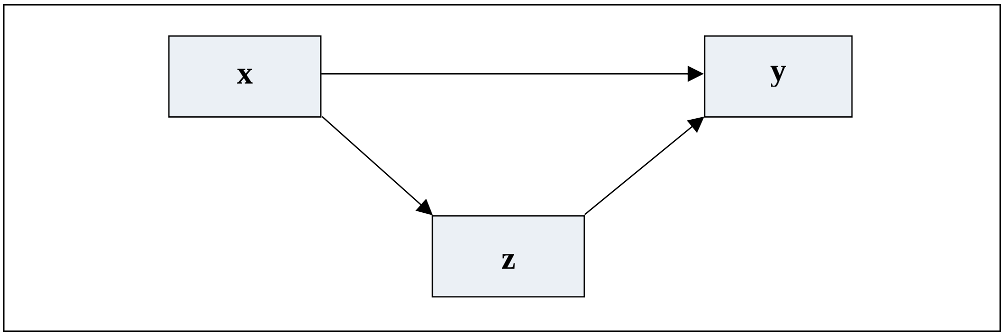
18. MEDIATION ANALYSIS

Outline

Content

In this chapter, we do practical exercises with mediation analysis, using the KHB method.

18.1 Introduction



A mediator is a variable that is influenced by the x-variable and influences the y-variable. In other words, some (it could be a little or a lot) of the effect of x on y is mediated through z.

Some examples
We want to examine the association between occupational class (x) and liver cirrhosis (y). We think that the association may be mediated by alcohol consumption (z). We are interested in the association between poverty (x) and all-cause mortality (y). It is reasonable that this association might to some extent be mediated by stress (z).

18.1.1 Type of regression analysis

In order to carry out a mediation analysis, we first we need to decide on which type of regression analysis that fits our outcome (y) – it could be any type (e.g. linear, logistic, ordinal, multinomial, or some other type). As we have described in this guide, the choice depends largely on the measurement scale of y. Performing mediating analysis in the traditional way (i.e. including covariates in a stepwise fashion and comparing the estimates across models) only works satisfactorily if we perform linear regression. If we do a non-linear regression (e.g. logistic, ordinal, multinomial, Poisson, or Cox), then we should consider a different approach.

18.1.2 Rescaling bias

Why should we avoid the traditional approach with non-linear models? Well, first of all, there are plenty of articles being published that still do mediation analysis in this way. A problem is nonetheless that the non-linear models are not directly comparable (due to rescaling between the models).

There are different types of mediation analysis that one can employ to overcome rescaling bias – one of them is the KHB method (KHB stands for Karlson-Holm-Breen). Just remember that mediation analysis is only as good as your analytical model! Mediation assumes causality, and while using proper mediation analysis is one step in the right direction, causal inference (see Section 9.4) is still an issue if your study is based on observational data.

18.2 Function

KHB can be used for several different types of non-linear regression models. This include logistic (logit), ordinal (ologit), and multinominal (mlogit) – but not Poisson or Cox). What it does is that it compares the effect of x on y in a model without any covariates ("reduced model") with a model with on or more covariates ("full model").

The method is not part of standard Stata, so unless you already have done so, install the KHB package:

```
ssc install khb
```

<table>
<tr><td><b>Basic command</b></td><td colspan="2">khb modeltype yvar xvar || zvar(s)</td></tr>
<tr><td><b>Useful options</b></td><td colspan="2">khb modeltype yvar xvar || zvar(s), disentangle
khb modeltype yvar xvar || zvar(s), summary
khb modeltype yvar xvar || zvar(s), or</td></tr>
<tr><td><b>Explanations</b></td><td>yvar</td><td>Insert the name of the y-variable that you want to use.</td></tr>
<tr><td></td><td>xvar</td><td>Insert the name of the x-variable that you want to use.</td></tr>
<tr><td></td><td>zvar(s)</td><td>Insert the name of the mediator as well as any confounder(s) that you want to include.</td></tr>
<tr><td></td><td>modeltype</td><td>Specify what kind of regression model you want to perform (e.g. logit, ologit, or mlogit).</td></tr>
<tr><td></td><td>disentangle</td><td>Disentangle the contribution of each z-variable.</td></tr>
<tr><td></td><td>summary</td><td>Summary of decomposition.</td></tr>
<tr><td></td><td>or</td><td>Display odds ratios.</td></tr>
<tr><td><b>More information</b></td><td colspan="2">help khb</td></tr>
</table>

18.2.1 Practical example with logistic regression

Here, we want to examine the extent to which the number of best friends (z) mediates the association between parental mental illness (x) and a poor grade point average (y). We also include sex as a confounder (z).

Poor grade point average does not exist as a variable in the dataset, so first we will have to create it based on the variable gpa:

```
gen gpa_dic=gpa
```

```
recode gpa_dic (1.0/2.0=1) (2.1/5.0=0)
```

Dataset: StataData1.dta

Name	**Label**
bestfriends	Number of best friends (Age 15, Year 1985)
gpa_dic	Poor grade point average (Age 15, Year 1985)
parmental	Parental mental illness (Age 15, Year 1985)
sex	Sex

Define the analytical sample

We start by defining the analytical sample:

```
gen pop_mediate1=1 if bestfriends!=. & gpa_dic!=. & parmental!=. & sex!=.
```

Let us have a quick look at the variables:

```
sum bestfriends gpa_dic parmental sex if pop_mediate1==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 bestfriends |      8,604    2.858903    1.111829          1          5
     gpa_dic |      8,604    .0528824    .2238117          0          1
   parmental |      8,604    .0774059    .2672499          0          1
         sex |      8,604    .5256857    .4993688          0          1
```

Regression models

Now, we can run the regression model with the khb command.

```
khb logit gpa_dic parmental || bestfriends sex if pop_mediate1==1, summary disentangle or
```

```
Decomposition using the KHB-Method

Model-Type:  logit                                   Number of obs     =    8604
Variables of Interest: parmental                     Pseudo R2         =    0.09
Z-variable(s): bestfriends sex
------------------------------------------------------------------------------
     gpa_dic |         or   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
parmental    |
     Reduced |   2.093289   .3080201     5.02   0.000     1.568837    2.793061
        Full |   1.817201   .2678855     4.05   0.000     1.361199    2.425963
        Diff |    1.15193   .0413875     3.94   0.000     1.073603    1.235973
------------------------------------------------------------------------------

Summary of confounding

        Variable | Conf_ratio    Conf_Pct   Resc_Fact
    -------------+-------------------------------------
       parmental |  1.2367986       19.15   1.0889539
    ---------------------------------------------------

Components of Difference

      Z-Variable |      Coef     Std_Err      P_Diff  P_Reduced
    -------------+-----------------------------------------------
    parmental    |
     bestfriends |  .1885663   .0341404     133.32      25.53
             sex | -.0471272   .0161484     -33.32      -6.38
    -----------------------------------------------------------
```

The model without any z-variables (the “reduced” model) shows that there is a positive (OR=2.09) and statistically significant association (95 % CI=1.57 to 2.79) between parmental and gpa_dic. In other words, individuals whose parents suffered from mental illness have higher odds of obtaining a poor grade point average. In the model where the z-variables bestfriends and sex are included (the “full” model), the association is weakened but remains stastistically significant (OR=1.82, 95% CI=1.36 to 2.43).

