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)**

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:

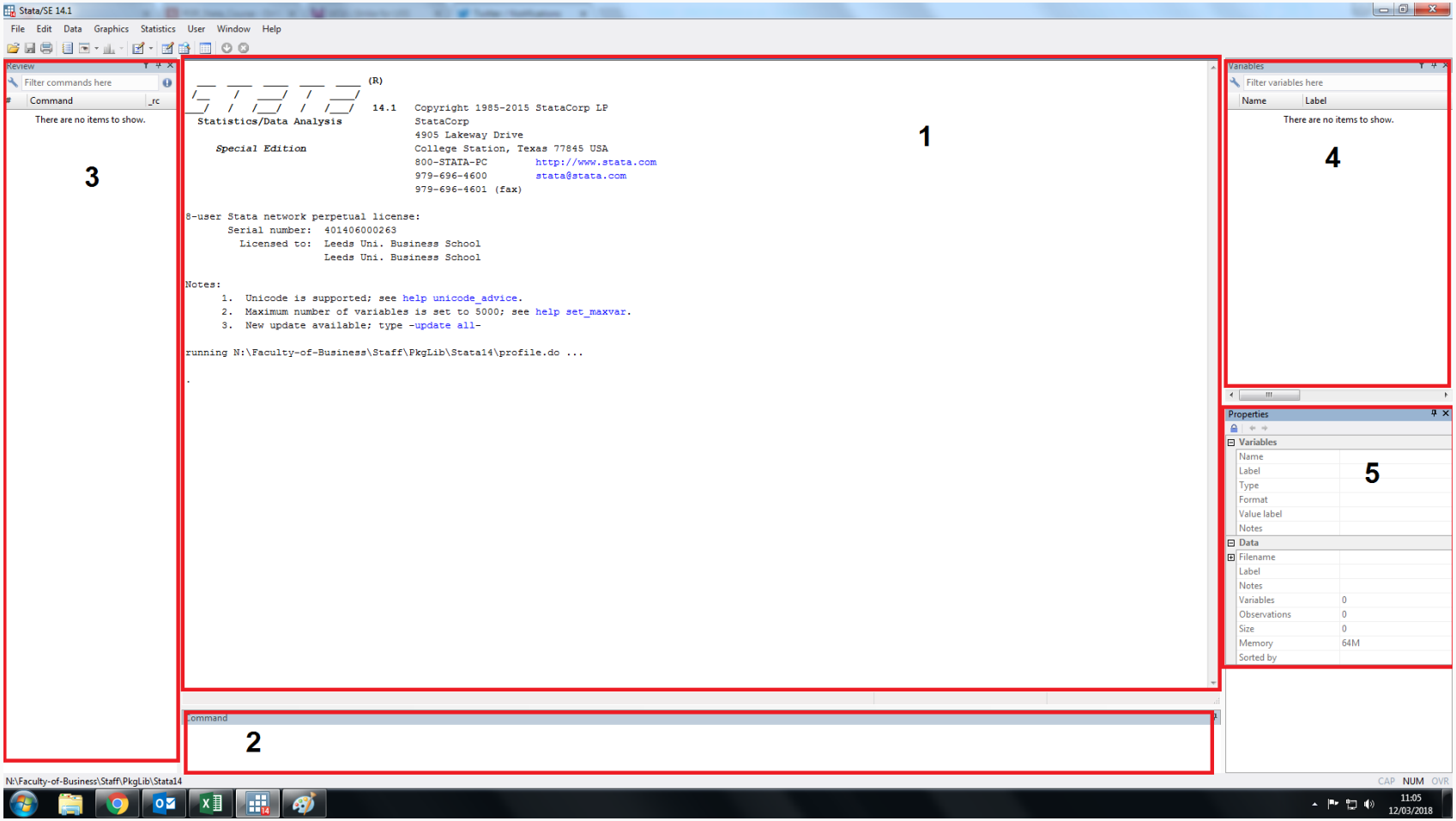
- IC (standard)
- SE (extended)
- MP (multiprocessing)

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 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

