University of Exeter

Applied Econometrics

Topic 1: Introduction

Jingya Zeng

University of Exeter

Outline

1. Overview of the Module
    - Aim of the module
    - Why Stata?
    - Format of the course
    - Course outline for the term
2. Introduction to Stata
    - Stata environment
3. Practical example 1 in Stata
    - Empirical analysis
    - Data in Stata
4. Practical example 2 in Stata
    - Stata Basics
    - Changing the data
    - Subscript and sort

University of Exeter

Overview of the Module

University of Exeter

Target audience & aim of the module

- Students familiar with intermediate statistics and econometrics on a theoretical level, and who want to learn how to perform statistical and econometric analysis.
- The course will introduce you to code into a professional statistics software.
- The course will use applications from various fields of economics, with a particular focus on labour economics.

University of Exeter

Skills developped in the course 1/2

Research skills: Translating a research question into data analysis tasks + assigning the correct procedure to the tasks + intepreting the results.

Empirical analysis skills: Conducting a well-structured and reproducible analysis + being able to assess the quality of an empirical analysis.

University of Exeter

Skills developped in the course 2/2

Data skills: Learn to visualize important properties of the data + generate summary statistics + dealing with missing values + uncover mistakes and inconsistencies in the data + manipulate, import and export data as well as results.

Applied econometrics skills: Learn how to perform in practice linear regressions, regressions for limited dependent variables, and panel data analysis + run regression diagnostics.

Technical skills: Learn how to employ the power of professional statistics software, including coding/programming skills.

University of Exeter

Why using Stata for this class?

- Many of you have used Stata?
- Stata is the most used software for econometric analysis
- Command line driven. Not point-and-click oriented
- Good balance between user-friendliness and programmability
- Good customer support, extensive documention, and active user community
- I like Stata (and so do a lot of other people).

University of Exeter

Lectures and seminars

- Lectures 2x1 hours:
  - Monday: 13h35 - 14h25
  - Tuesday: 13h35 - 14h25
- Seminars are held weekly from week 2:
  - Monday: 10h35 - 11h25 starting next week
- Office hours: Wednesdays 3-4 pm, Thursdays 3-4 pm.
- Practical exercises will features in both lectures and seminars.
- Optional: find a programming pair

University of Exeter

Assessment

- 80% of the marks will come from a practical exam (online option 4), to be held during the May examination period.
  - open book
  - three hours
- 20% of the marks will come from a graded homework, due for submission on 11th March.
  - You will have approx. a week to complete the tasks in this homework.

University of Exeter

Topics covered in the course 1/2

Exploratory data analysis: Summary statistics; plotting data; distribution of the data; Kernel density estimation; using logs; exporting tables and results etc.

Linear regression with cross-sectional data: Simple and multiple linear regression; interpretation of output; inference and regression diagnostics; heteroscedasticity.

Functional form: Linear-log, Log-linear, Log-log; polynomials; dummy and factor variables; interaction effects

University of Exeter

Topics covered in the course 2/2

Missing data: Possible causes of missing data; statistical and technical issues arising from missing data; potential remedies for missing data.

Regression models for limited dependent variables: Modelling binary outcomes; estimation of linear and non-linear models; interpretation; marginal effects.

Panel data: Introduction to panel data; descriptive statistics for panel data; estimation and interpretation of linear panel data models.

Time Series: Introduction to time series data; descriptive statistics for time series data; estimation and interpretation of time series data models.

University of Exeter

Reading List

The econometrics theory part is covered in most introductory/intermediate econometrics textbooks, e.g. Wooldridge textbooks (Introductory Econometrics: A Modern Approach, and Econometric Analysis of Cross Section and Panel Data).

Some textbooks focus on the use of Stata in their examples and exercises, e.g. (Microeconometrics Using Stata by Cameron and Trivedi, and Using Stata for Principles of Econometrics by Adkins and Hill).

All relevant material for the practical exam will be presented in the lectures and tutorials. Any additional reading will be made available on ELE.