In the table called Summary of confounding, we can see that the amount of the association explained by the z-variables (in this case, bestfriends and sex), is 19%. This amount is specified further in the table called Components of Difference. Here, we can see that the inclusion of bestfriends reduces the association by 26%. There is nonetheless a negative contribution of sex (-6%), meaning that the inclusion of this variable strengthens the association between parental mental illness and poor grade point average. This could (but does not have to) be an indication of an interaction effect by sex (which could be further examined with interaction analysis).

18.2.2 Practical example with ordinal regression

For this example, we want to see if grade point average (z) mediates the association between exposure to bullying (x) and educational level (y).

Dataset: StataData1.dta

Name	**Label**
educ	Educational level (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)
bullied	Exposure to bullying (Age 15, Year 1985)

Define the analytical sample

We start by defining the analytical sample:

```
gen pop_mediate2=1 if educ!=. & gpa!=. & bullied!=.
```

Let us have a quick look at the variables:

```
sum educ gpa bullied if pop_mediate1==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      7,991    2.203103    .7143586          1          3
         gpa |      7,991    3.214178    .6855603        1.1          5
     bullied |      7,991    .1032411    .3042926          0          1
```

Regression models

Now, we can run the regression model with the khb command.

```
khb ologit educ bullied || gpa if pop_mediate2==1, summary disentangle or
```

```
Decomposition using the KHB-Method

Model-Type:  ologit                                Number of obs     =    7991
Variables of Interest: bullied                     Pseudo R2         =    0.12
Z-variable(s): gpa
------------------------------------------------------------------------------
        educ |         or   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
bullied      |
     Reduced |   .6710792   .0484461    -5.53   0.000     .5825382    .7730776
        Full |     .95188   .0689296    -0.68   0.496     .8259298    1.097037
        Diff |    .705004   .0277197    -8.89   0.000      .652715    .7614819
------------------------------------------------------------------------------

Summary of confounding

        Variable | Conf_ratio    Conf_Pct   Resc_Fact
    -------------+-------------------------------------
         bullied |  8.0879499       87.64   1.1856254
    ---------------------------------------------------

Components of Difference

      Z-Variable |       Coef     Std_Err      P_Diff  P_Reduced
    -------------+----------------------------------------------
    bullied      |
             gpa | -.3495518    .0393185      100.00      87.64
    --------------------------------------------------------------------------
```

The model without any z-variables (the “reduced” model) shows that there is a negative (OR=0.67) and statistically significant association (95 % CI=0.58 to 0.77) between bullied and educ. This means that individuals who were exposed to bullying at age 15 have lower odds of attaining a high level of education as adults, in comparison to those who were not exposed to bullying. In the model where the z-variable gpa is included (the “full” model), the association is still negative but very weak and stastistically non-significant (OR=0.95, 95% CI=0.83 to 1.10).

In the table called Summary of confounding, we can see that the amount of the association explained by the z-variables (in this case, we only included gpa), is 88%. This is also shown specified further in the table called Components of Difference.

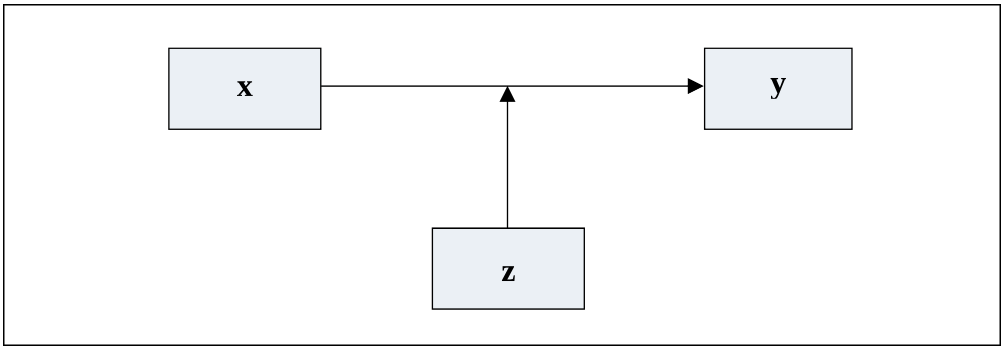
19. INTERACTION ANALYSIS

Outline

Content

In this chapter, we go through two approaches to performing interaction analysis in Stata.

19.1 Introduction



A moderator (or effect modifier) is a variable that influences the very association between the x-variable and the y-variable. Put differently, the association between x and y looks different depending on the value of z.

Some examples

We want to examine the association between social support (x) and mental health (y). We think that the association may be moderated by gender (z). For example, we may expect social support to be more important for mental health among women than among men.

We are interested in the association between mother's educational attainment (x) and babies' birth weight (y). It is reasonable that mother's smoking (z) affects that association: there may be an association between x and y if the mother smokes, but no association between x and y if the mother does not smoke.

19.1.1 Type of regression analysis

In order to carry out an interaction analysis, we first we need to decide on which type of regression analysis that fits our outcome (y) – it could be any type (e.g. linear, logistic, ordinal, multinomial, or some other type). As we have described in this guide, the choice depends largely on the measurement scale of y. Generally, interaction analysis works the same way irrespective of the type of regression analysis that we choose.

19.1.2 Primary approaches to interaction analysis

We will put forward two primary approaches to interaction analysis. They require the following independent variables to be included in the analysis:

	Approach A: **Interaction effect term**	**Approach B:** **Comparison of model fit**
Independent variables included in the model	x (main effect term)	x (main effect term)
	z (main effect term)	z (main effect term)
	x*z The product of x and z (interaction effect term)	x*z The product of x and z (interaction effect term) *or* All possible combinations of x and z includes as dummies

Note This chapter focuses solely on two-way interactions (i.e. interactions between two variables). While it is possible analyse interactions between more than two variables, the interpretation is usually not very straight-forward.