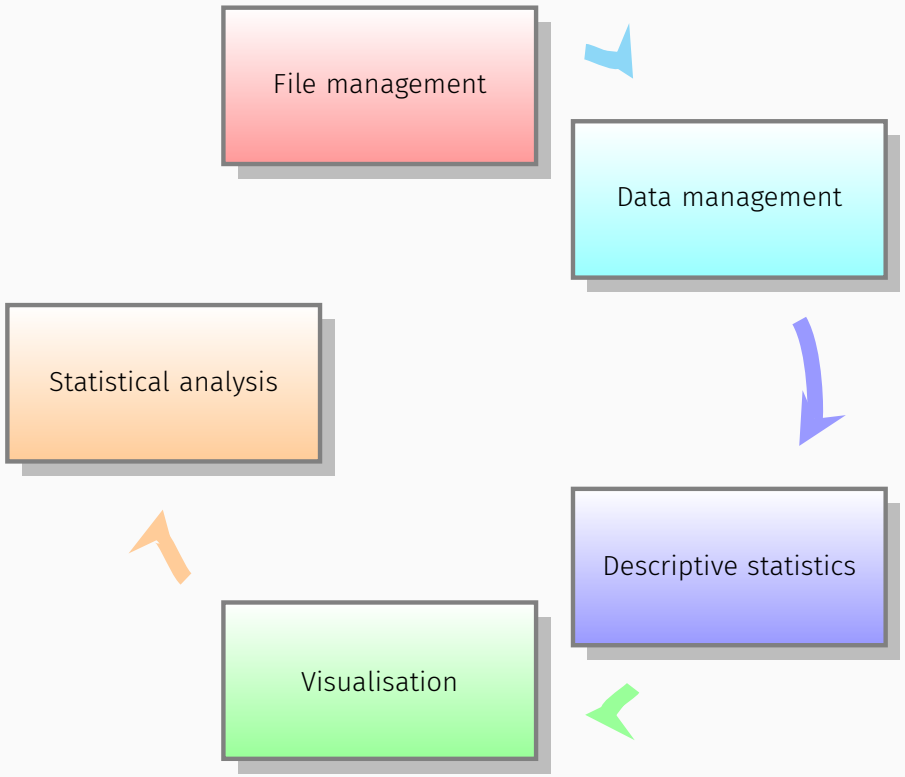
The review window (3), lists all the operations preformed since opening Stata. If you click on one of your past commands, you will see the command being displayed in the Command window and you can re-run it by hitting the enter key.

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



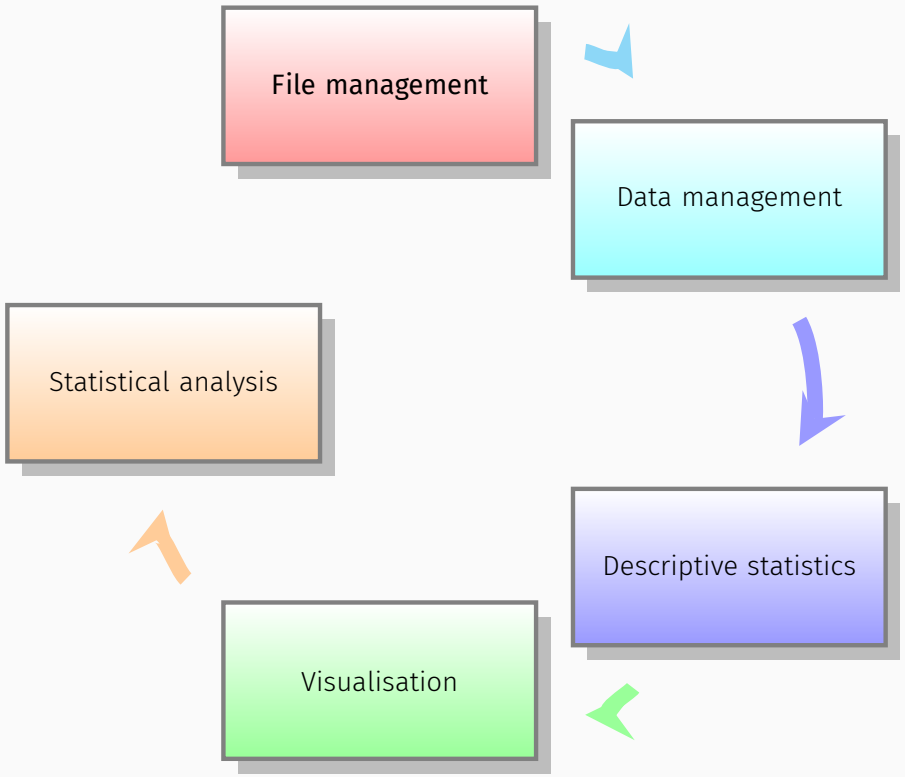
STATA WORKFLOW

- 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

FILE MANAGEMENT

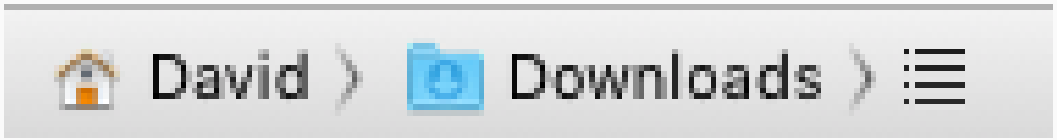
- What do we mean by **file management**?
  - Typically, when people (most) begin using Stata, they will just open some data and **do stuff**
- Questions that arise:
  - **Where is the data stored?**
  - **Where is the output stored?**
  - **Where is Stata currently working from?**
  - **Are we utilising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

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

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 → 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`