University of Exeter

Materials on the web

One of the best way to learn is to look for answers on the web. While conducting web searches, you may encounter:

1. Stata discussion forum: https://www.statalist.org/ (since April 2014; pre-2014 searchable archive at https://www.stata.com/statalist/archive/)
2. Stata FAQs are at http://www.stata.com/support/faqs/
3. Stata YouTube channel: https://www.youtube.com/channel/UCVk4G4nEtBS4tLOyHqustDA
4. UCLA Stata site: https://stats.oarc.ucla.edu/stata/
5. Other discussion forums, e.g. https://stackoverflow.com/questions
6. Some Profs or Universities websites, e.g. https://dss.princeton.edu/training/

University of Exeter

Materials on ELE

- Lecture notes, exercises and tutorial problem sets.
- Data files. **Please use the data files from this course only for the purpose of the course. You may not distribute data unless it is explicitly stated that you are allowed to do so. You may not use the data for commercial purposes, or attempt to sell the data.**
- Stata code (in the form of do-files and ado-files). (The copyright of any Stata code distributed in class lies with the authors.)
- Discussion board: feel free to discuss issues like general data analysis questions, Stata problems etc.
- More information regarding the practical exam and the graded homework are to follow soon.

University of Exeter

Introduction to Stata

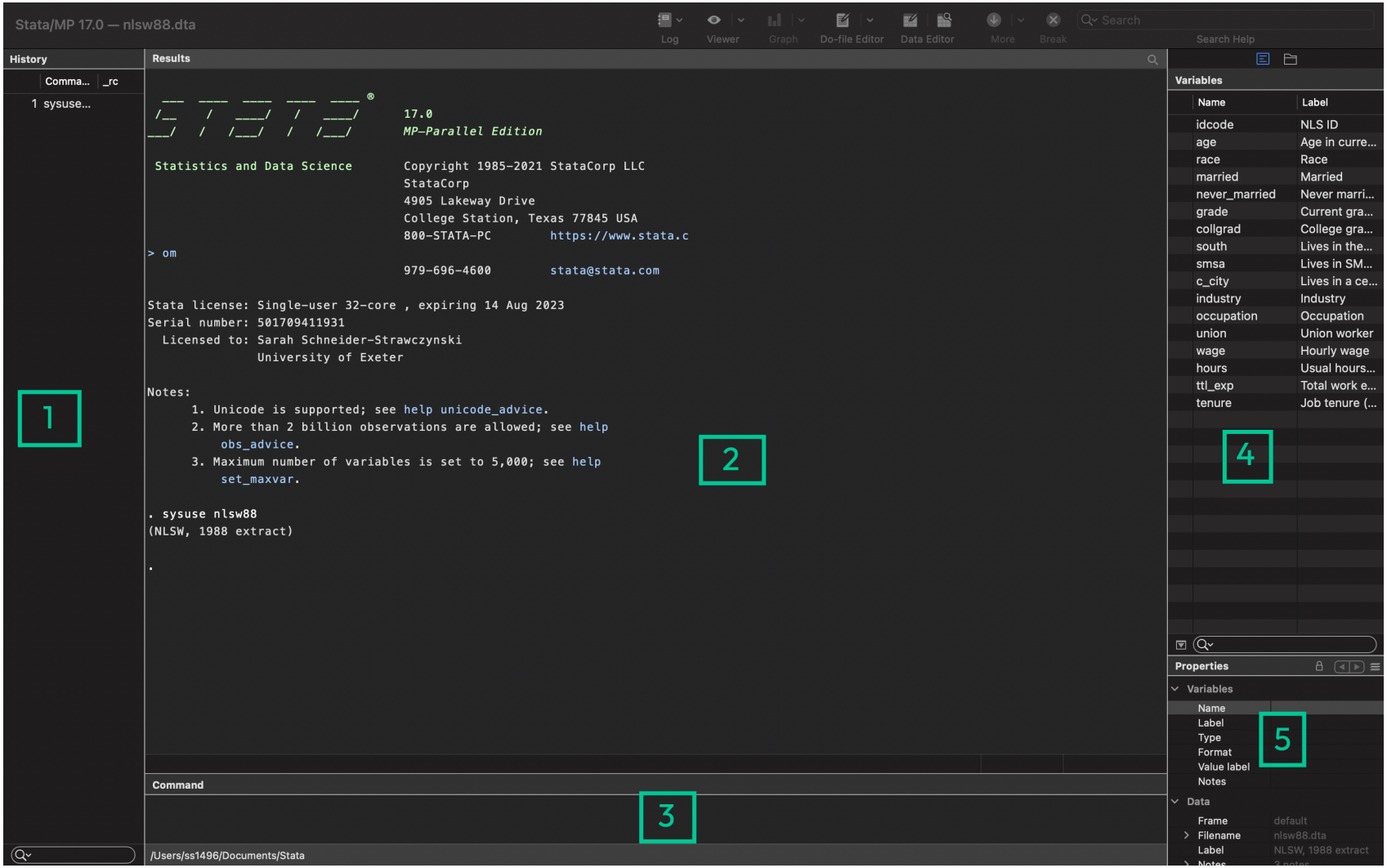
University of Exeter

Features

- One the most used software to apply econometrics techniques in applied economics
- Advantages for statistical analysis:
  - Many built-in statistical procedures (time series, cross sections, panel models, survey analysis etc..)
  - More intuitive for a first statistical analysis coding experience
  - Good graphics and table exports capabilities (publication quality)
- Other software/coding environments: SAS, R, Python. You can also integrate Python and R into Stata.

University of Exeter

Stata Environnement



Stata Environnement

1. **Review window** : show commands that were previously submitted during the open Stata session. A single click brings the command back to the command window. A double click on a command exectues it.
2. **Results window**: all results or messages from the submitted commands will appear in that window.
3. **Command window**: allow to type the command that can be executed directly by pressing
. By hitting the $\boxed{[\text{Page Up}]}$ key, you can recover and edit commands you have previously submitted.

University of Exeter

Stata Environnement

4. **Variables window**: lists all variables in the dataset, alongside with their label.
5. **Properties window**: display information about the dataset (number of observations, number of variables etc...), and information about a variable that has been selected in the Variables window. Also allow to edit some properties of a variable such as labels.

Stata Environnement