Approach A: Interaction effect term

By including the two main effects (x and z) as well as the interaction effect term in the same model, we can see if the interaction has any effect that goes beyond the main effects. In other words, is the interaction term statistically significant ($p<0.05$)? We also get information about in which direction the interaction effect goes, i.e. what it means, although this effect is not always easy to interpret.

Approach B: Comparison of model fit

This approach is more flexible than the previous one since it is based on comparison of model fit: does a model that includes the main effects as well as a) the interaction effect term *or* b) all possible combinations of x and y included as dummies, fit the data significantly better than a model that just includes the main effects? This can be formally tested by a likelihood ratio test. If the test produces a p-value that is below 0.05, it suggests that the model with the interaction fits the data better. Alternatively, or as a complement, one can compare the Akaike's Information Criterion (AIC), and the Bayesian Information Criterion (BIC) between the models. The model with the lowest values has the better fit.

19.1.3 Measurement scales

The measurement scales of our independent variables (i.e. x and z) are important since it affects what kind of approach that is possible to take.

Measurement scale: x	Measurement scale: z	Possible approach
Binary or Continuous	Binary or Continuous	A, B
Ordinal or Nominal (non-binary)	Ordinal or Nominal (non-binary)	B
Binary or Continuous	Ordinal or Nominal (non-binary)	B
Ordinal or Nominal (non-binary)	Binary or Continuous	B

In other words: having binary and/or continuous variables is the ideal situation since you can use both approaches. As soon as you include ordinal and/or nominal (non-binary) variables, everything becomes more difficult.

19.1.4 Two ways of generating the interaction term

Regardless of whether you choose Approach A or B, here are two ways that you can generate the product term or combination variable. Doing it manually – what we here call Approach 1, requires that you use gen, and sometimes recode and/or if. Approach 2 does it automatically. While we like to do it manually since we feel more in control of what is happening, doing it automatically is of course easier and faster.

Note The manual approach creates interaction terms in the dataset, whereas the automatic approach treats interaction terms as virtual (they do not actually exist in the dataset).

Approach 1: Manual

To illustrate what we mean by a manual approach, we will present two examples. For the first, we create a simple product term whereas, for the second, we create a combination variable.

Product term

Let us assume that we want to see the interaction effect between blood pressure and sex on some outcome. Blood pressure (bp) is a continuous variable whereas sex (sex) is a binary variable (0=Man, 1=Woman). We can simply multiply these terms: x*z

```
gen bp_sex=bp*sex
```

In our model, we would thus include the following independent variables: bp, sex, and bp_sex. This is what it would look like if we did a very basic logistic regression analysis:

```
logistic yvarname bp sex bp_sex
```

Combination variable

If one of our independent variables are ordinal or nominal (non-binary), we cannot multiply them. Instead, we have to create combinations of the variables. Let us now assume that we want to see the interaction effect between stress level and sex on some outcome. Stress level (stress) is an ordinal variable with three categories (1=Low, 2=Medium, 3=High), whereas sex (sex) is a binary variable (0=Man, 1=Woman). In other words, there are six possible combinations. There are many ways that we can use gen, recode, and if to create the combination variable, and this is one of them:

```
gen stress_sex=.
```

```
recode stress_sex (.=1) if stress==1 & sex==0
```

```
recode stress_sex (.=2) if stress==2 & sex==0
```

```
recode stress_sex (.=3) if stress==3 & sex==0
```

```
recode stress_sex (.=4) if stress==1 & sex==1
```

```
recode stress_sex (.=5) if stress==2 & sex==1
```

```
recode stress_sex (.=6) if stress==3 & sex==1
```

We would then include the following independent variables in the model: ib1.stress, sex, and ib1.stress_sex (the choice of reference categories is up to you). This is what it would look like if we did a very basic logistic regression analysis:

```
logistic yvarname ib1.stress sex ib1.stress_sex
```

Approach 2: Automatic

To illustrate what we mean by automatic, we first have to further discuss what factor variables are in Stata.

We have already showed earlier in this guide how factor variables can be used as a way of specifying the reference category of categorical (non-binary) variables that we include in regression analysis (also see Section 11.2.2).

However, this is just one application. We can also use factor variables to denote interactions. There are five factor-variable operators (i.e. prefix) that are possible to use:

Operator	Explanation
i.	Specify an indicator variable.
c.	Specify a continuous variable.
o.	Specify omitted levels (categories) of a variable.
#	Binary operator to specify an interaction.
##	Binary operator to specify factorial interactions.

Product term and combination variable

Let us assume that we want to see the interaction effect between blood pressure and sex on some outcome. Blood pressure (bp) is a continuous variable whereas sex (sex) is a binary variable (0=Man, 1=Woman). If we would specify the interaction with a binary operator, it would look like this:

c.bp#i.sex

In our model, we would include the following: bp, sex, and c.bp#i.sex. This is what it would look like if we did a very basic logistic regression analysis:

```
logistic yvarname bp sex c.bp#i.sex
```

Alternatively, we could have made use of Stata’s factorial interactions:

```
logistic yvarname c.bp##i.sex
```

This would produce exactly the same output.

As you probably figured out already, we do the same if one of our independent variables are ordinal or nominal (non-binary). Let us assume that we want to see the interaction effect between stress level and sex on some outcome. Stress level (stress) is an ordinal variable with three categories (1=Low, 2=Medium, 3=High), whereas sex (sex) is a binary variable (0=Man, 1=Woman). In other words, there are six possible combinations. If we would specify the interaction with a binary operator, it would like this:

i.stress#i.sex

In our model, we would include the following: ib1.stress, sex, and i.stress#i.sex. This is what it would look like if we did a very basic logistic regression analysis:

```
logistic yvarname ib1.stress sex i.stress#i.sex
```

And, of course, we could have made use of Stata's factorial interactions instead:

```
logistic yvarname i.stress##i.sex
```

This would produce exactly the same output.

Note It is possible also to specify the reference category (base level) of interaction terms. For example: c.bp#ib1.sex or ib2.stress##ib0.sex

19.1.5 Interpretation

The most complicated part about interaction analysis is the interpretation. It is important that you keep track how your variables are coded, if you want to say something about what the interaction means. The example below is based on Approach A.

Example

We want to examine the association between social support (x) and happiness (y). We think that the association may be moderated by gender (z). The following hypotheses are formulated: 1) Those with higher levels of social support are more likely to be happy, 2) Women are more likely to be happy, and 3) Social support is more strongly associated with happiness among women than among men.