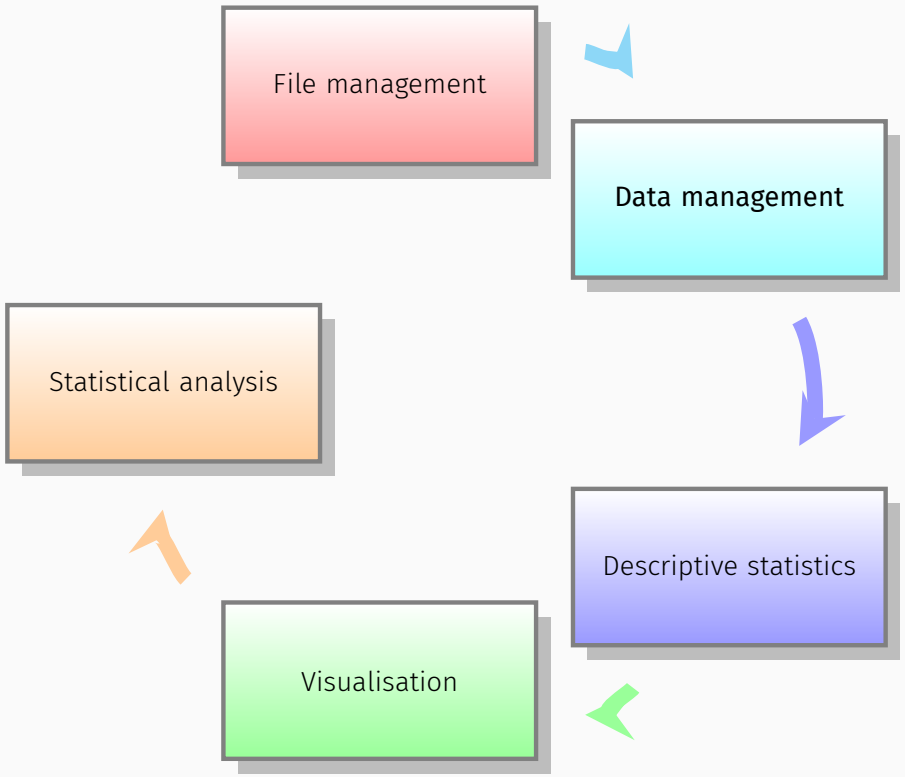
  Tells Stata to **create a new folder in this directory**
- `"/Users/David/Documents/Projects/Thesis_Paper_Two"`

  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**

```
The theme provides sensible defaults to \emph{emphasis}
text, \alert{accent} parts or show \textbf{bold} results.
```

becomes

The theme provides sensible defaults to emphasis text, accent parts or show **bold** results.

LISTS

Items

- Milk
- Eggs
- Potatos

Enumerations

1. First,
2. Second and
3. Last.

DESCRIPTIONS

PowerPoint Meeh.

Beamer Yeeeha.

ANIMATION

- This is important

ANIMATION

- This is important
- Now this

ANIMATION

- This is important
- Now this
- And now this

ANIMATION

- This is really important
- Now this
- And now this

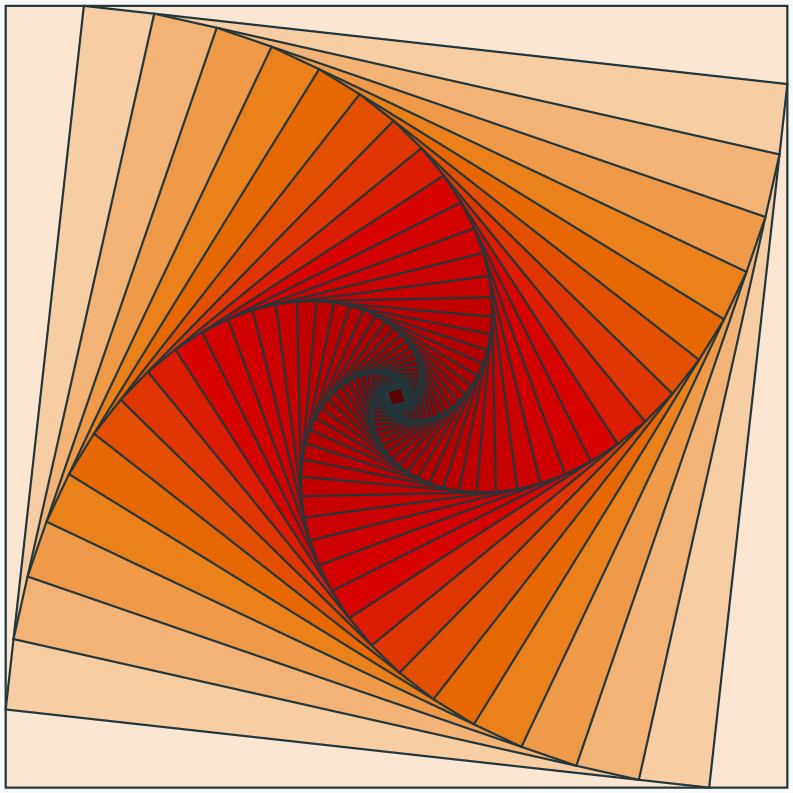


Figure: Rotated square from texample.net.

Table: Largest cities in the world (source: Wikipedia)

City	Population
Mexico City	20,116,842
Shanghai	19,210,000
Peking	15,796,450
Istanbul	14,160,467

BLOCKS

This is a block title

This is soothing.

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

QUOTES

Veni, Vidi, Vici

DARK BACKGROUND

CONCLUSION

SUMMARY

Get the source of this theme and the demo presentation from

`github.com/matze/mtheme`

Questions?