Some other windows appear only when a specific action triggers their opening:

- **Viewer** window: appear when the user do a help request or when the user request the visualisation of log files.
- **Do-file Editor** window: appear when the user want to open Stata's code file editor or do-files.
- **Data Browser** window: allow to see the data.
- **Data Editor** window: allow to see and modify the data.
- **Variable Manager** window: all to create and manage variables (names, labels, formats...)
- **Graph Editor** window: allo to edit manually Stata graphs.

University of Exeter

Stata working directory

First, you need to define your working directory (the files in your computer where you are working for this Stata session).

To find out your current working directory, type:

```
pwd
```

Stata working directory

To change your current working directory, type for instance:

Mac example

```
cd "/Users/Jingya/BEE2032/workspace"
```

PC example

```
cd "C:/Users/Jingya/BEE2032/workspace"
```

You need to adapt the file path to your own computer

University of Exeter

Using logs

Create a log file allows to record your Stata session (commands + results), that is everything you see in the results window (including errors).

```
log using lecture1.log, replace
```

You can close the log file: `log close`

If you want to continue adding content to a same log file that was previously closed:

```
log using lecture1.log, append
```

University of Exeter

Using logs

You can also create another log file with only the commands you run:

```
cmdlog using lecture1.do
```

You can close this log file of commands: `cmdlog close`

University of Exeter

How to get help

Online help: type the `help` command that will displat the help fil for the commande. Example: `help tabulate`

Search: in case you don't know the name for a command, or you need help for function : `search`

Links in help files, follow the links to see help files for related entries

Manuals: the manuals contains additional information, along with examples illustrating the use of the commands.

Practical example 1 in Stata

University of Exeter

Is there a college premium in the wages of American women?

Step 1 Formulate the research question.

Step 2 Find appropriate data to help answer research question.

First clear the workspace to be sure you start fresh without already loaded data

```
clear
```

Then load the data nlsw88 that is already available directly through Stata. It contains information on a group of women in their 30s and early 40s to study labour force patterns.

```
sysuse nlsw88
```

University of Exeter

Is there a college premium in the wages of American women?

Step 3 Analyse the effect of being a college graduate on the wage individuals earn.

```
regress wage collgrad
```

You can also use the short version of the command `regress`:

```
reg wage collgrad
```

Step 4 Interpret the findings and report them.