Since the outcome is binary (0=Not happy and 1=Happy), we choose logistic regression analysis. Social support ranges between 0 and 10, where higher values reflect higher levels of social support. Gender has the values 0=Man and 1=Women.

To begin with, we examine the association between x and y: the odds ratio for social support is 1.20, which confirmed our first hypothesis. Next, we examine the association between z and y: the odds ratio for gender is 1.17, which confirms the second hypothesis. Finally, we include x and z as well as the interaction in the model. The interaction term is statistically significant ($p<0.05$) and the odds ratio is 1.45, which means that the combination of having higher levels of social support and being a woman is associated with increasing chances of being happy.

If the interpretation of the interaction analysis is difficult, you may improve your understanding by doing a separate regression analysis for each category of the z-variable (this is of course only possible if you have a rather large dataset – and thus enough power – and not too many categories in your z-variable). This is sometimes referred to as stratified analyses. However, stratification can also mean many other things in statistics. To make things clear, we will refer to it these kind of separate regression analyses as “specific” – e.g. sex-specific regression analysis.

We can go back to the example to illustrate what specific regression analyses can look like:

Example

We want to examine the association between social support (x) and happiness (y). We think that the association may be moderated by gender (z). The following hypotheses are formulated: 1) Those with higher levels of social support are more likely to be happy, 2) Women are more likely to be happy, and 3) Social support is more strongly associated with happiness among women than among men.

Since the outcome is binary (0=Not happy and 1=Happy), we choose logistic regression analysis. Social support ranges between 0 and 10, where higher values reflect higher levels of social support. Gender has the values 0=Man and 1=Women.

To begin with, we examine the association between x and y *among men only*: the odds ratio for social support is 1.04. Next, we examine the association between x and y *among women only*: the odds ratio for social support is 1.76. Thus, we now see that we have a stronger effect of social support on happiness among women than among men (just like the interaction analysis said).

Remember, however: these kinds of specific or separate analyses are perhaps easier to understand, but if you want to say that any differences between groups (i.e. categories of the z-variable) are statistically significant, you should do a proper interaction analysis.

19.2 Approach A

In the following sections, we will explore how to perform interaction analysis with product terms.

Note We could also have used Approach B in these examples (Approach B can be applied to any type of interaction analysis, whereas the Approach A can only be applied to analyses where we include product terms).

19.2.1 Practical example with linear regression

For this example, we will use Approach A to conduct an interaction analysis based on linear regression. We want to see if sex (z) moderates the association between grade point average (x) and income (y).

Dataset: StataData1.dta

Name	**Label**
income	Annual salary income (Age 40, Year 2010)
gpa	Grade point average (Age 15, Year 1985)
sex	Sex

Define the analytical sample

We start by defining the analytical sample:

```
gen pop_interact1=1 if income!=. & gpa!=. & sex!=.
```

Let us have a quick look at the variables:

```
sum income gpa sex if pop_interact1==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      income |      7,704    200228.5      114478      10000     790000
         gpa |      7,704    3.210228    .6869562          1          5
         sex |      7,704     .498053    .5000287          0          1
```

Simple regression models

First, we will run the simple models, one for gpa and income, and one for sex and income.

```
reg income gpa if pop_interact1==1
```

```
      Source |       SS           df       MS      Number of obs   =     7,704
-------------+----------------------------------   F(1, 7702)      =    405.78
       Model |  5.0524e+12         1  5.0524e+12   Prob > F        =    0.0000
    Residual |  9.5897e+13     7,702  1.2451e+10   R-squared       =    0.0500
-------------+----------------------------------   Adj R-squared   =    0.0499
       Total |  1.0095e+14     7,703  1.3105e+10   Root MSE        =    1.1e+05

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   37281.09   1850.725    20.14   0.000     33653.16    40909.01
       _cons |   80547.65   6075.739    13.26   0.000     68637.55    92457.75
------------------------------------------------------------------------------
```

```
reg income sex if pop_interact1==1
```

```
      Source |       SS           df       MS      Number of obs   =     7,704
-------------+----------------------------------   F(1, 7702)      =    967.44
       Model |  1.1265e+13         1  1.1265e+13   Prob > F        =    0.0000
    Residual |  8.9684e+13     7,702  1.1644e+10   R-squared       =    0.1116
-------------+----------------------------------   Adj R-squared   =    0.1115
       Total |  1.0095e+14     7,703  1.3105e+10   Root MSE        =    1.1e+05

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -76479.13   2458.847   -31.10   0.000    -81299.14   -71659.12
       _cons |   238319.1   1735.279   137.34   0.000     234917.5    241720.7
------------------------------------------------------------------------------
```

There are statistically significant associations in both simple models. More specifically, the B coefficient for gpa is 37281 (95 % CI: 33653 to 40909) and the B coefficient for sex is -76479 (95% CI: -81299 to -71659).

Multiple regression model

Next, we run a model with both independent variables included:

```
reg income gpa sex if pop_interact1==1
```

```
      Source |       SS           df       MS      Number of obs   =     7,704
-------------+----------------------------------   F(2, 7701)      =    883.01
       Model |  1.8831e+13         2  9.4157e+12   Prob > F        =    0.0000
    Residual |  8.2118e+13     7,701  1.0663e+10   R-squared       =    0.1865
-------------+----------------------------------   Adj R-squared   =    0.1863
       Total |  1.0095e+14     7,703  1.3105e+10   Root MSE        =    1.0e+05

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   46087.42   1730.152    26.64   0.000     42695.85    49478.99
         sex |  -85444.38   2376.941   -35.95   0.000    -90103.83   -80784.93
       _cons |   94833.13    5636.71    16.82   0.000     83783.64    105882.6
-------------------------------------------------------------------------------
```

We can note that the B coefficients increase quite a lot (i.e. become further from 0).

Multiple regression model with interaction effect

In this step, we will include the interaction term using Approach 2 (two hashtags mean that we specify the main effects and the interaction effect at the same time):

```
reg income c.gpa##i.sex if pop_interact1==1
```

```
      Source |       SS           df       MS      Number of obs   =     7,704
-------------+----------------------------------   F(3, 7700)      =    634.80
       Model |  2.0017e+13         3  6.6722e+12   Prob > F        =    0.0000
    Residual |  8.0933e+13     7,700  1.0511e+10   R-squared       =    0.1983
-------------+----------------------------------   Adj R-squared   =    0.1980
       Total |  1.0095e+14     7,703  1.3105e+10   Root MSE        =    1.0e+05

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   64007.13   2407.963    26.58   0.000     59286.86    68727.39
             |
         sex |
      Woman  |   31764.21   11287.05     2.81   0.005     9638.512     53889.9
             |
   sex#c.gpa |
      Woman  |  -36487.05   3436.005   -10.62   0.000    -43222.56   -29751.55
             |
       _cons |   39042.93   7675.958     5.09   0.000     23995.96    54089.89
------------------------------------------------------------------------------
```

In the table above, we can see that the estimate for the interaction term has a p-value below 0.05 (0.000). This suggests that there is a statistically significant interaction effect between grade point average and sex on income.

