

# DATA MANAGEMENT AND INTRODUCTION TO STATA

---

David Clark

Thursday 12th April

University of Leeds

# INTRODUCTIONS AND PREAMBLE

---

## 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)**

# WHO IS THIS COURSE FOR?

For those who want to:

- **Organise and manage data**
  - Generating, reshaping, dropping and recoding
- **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)

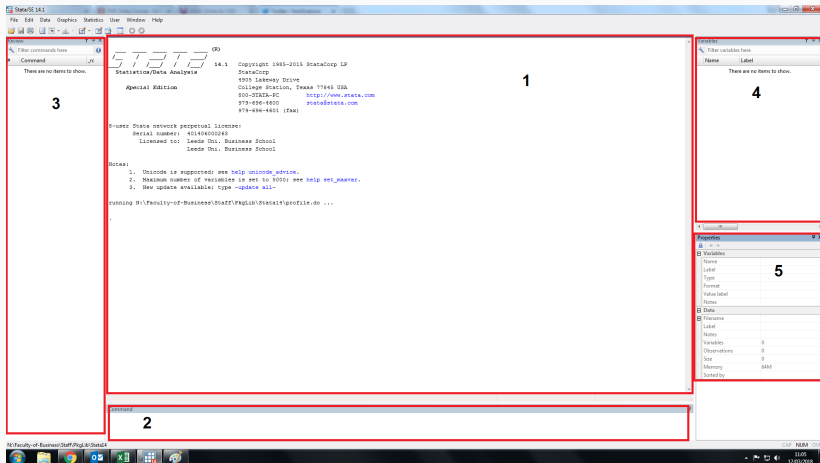
## WHY NOT USE X?

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

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



# STATA 14 FRONT END GRAPHIC USER INTERFACE (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.

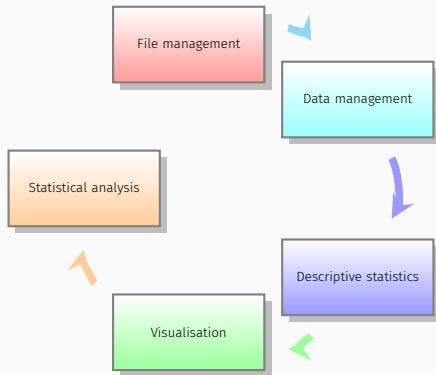
The **review window** (3), lists all the operations preformed since opening Stata. If you click on one of your past commands, you will see the command being displayed in the Command window and you can re-run it by hitting the enter key.

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

---

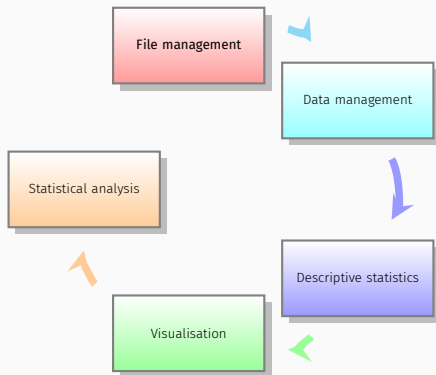


- File management
  - Data management
  - Descriptive statistics
  - Visualisation
  - Statistical analysis
- } These two stages will consume the **most time** in any research project

# FILE MANAGEMENT

---

# STATA WORKFLOW: FILE MANAGEMENT





- 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

- 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

## 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

### 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

- 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

- 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

### Using the drop down menus in the GUI

- If you follow this menu path:

File → 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.

### 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**

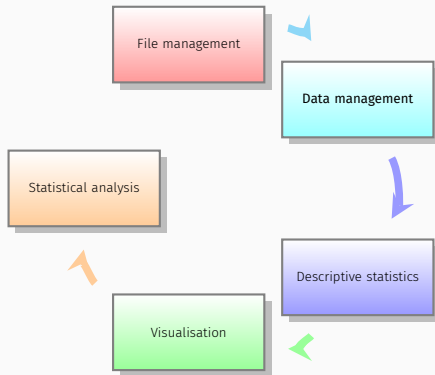
# DEMO: CHANGING AND CREATING DIRECTORIES

---

# DATA MANAGEMENT

---

# STATA WORKFLOW: DATA MANAGEMENT



- 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**

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 be 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

### 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 Stata is to check the **variables window**
- If you see variables names in there, **you've successfully opened your Stata datafile**



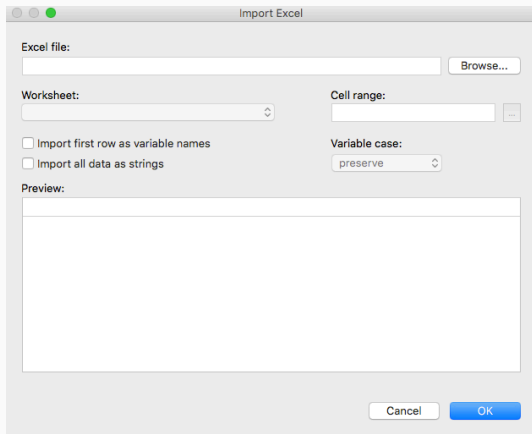
### Excel datafiles (.xls/x files)

- 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 straightforward 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:



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

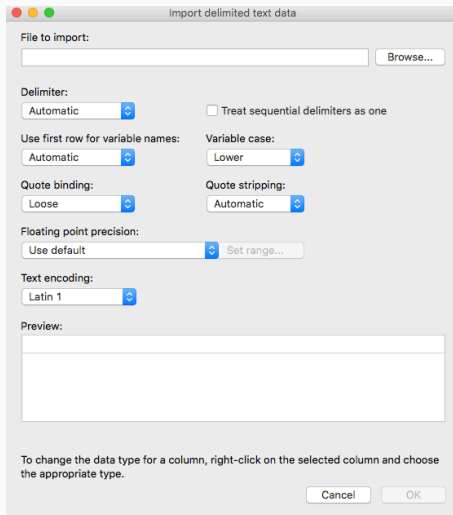
### Comma-Separate Values datafiles (.csv files)

- Other than Excel datafile types, another popular datafile format is the comma-separate 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:



The screenshot shows a dialog box titled "Import delimited text data". It contains several configuration options for importing a file:

- File to import:** A text input field with a "Browse..." button to its right.
- Delimiter:** A dropdown menu set to "Automatic". To its right is a checkbox labeled "Treat sequential delimiters as one" which is currently unchecked.
- Use first row for variable names:** A dropdown menu set to "Automatic".
- Variable case:** A dropdown menu set to "Lower".
- Quote binding:** A dropdown menu set to "Loose".
- Quote stripping:** A dropdown menu set to "Automatic".
- Floating point precision:** A dropdown menu set to "Use default" with a "Set range..." button to its right.
- Text encoding:** A dropdown menu set to "Latin 1".
- Preview:** A large empty text area for previewing the data.

At the bottom of the dialog, there is a note: "To change the data type for a column, right-click on the selected column and choose the appropriate type." Below this note are "Cancel" and "OK" buttons.

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

---



# CONCLUSION

---



QUESTIONS?