DATA MANAGEMENT AND INTRODUCTION TO STATA

David Clark

Thursday 12th April

University of Leeds

INTRODUCTIONS AND PREAMBLE

INTRODUCTIONS

David Clark

- Teaching Fellow in Economics
- Used Stata for 6(-ish) years … still learning!

Kausik Chaudhuri

- Senior Lecturer in Economics
- Used Stata for .....

WHO IS THIS COURSE FOR?

Targeted at anyone who has **no to little** experience using Stata

Primarily for those engaging in quantitative research (MRes/PhD)

What to learn to use a statistical package that allows for both use of **point-and-click GUI** and **Stata's Markup and Control Language (SMCL)**

COURSE MATERIALS

All the materials for the course are hosted in my GitHub repo:

https://github.com/davidptclark/pgr_stata_course

Press the green button labelled Clone or download and then press Download ZIP

The structure of the directories and folders is as follows:

- Course_Slides: Contains the slides for both part of the course:
    - part_1: Morning session
    - part_2: Afternoon session
- practical_files: Contains the exercises and data for the course:
    - part_1: Morning session
    - part_2: Afternoon session

WHO IS THIS COURSE FOR?

For those who want to:

- Organise and manage data
    - Generating, keeping and dropping, reshaping
- Visualise data
    - Scatter and line graphs
    - Histograms
- Analyse data
    - ANOVA
    - Regression analysis
- Automate and reproduce workflow
    - Log and do-files
    - Loops

WHAT IS STATA?

WHAT IS STATA?

Stata is a powerful statistical package with:

- smart data-management facilities
- a wide array of up-to-date statistical techniques
- an excellent system for producing publication-quality graphs

Available on a variety of operating systems (Windows, Mac OS and Linux distributions)

Also available in different varieties:

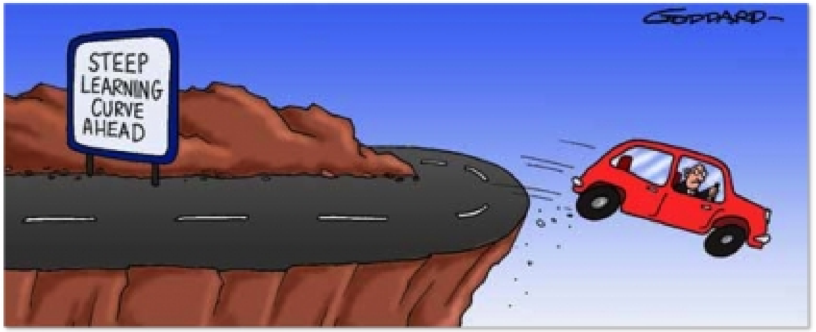
- IC (standard)
- SE (extended)
- MP (multiprocessing)

WHY NOT USE X?

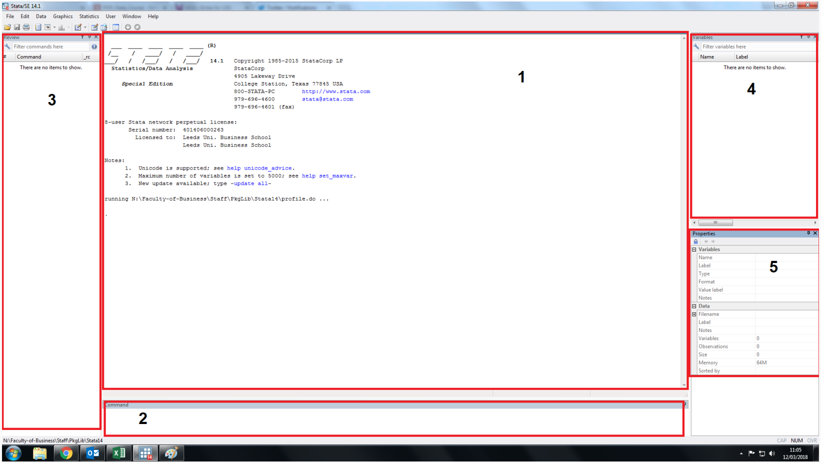
There are alternative statistical software packages you can use (to name a few):

- R
- Matlab
- SAS
- SPSS
- Gauss
- Gretl
- Eviews

CRIPPLING SELF-DOUBT



STATA 14 FRONT END GRAPHIC USER INTERFACE (GUI)



There are no items to show.