Interpretation

How can we understand this interaction effect that we found? Since our x-variable – sex – is binary, the easiest strategy for gaining more insight would be to do sex-specific analyses of the association between gpa and income.

We start with a model for men, and then continue with the same for women.

```
reg income gpa if pop_interact1==1 & sex==0
```

```
      Source |       SS           df       MS      Number of obs   =     3,867
-------------+----------------------------------   F(1, 3865)      =    514.26
       Model |  7.4266e+12         1  7.4266e+12   Prob > F        =    0.0000
    Residual |  5.5816e+13     3,865  1.4441e+10   R-squared       =    0.1174
-------------+----------------------------------   Adj R-squared   =    0.1172
       Total |  6.3242e+13     3,866  1.6359e+10   Root MSE        =   1.2e+05

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   64007.13   2822.522    22.68   0.000     58473.35     69540.9
       _cons |   39042.93   8997.463     4.34   0.000      21402.7    56683.15
------------------------------------------------------------------------------
```

```
reg income gpa if pop_interact1==1 & sex==1
```

```
      Source |       SS           df       MS      Number of obs   =     3,837
-------------+----------------------------------   F(1, 3835)      =    202.31
       Model |  1.3250e+12         1  1.3250e+12   Prob > F        =    0.0000
    Residual |  2.5117e+13     3,835  6.5494e+09   R-squared       =    0.0501
-------------+----------------------------------   Adj R-squared   =    0.0499
       Total |  2.6442e+13     3,836  6.8931e+09   Root MSE        =     80928

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   27520.07   1934.828    14.22   0.000     23726.68    31313.46
       _cons |   70807.13   6532.148    10.84   0.000     58000.31    83613.95
------------------------------------------------------------------------------
```

The sex-specific models show that the slope in income according to grade point average is steeper among men compared to among women.

Illustration

In order to illustrate the interaction, we can use the margins command. The first step is re-run the model. The quietly option is included in the beginning of the command to suppress the output.

```
quietly reg income c.gpa##i.sex if pop_interact1==1
```

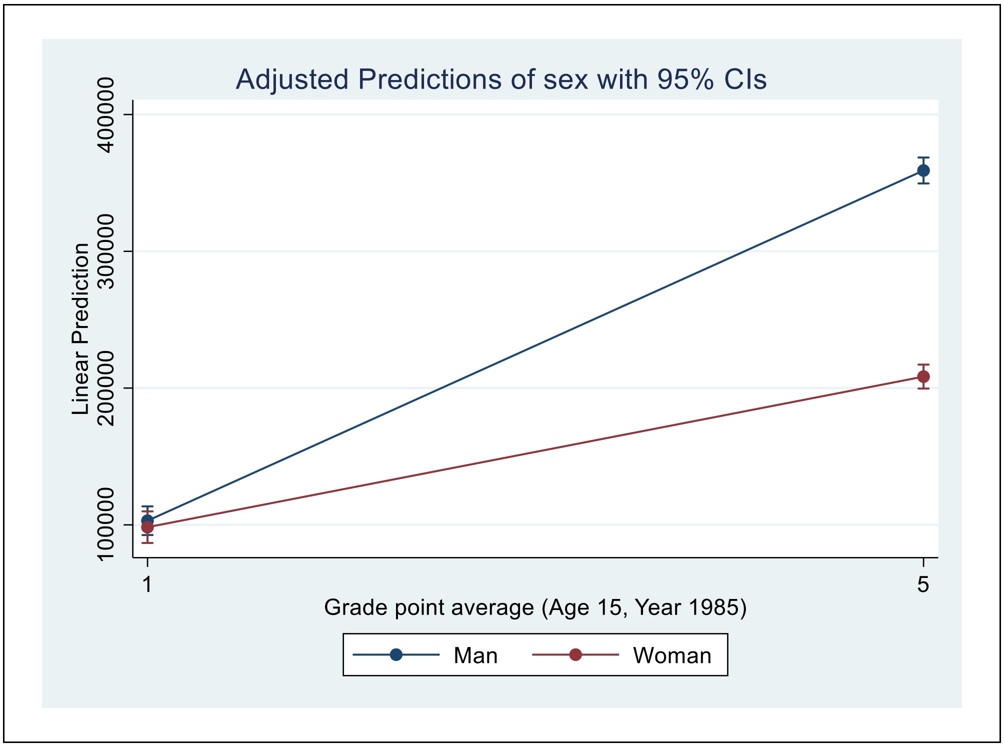
Then we can produce the margins (the quietly option is included here as well):

```
quietly margins sex, at(gpa=(1 5))
```

Note We specify 1 and 5 here since they represent the lowest and highest values for the variable gpa.

And then, finally, it is time to produce the marginsplot:

```
marginsplot
```



Summary
There is a positive, statistically significant association between grade point average at age 15 and income at age 40. While this association exists among men and women alike, the slope is steeper among men.

19.2.2 Practical example with logistic regression

For this example, we will use Approach A to conduct an interaction analysis based on logistic regression. We want to see if sex (z) moderates the association between out-patient care due to cardiovascular disease (x) and early retirement (y).

Dataset: StataData1.dta

Name	**Label**
earlyret	Early retirement (Age 50, Year 2020)
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
sex	Sex

Define the analytical sample

We start by defining the analytical sample:

```
gen pop_interact2=1 if earlyret!=. & cvd!=. & sex!=.
```

Let us have a quick look at the variables:

```
sum earlyret cvd sex if pop_interact2==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earlyret |      8,773    .1371253    .3439992          0          1
         cvd |      8,773    .0457084    .2088639          0          1
         sex |      8,773    .4934458    .4999855          0          1
```

Simple regression models

First, we will run the simple models, one for cvd and earlyret, and one for sex and earlyret.

```
logistic earlyret cvd if pop_interact2==1
```

```
Logistic regression                             Number of obs     =      8,773
                                                LR chi2(1)        =     231.09
                                                Prob > chi2       =     0.0000
Log likelihood = -3391.1166                     Pseudo R2         =     0.0329

------------------------------------------------------------------------------
    earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cvd |   5.594374   .5930068    16.24   0.000     4.544893    6.886196
       _cons |    .139823   .0046581   -59.05   0.000     .1309849    .1492574
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

```
logistic earlyret sex if pop_interact2==1
```

```
Logistic regression                             Number of obs     =      8,773
                                                LR chi2(1)        =      43.76
                                                Prob > chi2       =     0.0000
Log likelihood = -3484.7796                     Pseudo R2         =     0.0062

------------------------------------------------------------------------------
    earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.511296   .0949347     6.57   0.000     1.336226    1.709304
       _cons |   .1276326   .0060431   -43.48   0.000     .1163212    .1400439
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Therese are statistically significant associations in both simple models. More specifically, the OR for cvd is 5.59 (95% CI: 4.54-6.89) and the OR for sex is 1.51 (95% CI: 1.34-1.71).

Multiple regression model

Next, we run a model with both independent variables included:

```
logistic earlyret cvd sex if pop_interact2==1
```

```
Logistic regression                               Number of obs     =      8,773
                                                  LR chi2(2)        =     295.99
                                                  Prob > chi2       =     0.0000
Log likelihood = -3358.6649                       Pseudo R2         =     0.0422

------------------------------------------------------------------------------
    earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cvd |   6.228446   .6726551    16.94   0.000     5.040252    7.696745
         sex |   1.678049   .1089692     7.97   0.000     1.477506    1.905811
       _cons |   .1052034   .0053898   -43.95   0.000     .0951527    .1163158
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Actually, both ORs increase a bit in this model.

Multiple regression model with interaction effect

In this step, we will include the interaction term using Approach 2 (two hashtags mean that we specify the main effects and the interaction effect at the same time):

```
logistic earlyret i.cvd##i.sex if pop_interact2==1
```

```
Logistic regression                             Number of obs     =      8,773
                                                LR chi2(3)        =     296.14
                                                Prob > chi2       =     0.0000
Log likelihood = -3358.5926                     Pseudo R2         =     0.0422

------------------------------------------------------------------------------
    earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cvd |
        Yes  |   6.427069   .8723966    13.71   0.000     4.925754    8.385969
             |
         sex |
      Woman  |   1.691375   .1153998     7.70   0.000     1.479666    1.933374
             |
     cvd#sex |
  Yes#Woman  |   .9184989    .205248    -0.38   0.704     .5927464    1.423273
             |
       _cons |   .1047065   .0055303   -42.72   0.000     .0944095    .1161266
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

In the table above, we can see that the estimate for the interaction term has a p-value above 0.05 (0.704). This suggests that there is no statistically significant interaction effect between out-patient care due to CVD and sex on early retirement.

Note The reference category for the interaction term is by default combination with the smallest value, in this case No#Man. The reason that some combinations are omitted is because they correlate perfectly with the main effect terms.

Summary
Sex does not seem to moderate the association between out-patient care due to CVD and early retirement.

19.3 Approach B

In the following sections, we will explore how to perform interaction analysis based on comparison of model fit.

19.3.1 Practical example with logistic regression

For this example, we will use Approach B to conduct an interaction analysis based on logistic regression. We want to see if educational level (z) moderates the association between marital status (x) and early retirement (y).

Dataset: StataData1.dta

Name	**Label**
earlyret	Early retirement (Age 50, Year 2020)
marstat40	Marital status (Age 40, Year 2010)
educ	Educational level (Age 40, Year 2010)

Define the analytical sample

We start by defining the analytical sample:

```
gen pop_interact3=1 if earlyret!=. & marstat40!=. & educ!=.
```

Let us have a quick look at the variables:

```
sum earlyret marstat40 educ if pop_interact3==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earlyret |      8,668    .1363636    .3431941          0          1
   marstat40 |      8,668    1.693932    .8151531          1          4
        educ |      8,668    2.185971    .7223392          1          3
```

Simple regression models

First, we will run the simple models, one for marstat40 and earlyret, and one for educ and earlyret.

```
logistic earlyret i.marstat40 if pop_interact3==1
```

```
Logistic regression                             Number of obs     =      8,668
                                                LR chi2(3)        =     207.50
                                                Prob > chi2       =     0.0000
Log likelihood = -3348.7775                     Pseudo R2         =     0.0301

------------------------------------------------------------------------------
    earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   marstat40 |
  Unmarried  |   2.541609   .1887313    12.56   0.000     2.197361    2.939788
   Divorced  |   2.341314   .1921366    10.37   0.000     1.993458    2.749871
    Widowed  |   4.220408   1.043441     5.82   0.000     2.599596    6.851773
             |
       _cons |   .0947776   .0050087   -44.59   0.000      .085452    .1051209
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

```
logistic earlyret i.educ if pop_interact3==1
```

```
Logistic regression                             Number of obs     =      8,668
                                                LR chi2(2)        =     203.20
                                                Prob > chi2       =     0.0000
Log likelihood = -3350.9277                     Pseudo R2         =     0.0294

----------------------------------------------------------------------------------
        earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
            educ |
Upper secondary  |   .6288499   .0472754    -6.17   0.000     .5426949    .7286824
     University  |   .2938966   .0263025   -13.68   0.000     .2466128    .3502463
                 |
           _cons |   .2829736   .0170354   -20.97   0.000     .2514794    .3184121
----------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

We see from the output in the tables above that there are quite clear (and statistically significant) associations between marital status and early retirement on the one hand, and educational level and early retirement on the other hand.

Multiple regression model

Next, we run a model with both independent variables included:

```
logistic earlyret i.marstat40 i.educ if pop_interact3==1
```

```
Logistic regression                             Number of obs     =      8,668
                                                LR chi2(5)        =     363.28
                                                Prob > chi2       =     0.0000
Log likelihood = -3270.8877                     Pseudo R2         =     0.0526

---------------------------------------------------------------------------------
        earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+---------------------------------------------------------------
       marstat40 |
      Unmarried  |   2.302619   .1732414    11.09   0.000     1.986921    2.668478
       Divorced  |   2.101947   .1747253     8.94   0.000     1.785933    2.473878
        Widowed  |   3.994096   1.003722     5.51   0.000     2.440678    6.536219
                 |
            educ |
Upper secondary  |   .6614076   .0503735    -5.43   0.000     .5696928    .7678875
     University  |   .3349882   .0304173   -12.04   0.000      .280375    .4002392
                 |
           _cons |   .1687758   .0129286   -23.23   0.000     .1452467    .1961164
---------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

The ORs for marstat40 have decreased a bit, whereas the ORs for educ are actually slightly larger.

Now, we need to save the estimates from this model:

```
estimates store model1
```

Multiple regression model with interaction effect

In this step, we will include the interaction term using Approach 2 (two hashtags mean that we specify the main effects and the interaction effect at the same time):

```
logistic earlyret i.marstat40##i.educ if pop_interact3==1
```

```
Logistic regression                             Number of obs     =      8,668
                                                LR chi2(11)       =     366.78
                                                Prob > chi2       =     0.0000
Log likelihood =  -3269.136                     Pseudo R2         =     0.0531

------------------------------------------------------------------------------------------
                  earlyret | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------------+--------------------------------------------------------------
                 marstat40 |
                Unmarried  |   2.469909   .3572869     6.25   0.000     1.860157    3.279535
                 Divorced  |   1.828658   .2992402     3.69   0.000     1.326913    2.520127
                  Widowed  |   4.717526   2.296895     3.19   0.001     1.816672    12.25045
                           |
                      educ |
          Upper secondary  |   .6494543   .0870426    -3.22   0.001     .4994212    .8445596
               University  |   .3397788   .0503307    -7.29   0.000     .2541615    .4542373
                           |
            marstat40#educ |
Unmarried#Upper secondary  |   .9062763   .1638395    -0.54   0.586     .6358837    1.291646
     Unmarried#University  |    .901407   .1899455    -0.49   0.622     .5964235    1.362345
 Divorced#Upper secondary  |   1.246885   .2497817     1.10   0.271     .8419957    1.846473
      Divorced#University  |   1.140599   .2742129     0.55   0.584     .7120237    1.827138
  Widowed#Upper secondary  |   .9238524   .5591375    -0.13   0.896     .2821209    3.025311
       Widowed#University  |   .5886182   .4313178    -0.72   0.470     .1399924     2.47493
                           |
                     _cons |   .1695804   .0186211   -16.16   0.000     .1367439     .210302
------------------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Similar to the previous model, we have to save the estimates from this one:

```
estimates store model2
```

Next, we compare the fit of the two models:

```
lrtest model1 model2, stats
```

```
Likelihood-ratio test                                 LR chi2(6)  =      3.50
(Assumption: model1 nested in model2)                 Prob > chi2 =    0.7435

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
      model1 |      8,668  -3452.526  -3270.888       6    6553.775    6596.18
      model2 |      8,668  -3452.526  -3269.136      12    6562.272   6647.081
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

Note We called our saved models “model1” and “model2”, but you can choose any name you like.

In the table above, we can see that the p-value for the likelihood ratio test is above 0.05 (0.7435), which suggests that model that contains that interaction term (model2) does not fit the data better than the model without the interaction term (model1). This is also confirmed by the values for AIC and BIC, which are lower for model1. We can thereby conclude that there is no statistically significant interaction between marital status and educational level in the effect on early retirement.

Summary

Educational attainment does not seem to moderate the association between marital status and early retirement.

19.3.2 Practical example with Cox regression

For this example, we will use Approach B to conduct an interaction analysis based on Cox regression. We want to see if sex (z) moderates the association between body mass index (x) and out-patient care due to CVD (y).

Note Here, we use the same stset as in Chapter 17.

Dataset: StataData1.dta

Name	**Label**
cvd	Out-patient care due to CVD (Ages 41-50, Years 2011-2020)
bmi	Body mass index (Age 20, Year 1900)
sex	Sex

Define the analytical sample

We start by defining the analytical sample:

```
gen pop_interact4=1 if cvd!=. & bmi!=. & sex!=.
```

Let us have a quick look at the variables:

```
sum cvd bmi sex if pop_interact4==1
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         cvd |      8,385     .047585    .2128992          0          1
         bmi |      8,385    22.64526     3.50581   10.97624   39.25653
         sex |      8,385    .5177102     .499716          0          1
```

Simple regression models

First, we will run the simple models, one for bmi and cvd, and one for sex and cvd.

```
stcox bmi if pop_interact4==1
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,385                  Number of obs    =      8,385
No. of failures =          399
Time at risk    =  81871.70157
                                                LR chi2(1)       =       0.05
Log likelihood  =   -3594.9981                  Prob > chi2      =     0.8162

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   1.003322   .0143091     0.23   0.816     .9756654    1.031763
------------------------------------------------------------------------------
```

```
stcox sex if pop_interact4==1
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,385                  Number of obs    =      8,385
No. of failures =          399
Time at risk    =  81871.70157
                                                LR chi2(1)       =      45.59
Log likelihood  =    -3572.231                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5015805   .0525306    -6.59   0.000      .408502    .6158672
------------------------------------------------------------------------------
```

The tables above show that there is no association between bmi and cvd, whereas there is a clear and statistically significant association between sex and cvd.

Multiple regression model

Next, we run a model with both independent variables included:

```
stcox bmi sex if pop_interact4==1
```

```
Cox regression -- Breslow method for ties

No. of subjects =        8,385                  Number of obs    =      8,385
No. of failures =          399
Time at risk    =  81871.70157
                                                LR chi2(2)       =      46.60
Log likelihood  =   -3571.7276                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9851039   .0147709    -1.00   0.317     .9565748    1.014484
         sex |   .4920975   .0524188    -6.66   0.000     .3993743    .6063482
------------------------------------------------------------------------------
```

The HRs for bmi and sex have decreased a tiny bit (in this case, become further from 1).

Now, we need to save the estimates from this model:

```
estimates store model1
```

Multiple regression model with interaction effect

In this step, we will include the interaction term using Approach 2 (two hashtags mean that we specify the main effects and the interaction effect at the same time):

```
stcox c.bmi##i.sex if pop_interact4==1
```

```
Cox regression -- Breslow method for ties

No. of subjects =         8,385                 Number of obs    =      8,385
No. of failures =           399
Time at risk    =   81871.70157
                                                LR chi2(3)       =      51.80
Log likelihood  =    -3569.1249                 Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9562677   .0191093    -2.24   0.025     .9195381    .9944644
             |
         sex |
      Woman  |   .1044956   .0719111    -3.28   0.001     .0271221    .4025988
             |
   sex#c.bmi |
      Woman  |   1.071326   .0322614     2.29   0.022     1.009925    1.136461
------------------------------------------------------------------------------
```

Similar to the previous model, we have to save the estimates from this one:

```
estimates store model2
```

Note Since bmi is continuous and sex is binary, we actually do not need to compare the model fit using Approach B – it would be sufficient with Approach A (we can already see in the table that the interaction term is statistically significant). But since Approach B is what the example is about, we will do it for the sake of practice.

Next, we compare the fit of the two models:

```
lrtest model1 model2, stats
```

```
Likelihood-ratio test                                 LR chi2(1)  =      5.21
(Assumption: model1 nested in model2)                 Prob > chi2 =    0.0225

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
      model1 |      8,385  -3595.025  -3571.728       2    7147.455   7161.524
      model2 |      8,385  -3595.025  -3569.125       3     7144.25   7165.352
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

Note We called our saved models “model1” and “model2”, but you can choose any name you like.