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   14.1   Copyright 1985-2015 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

8-user Stata network perpetual license:
       Serial number:  401406000263
         Licensed to:  Leeds Uni. Business School
                       Leeds Uni. Business School

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
      3.  New update available; type -update all-

running N:\Faculty-of-Business\Staff\PkgLib\Stata14\profile.do ...

.
```

There are no items to show.

Variables
- Name
- Label
- Type
- Format
- Value label
- Notes

Data
- Filename
- Label
- Notes
- Variables 0
- Observations 0
- Size 0
- Memory 64M
- Sorted by

N:\Faculty-of-Business\Staff\PkgLib\Stata14

STATA 14 FRONT END GUI

Stata has an menu bar on the top and 5 internal windows.

The main window is the one in the middle (1 on the previous slide). It gives you all the output of your operations in Stata.

The command window (2) executes commands.

- You can type commands directly in this window as an alternative to using the menu system.
- Stata will show you what the written command is for each action performed using the drop-down menus.

STATA 14 FRONT END GUI

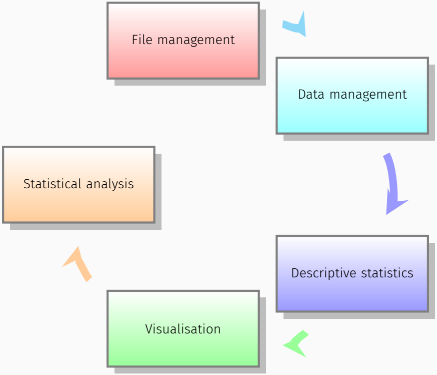
The review window (3), lists all the operations preformed since opening Stata. If you click on one of your past commands, you will see the command being displayed in the Command window and you can re-run it by hitting the enter key.

The variables window (4) lists the variables in the current dataset (and their descriptions). When you double-click on the variable, it appears in the Command window.

The properties window (5) gives information about your dataset and your variables.

STATA WORKFLOW

STATA WORKFLOW



File management

Data management

Descriptive statistics

Visualisation

Statistical analysis

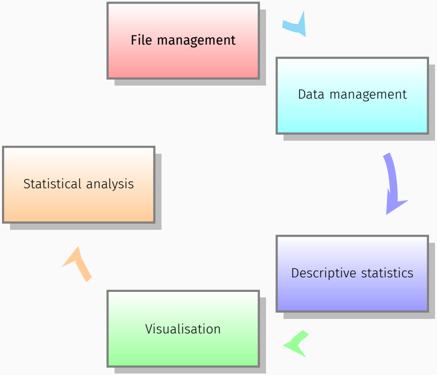
STATA WORKFLOW

- File management
- Data management
- Descriptive statistics
- Visualisation
- Statistical analysis

} These two stages will consume the most time in any research project

FILE MANAGEMENT

STATA WORKFLOW: FILE MANAGEMENT



File management

Data management

Descriptive statistics

Visualisation

Statistical analysis

FILE MANAGEMENT

- This is often an aspect of using Stata that is **wrongly** overlooked
- Usually a facet that people return to after learning the syntax
- As researchers, one of our primary objectives:

Replicability and reliability

- If, after testing your research hypothesis, using data, you discover some results of interest, **what use is this if they cannot be reproduced by others?**
- Hence, engraining good practices from the beginning, promotes higher-quality research in future work

FILE MANAGEMENT

- What do we mean by **file management**?
    - Typically, when people (most) begin using Stata, they will just open some data and **do stuff**
- Questions that arise:
    - **Where is the data stored?**
    - **Where is the output stored?**
    - **Where is Stata currently working from?**
    - **Are we utilising one or many directories?**
- File management is knowing the answer to these questions constantly and having a good justification for their placement

FILE MANAGEMENT

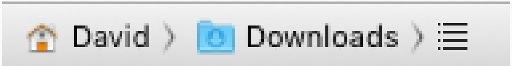
Where is Stata currently working from?

- Definition: **working directory**
  - The (current) working directory is the file within the computer's hierarchical file structure that a program is working from
- That is to say, anything you ask Stata to open or to save will be accessed or stored in this working directory

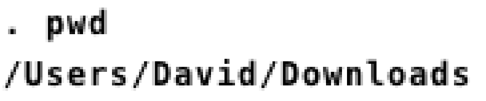
FILE MANAGEMENT

Where is Stata currently working from?

- There are two ways of finding out what the current working directory is in Stata:
  - Look at the bottom-left hand corner of Stata



  - Type the command **pwd** into the command window in Stata



```
. pwd
/Users/David/Downloads
```

- Both are telling us that we are working out of the **Downloads** folder

FILE MANAGEMENT

- On the University system, this usually is set as a default to the personal drive (M:/)
- In either case, **is it a good idea to work out of an indiscriminate folder?**
  - **Almost always, no!**
  - Why? → There will be unrelated files that will make it complex to keep track of related files and output

  So, we have two options what we can proceed with that adhere to **good practice**:
  - Change to a directory that already exists
  - Create a directory to work from

FILE MANAGEMENT: CHANGING DIRECTORY

- If the folder that you want to work from already exists, we can tell Stata to change the working directory to this folder.
- For example, imagine I have a folder called **Thesis_Paper_One** and here is the path (note, this was the file path on my Mac, it will look slightly different on Windows PCs):

Users → **David** → **Documents** → **Projects** → **Thesis_Paper_One**

- This can be done in two ways:
    - Using the drop down menus in the GUI
    - Using the **cwd** command directly

FILE MANAGEMENT: CHANGING DIRECTORY

Using the drop down menus in the GUI

- If you follow this menu path:

File $\rightarrow$ Change working directory...

- Stata will then open a file explorer window where you can navigate to, and choose, the folder you wish to set as the current working directory
- This is a useful method if **you do not have the exact file path to hand**
- Notice, Stata will then print the exact file path in the output window after changing working directory successfully.

FILE MANAGEMENT: CHANGING DIRECTORY

Using the drop down menus in the GUI

- If you already happen to know the file path to the directory, we can type the change directory command directly into the command prompt:

```
cd ”/Users/David/Documents/Projects/Thesis_Paper_One”
```

Breakdown

- `cd`

  Tells Stata to **change directory**

- `”/Users/David/Documents/Projects/Thesis_Paper_One”`

  Provides Stata with the file path to the directory that you will want to change to

FILE MANAGEMENT: CREATING A DIRECTORY

- Perhaps you want to create the folder, as part of a new project, which we'll call **Thesis_Paper_Two**
- Here, we can only use the command prompt, by typing the following command

```
mkdir "/Users/David/Documents/Projects/Thesis_Paper_Two"
```

Breakdown

- `mkdir`

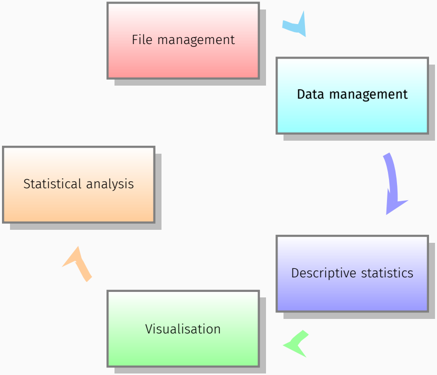
  Tells Stata to **create a new folder in this directory**
- `"/Users/David/Documents/Projects/Thesis_Paper_Two"`

  Provides Stata with the file path to the directory that you will want to move to (Projects) and create a folder in there called Thesis_Paper_Two

DEMO: CHANGING AND CREATING DIRECTORIES

DATA MANAGEMENT

STATA WORKFLOW: DATA MANAGEMENT



DATA MANAGEMENT

- As stated previously, the data management aspect of the workflow is arguably **one of the most important (and time-consuming) stages of a research project**
- **Why?**
  - Data might not be native to Stata, so it must be imported correctly
  - Datasets, particularly survey data, may have some errors in their reporting and may require our attention
  - You may want to gather data from different datasets and consolidate them into one master dataset
  - Perhaps you want to create new variables based on the original data
- Taking the time to carry out this stage properly will **save you time in the long run**

DATA MANAGEMENT: IMPORTING DATA

We can characterise datasets into two broad sets:

- **Stata datafiles** (.dta files)
  - This is the file type that Stata can read and work with natively; as such it requires little work to open
- **All other types of datafile**
  - This is a huge set but these have to imported as a foreign datafile
  - While the list is endless, we will focus only on:
    - CSV
    - XLS
    - XLSX
- If there are other datasets from particular programs you'd like to import, consider the program **StatTransfer -** not free

DATA MANAGEMENT: A QUICK ASIDE

Before we go ahead and learn to import and open different datafile types into Stata, **we need to make sure we're working on a blank canvas**, so to speak If we don't do this beforehand:

1. Stata will **refuse** to open other data as it will risk losing data that is already in memory
2. It may, in theory, **corrupt** the imported data and/or the data in current memory

So to avoid the risk of either of the two above outcomes, **we type in the following command into Stata before importing anything**:

```
clear all
```

This will clear literally everything from Stata's working memory, including, most importantly, any open datafiles

DATA MANAGEMENT: IMPORTING STATA DATAFILES

Stata datafiles (.dta files)

- As previously stated, these are native to Stata and so are very easily open and read by Stata
- **Assuming that the datafile is stored in the current working directory**, use the following drop down menu path:

File → Open

- Once you've done this, you can navigate to the datafile and then double-click to open
- An easy way to verify the data is now loaded into to Stata is to check the variables window
- If you see variables names in there, you've successfully opened your Stata datafile

DATA MANAGEMENT: IMPORTING XLS/X DATAFILES

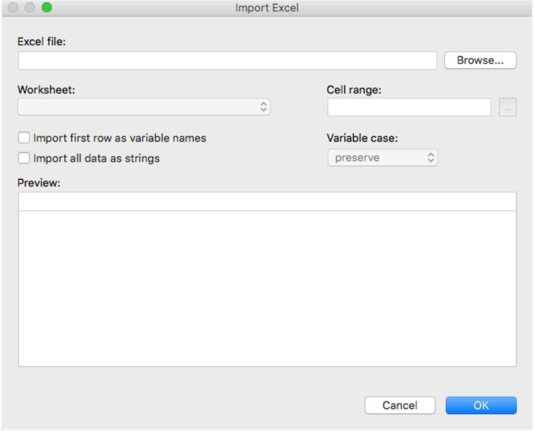
Excel datafiles (.xls/x files)

- Despite secondary data being provided in a .dta format more frequently - **it is more common for the data to be provided as an Excel file** (either .xls or .xlsx file types)
- This is not quite as strightforward as opening a .dta file but it is luckily not too complex an operation
- **Assuming that the datafile is stored in the current working directory**, use the following drop down menu path:

File → Import → Excel Spreadsheet (*.xls;*.xlsx)

DATA MANAGEMENT: IMPORTING XLS/X DATAFILES

After following the drop-down path, you'll be presented with the following window:



☐ Import first row as variable names
☐ Import all data as strings

Variable case: preserve

DATA MANAGEMENT: IMPORTING XLS/X DATAFILES

Complete the following steps to import the datafile:

1. Direct Stata to the datafile by selecting **Browse..**
2. Choose the worksheet that your data is in within the workbook
3. Typically, the heading of each column of the spreadsheet will contain the variable name. Select the option, **Import first row as variable names**
4. In the preview pane at the bottom of the window, you can confirm here that the data is being imported correctly
5. Finally, press OK and Stata should have then successfully imported the Excel datafile

DATA MANAGEMENT: IMPORTING CSV DATAFILES

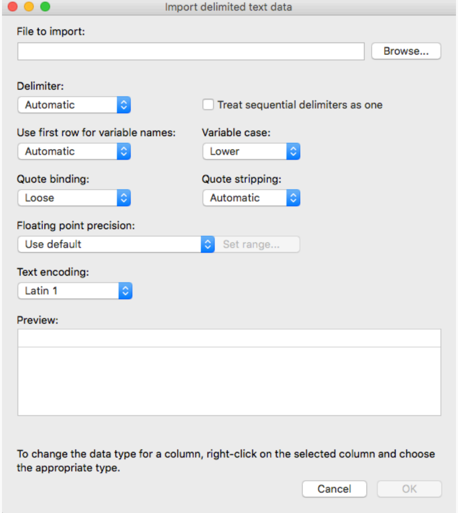
Comma-Seperate Values datafiles (.csv files)

- Other than Excel datafile types, another popular datafile format is the comma-seperate values (CSV) datafiles
- They're popular as they do not require proprietary software to open or edit
- **Assuming that the datafile is stored in the current working directory**, use the following drop down menu path:

File → Import → Text data (delimited;*.csv)

DATA MANAGEMENT: IMPORTING CSV DATAFILES

After following the drop-down path, you'll be presented with the following window:



Import delimited text data

File to import:

Delimiter: Automatic

□ Treat sequential delimiters as one

Use first row for variable names: Automatic

Variable case: Lower

Quote binding: Loose

Quote stripping: Automatic

Floating point precision: Use default

Text encoding: Latin 1

Preview:

To change the data type for a column, right-click on the selected column and choose the appropriate type.

DATA MANAGEMENT: IMPORTING CSV DATAFILES

Complete the following steps to import the datafile:

1. Direct Stata to the datafile by selecting **Browse..**
2. Choose the delimiter (the symbol that Stata recognises as separating individual data observations) - automatic is usually best
3. Just in the case with XLS/X files, if the first row represents the variable names, choose always under the appropriate option
4. Observe the preview and if everything looks in its proper format, press OK

DEMO: IMPORTING DATAFILES

BROWSING AND MANIPULATING DATA

BROWSING AND MANIPULATING DATA

So, regardless of how your data had been stored originally, **you should be aware** of how to import the datafile into Stata

Now what?

This is where we get started with Stata!

What might you want to do immediately after importing data?

1. Browse the data
2. Edit the data
3. Generate new variables

Keep in mind, before we've imported and formatted the data correctly, **we should not be embarking on any type of statistical analysis**

BROWSING DATA

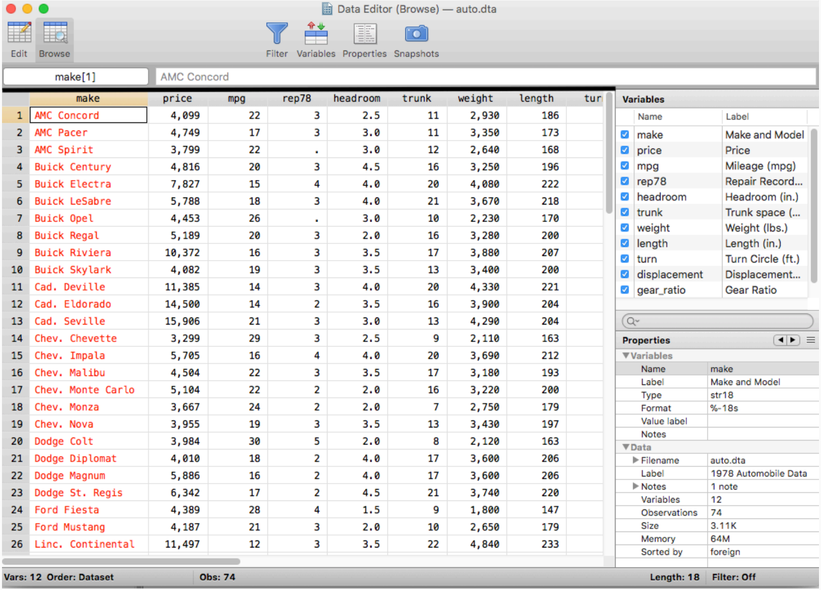
Despite the fact we saw a preview of the data when importing CSV/XLS datafiles, we don’t really know how the data looks

Something we can do is browse the dataset by simply typing:

```
browse
```

When you do so, you should see the following window open on your screen...

BROWSING DATA



make[1] AMC Concord

	make	price	mpg	rep78	headroom	trunk	weight	length	tur
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	
4	Buick Century	4,816	20	3	4.5	16	3,250	196	
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	
25	Ford Mustang	4,187	21	3	2.0	10	2,650	179	
26	Linc. Continental	11,497	12	3	3.5	22	4,840	233	

Variables

	Name	Label
☑	make	Make and Model
☑	price	Price
☑	mpg	Mileage (mpg)
☑	rep78	Repair Record...
☑	headroom	Headroom (in.)
☑	trunk	Trunk space (...
☑	weight	Weight (lbs.)
☑	length	Length (in.)
☑	turn	Turn Circle (ft.)
☑	displacement	Displacement...
☑	gear_ratio	Gear Ratio

Properties

▼Variables	
Name	make
Label	Make and Model
Type	str18
Format	%-18s
Value label	
Notes	
▼Data	
▶ Filename	auto.dta
Label	1978 Automobile Data
▶ Notes	1 note
Variables	12
Observations	74
Size	3.11K
Memory	64M
Sorted by	foreign

Vars: 12 Order: Dataset Obs: 74 Length: 18 Filter: Off

BROWSING DATA

That's great but **what's going on in this huge window?**

1. **Main window**
    - This is main data browser pane within the browse window
    - Here, we can see all our **variables and the observations for each individual unit in the dataset** (notice some are in red? We'll come back to this)
2. **Top-right corner**
    - This is the **variables pane**
    - Here you can see each of the variables within the dataset as well as their label
    - You can (un)tick them to filter the data in the browser pane
3. **Bottom-right corner**
    - This is the **properties** pane
    - It displays information about the selected variable - notably the **variable type**

A QUICK ASIDE - VARIABLE TYPES

In the last window, you've probably noticed that the observations for the variable make were all in **red**, whilst the others were just in **black** text

Any ideas why?

We could go into **a lot** of detail about this but simply put:

- When you see the observations of a variable in **red**, this signifies that the variable is of type **string** - that's to say the data is a word or contains text
- Otherwise, when the data is in black, this usually signifies that data is stored as a **numeric type** (e.g. float, integer, double)

MANIPULATING DATA: EDITING

When it comes to **manipulating/changing the data**, for whatever reason, there are a few ways of doing this:

1. Editing the data within the browser directly (**not recommended**)
   - Why? → piecemeal changes of the data doesn't speak to **replicability and reliability**
2. Using a suite of commands that allow you to change aspects of a variable (**recommended**) - these include:
   - `rename`

     - renaming variables
   - `keep`

     - keep a variable or observations in a range
   - `drop`

     - drop a variable or observations in a range
   - `generate`

     - create a new variable

MANIPULATING DATA: RENAMING AND LABELLING VARIABLES

Renaming variables:

Syntax