In the table above, we can see that the p-value for the likelihood ratio test is below 0.05 (0.0225), which suggests that model that contains that interaction term (model2) fits the data better than the model without the interaction term (model1). This is also confirmed by the values for AIC, which are lower for model2. The BIC value is not lower for model2 than model1, but it should be noted that BIC tend to penalise complex models more than AIC does.

We can thereby conclude that there is a statistically significant interaction between body mass index and sex in the effect on out-patient care due to CVD.

Interpretation

How can we understand this interaction effect that we found? Since our x-variable – sex – is binary, the easiest strategy for gaining more insight would be to do sex-specific analyses of the association between bmi and cvd.

We start with a model for men, and then continue with the same for women.

```
stcox bmi if pop_interact4==1 & sex==0
```

```
Cox regression -- Breslow method for ties

No. of subjects =        4,044                  Number of obs    =      4,044
No. of failures =          258
Time at risk    =  39187.84668
                                                LR chi2(1)       =       5.08
Log likelihood  =   -2131.7739                  Prob > chi2      =     0.0243

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9562525   .0191098    -2.24   0.025      .919522    .9944501
------------------------------------------------------------------------------
```

```
stcox bmi if pop_interact4==1 & sex==1
```

```
No. of subjects =        4,341                  Number of obs    =      4,341
No. of failures =          141
Time at risk    =  42683.85489
                                                LR chi2(1)       =       1.14
Log likelihood  =   -1178.1281                  Prob > chi2      =     0.2857

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   1.024468   .0230777     1.07   0.283     .9802202    1.070713
------------------------------------------------------------------------------
```

The sex-specific models show that the association between body mass index and out-patient care due to CVD is actually negative (and statistically significant) for men and positive (and statistically non-significant) for women. This explains why we did not find any association at all in our first simple model.

Illustration

We would also like to show how the interaction effect can be illustrated. Unfortunately, it not very straight-forward to graph interactions for non-linear outcomes. A solution here might be is to categorise bmi. For example, we could categorise it into underweight, normal weight, overweight, and obese in according to WHO standards.

```
gen bmi_cat=bmi
```

```
recode bmi_cat (0/18.4999=1) (18.5000/24.999=2) (25.000/29.999=3) (30.000/40=4)
```

Let us also create and apply some value labels:

```
label define bmi_cat 1 "Underweight" 2 "Normal weight" 3 "Overweight" 4 "Obese"
```

```
label values bmi_cat bmi_cat
```

We choose Normal weight (value 2) as out reference category for bmi, and Women (value 1) as our reference category for sex.

Note Even though we usually do not have to specify a reference category for binary variables (such as sex), margins will not work without it.

First, we run a Cox regression model with these two variables. We use the quietly option to supress the output.

```
quietly stcox ib2.bmi_cat##ib1.sex if pop_interact4==1
```

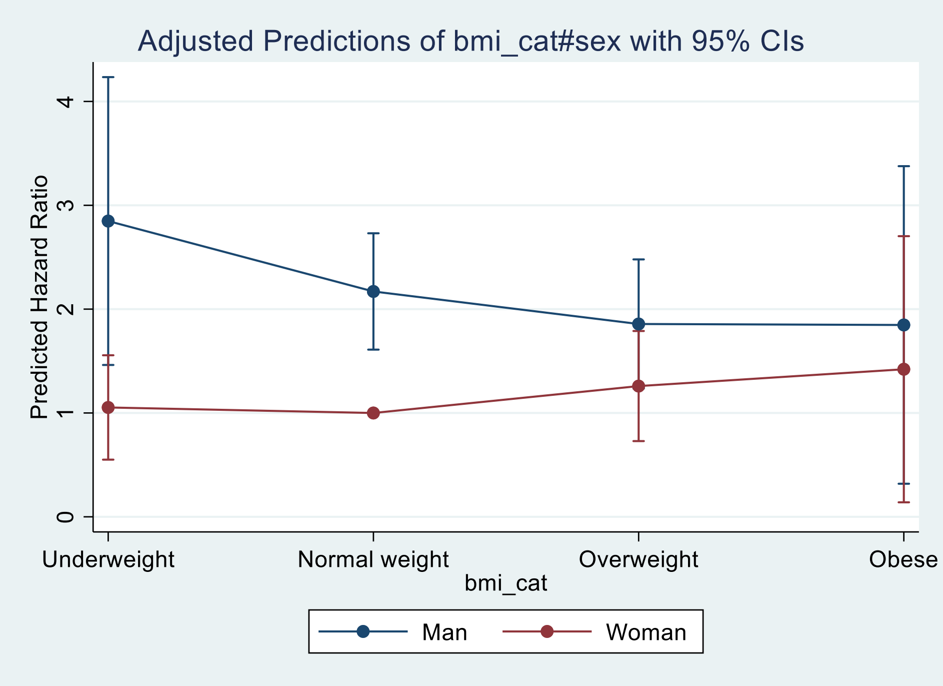
Then we type the margins command, with suppressed output here as well:

```
quietly margins ib2.bmi_cat#ib1.sex
```

Note We do not use double hashtags here to specify the main effects and the interaction effect, since it produces an additional line in the marginsplot.

And run the marginsplot:

```
marginsplot
```



Note The reference combination (predicted HR=1) includes women who are normal weight. All other estimates are compared to this combination. If we were to choose another reference combination, the graph would look quite different.

This graph illustrates the negative association for men and the positive association for women. Another way of looking at it is that the difference between men and women when it comes to the association between bmi_cat and cvd is largest among the underweight but becomes smaller as body mass index increases. Generally, the results might not be what we would expect (but remember that this dataset is fictional).

Summary
Gender moderates the association between body mass index and out-patient care due to CVD. More specifically, a higher body mass index is associated with a lower risk of experiencing out-patient care due to CVD among men. Among women, the opposite association was observed.

Research Reports in Public Health Sciences

Stockholm University, 106 91 Stockholm
www.su.se/publichealth | ISSN 2003-0142

Stockholm
University