```
rename [oldname] [newname]
```

Example

```
rename year Year
```

Labelling variables:

Syntax

```
label variable [varname] ”[label]”
```

Example

```
label variable Year ”Year survey took place”
```

MANIPULATING DATA: KEEPING AND DROPPING VARIABLES

- The original dataset may contain variables you are **not interested in** or **observations you don't want to analyse**
- It's a good idea to get rid of these first - that way, they won't use up valuable memory and these data will not inadvertently be included into your statistical analysis
- This can be done using either:
  - `keep`
  - `drop`

Syntax (applies to drop also)

```
keep [varlist]
```

Example

```
keep id age year health
```

MANIPULATING DATA: KEEPING AND DROPPING VARIABLES

- The previous slide shows the most straightforward implementation of the keep and drop commands, however we can make these commands more complex by using what are called **relational** and **logical operators**
- This is a better alternative to piecemeal edits to the data, as **you can set particular conditions to which data is kept or dropped** given your own research ideals
- While you may not use all of them, the next slide lists all the available operators that **you can use with commands in Stata that have the option of setting a conditional**

RELATIONAL AND LOGICAL OPERATORS

Relational

- ==

  equal to

- !=

  not equal to

- >

  greater than

- >=

  greater than or equal to

- <

  less than

- <=

  less than or equal to

Logical

- &

  and

-

  | or

- ~

  not

- !

  not

MANIPULATING DATA: USING CONDITIONALS

It may seem like it's starting to get complicated now but this is a really small addition

For example, the general syntax for keep is as follows:

```
keep [varlist] if [condition]
```

This addition at the end of the required syntax is applicable to **most** Stata commands

A good way to check is to type

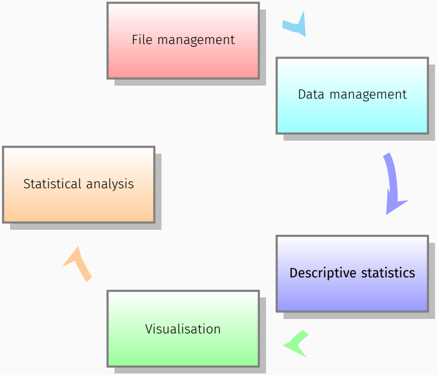
```
help [command]
```

This will bring up a help file for a given command (and is very useful!)

DEMO: USING CONDITIONALS

DESCRIPTIVE STATISTICS

STATA WORKFLOW: DESCRIPTIVE STATISTICS



File management

Data management

Descriptive statistics

Visualisation

Statistical analysis

DESCRIPTIVE STATISTICS

It may seem like we've go though a lot to get to this point but again, thorough procedure pays off

Up to this point, we should be able to:

- Set working directories for new projects
- Import various datafiles
- Manipulate data and generate new variables

Only now should we proceed with looking at and analysing the data, **not before!**

DESCRIPTIVE STATISTICS

So, what do we mean by **descriptive statistics**?

These include but are not limited to:

- Sample size
    - Overall
    - Sub-group
- Outcome frequencies
- Measures of central tendency
    - Mean
    - Median
    - Mode
- Measures of variability
    - Standard deviation
    - Variance
- Variable relation
    - Correlation

DESCRIPTIVE STATISTICS

Some of you (hypothetically) may ask, ”What’s the point in these descriptive statistics?”

Short answer: **A LOT!**

It can tell us:

- The dimensions of the dataset
- What the distribution is of each variable
- How variables are related with one another and to what degree

The reason that these are important before modelling is because a lot about what we want to know about the data between groups or within groups are expressed here!

DESCRIPTIVE STATISTICS

These properties of the data are attained using a few key commands in Stata:

- `describe`

  ”describes” the data
- `summarize`

  summarises the data
- `tabulate`

  tabulates a/the variable(s)
- `correlate`

  creates a correlation matrix of specified variables

DESCRIPTIVE STATISTICS: DESCRIBE

```
describe [varlist], [options]
```

This reports some basic information about the dataset and its variables (size, number of variables and observations, storage types of variables etc.

Notice, there is the ability to add an option to the end of the command - These are option but may be useful

- `simple`
  display only variable names
- `short`
  display only general information
- `fullnames`
  do not abbreviate variable names
- `numbers`
  display variable number along with name

DESCRIPTIVE STATISTICS: SUMMARIZE

```
summarize [varlist] if [condition], [options]
```

summarize calculates and displays a variety of univariate summary statistics. If no varlist is specified, summary statistics are calculated for all the variables in the dataset

The main option that can be used here is detail, which will produce additional summary statistics, such as third and fourth-order moments - skewness and kurtosis

Notice, the **conditional option** is available here, which **allows for summary statistics for a sub-sample of the data**

DESCRIPTIVE STATISTICS: TABULATE

The general syntax for tabulate is as follows:

```
tabulate [varlist] if [condition], [options]
```

tabulate produces a one(two)-way table of frequency counts

Again, there is the option to **tabulate the data on a condition** that you specify about the data

There are options here but they are used infrequently, so the standard syntax is usually enough

DESCRIPTIVE STATISTICS: CORRELATE

```
correlate [varlist] if [condition], [options]
```

The correlate command displays the correlation matrix or covariance matrix for a group of variables. If varlist is not specified, the matrix is displayed for all variables in the dataset.

Again, there is the option to **look at correlations between variables on a condition** that you specify about the data

Moreover, like tabulate, there are options here but they are used infrequently, so the standard syntax is usually enough

DEMO: DESCRIPTIVE STATISTICS

Questions?