Is there a college premium in the wages of American women?

```
. regress wage collgrad

      Source |       SS           df       MS      Number of obs   =     2,246
-------------+----------------------------------   F(1, 2244)      =    172.44
       Model |  5307.01034         1  5307.01034   Prob > F        =    0.0000
    Residual |  69060.9571     2,244  30.7758276   R-squared       =    0.0714
-------------+----------------------------------   Adj R-squared   =    0.0709
       Total |  74367.9674     2,245  33.1260434   Root MSE        =    5.5476

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    collgrad |    3.615502   .2753268    13.13   0.000      3.07558    4.155424
       _cons |    6.910561   .1339984    51.57   0.000     6.647788    7.173335
------------------------------------------------------------------------------
```

The college premium is estimated to be 3.62 US Dollars.

Data that are shipped with Stata

Stata comes with several data sets. These are very useful to:

- work out the examples in the manuals and online help files;
- quickly demonstrate a Stata command;
- demonstrate a Stata problem you ran into with your own data. If you manage to replicate it with one of the built-in data files, it is easier for someone else to help you.

To see a list of system data files: `sysuse dir`
To load a system data file into memory: `sysuse name`

The most commonly used system data file is **auto.dta** which contains the measurements, price and mileage of 74 cars. Many posts on statalist use the auto data to illustrate a problem or a solution.

Advice: Never ever overwrite system data files. In general, be warry of saving the dataset once you performed some computations on it.

University of Exeter

On notation: reading data into Stata

Apart from datasets that was shipped with Stata:

- DOUBLE CLICK on a file in Stata format ***.dta** will open the dataset in Stata. (Make sure the working directory is what you want it to be.)
- Start the software, change your working directory to where you have stored the Stata data files, and then issue the command `use filename`.

Stata can only have one data set at a time in memory.

- `clear` what in memory to use a new data set.
- To save your data, type `save [filename]`. This will write a file in the current working directory.

Reading data into stata (cont)

- Stata can import data from spreadsheets (various formats, one observation per line).
  - To find the command line you need to use , you can use the click button (`File/Import/...`).
  - But it is more efficient and reproducible to import directly with code.
  - The click button approach won't be possible to use for this class as the evaluation is performed on you writing the code for everything you do.
  - To import text delimited data (i.e. csv files): `import delimited`
  - To import excel spreadsheets: `import excel`
- Data can be entered directly into Stata, or copied and pasted from a spreadsheet. Type `edit` to invoke the data editor.

Practical example 2 in Stata

University of Exeter

Basic data reporting

Reading the intro survey data in ELE into Stata:

`browse` the data in the Data Editor. Note the different colors

Show the basic features of the data: number of observations and number of variables in the result window:

`describe`

University of Exeter

Basic syntax

List all observations and variables in the result window.

```
list
```

To use commands like `list` in a more specific manner, we need to undertsand the basic syntax shared by most Stata commands.

The syntax of `list` :

```
list [varlist] [if] [in] [,options]
```

All components in square brackets are optional (we saw that `list` on its own works fine) and are used to add more specific constraints to the command.

Basic syntax

```
list [varlist] [if] [in] [,options]
```

`[varlist]` can be used to restrict the execution of the command to only one or more variables.

Example

For instance to only list variables related to the predicted scores and gender.

```
list score gender
```

Basic syntax

`list [varlist] [if] [in] [,options]`

`[if]` can be used to restrict the execution of the command only to the observations for which a logical expression is true (= that fulfill a special condition).

Example

For instance, to only list observations for which the predicted score is lower than 50.

`list if score<50`

Logical expressions in Stata may include:

- `>` : greater than
- `>=` : greater than or equal to
- `==` : equal to
- `<` : less than
- `<=` : less than or equal to
- `!=` : not equal to
- `&` : and
- `|` : or

University of Exeter

Basic syntax

`list [varlist] [if] [in] [,options]`

`[in]` can be used to restrict the execution of the command only to the observations that are in a specified range (according to their numbered place in the dataset). Note that `in` refers to the current order of the data.

Example

For instance, to only list observations that are between the first and the 5th observations: `list in 1/5` To list only the observation at the 93rd place: `list in 93`

Basic syntax

```
list [varlist] [if] [in] [,options]
```

Several constraints can be specified together.

Example

For instance, to list variables related to the predicted scores and gender among the participants interest in basic data analysis, we could write:

```
list scores gender if why==”Basic Data analysis”
```

Basic data description

In addition to `describe` and `list`, the following commands are very useful.

- To browse data: `browse`
- To produce summary statistics: `summarize`
- To produce one-way or two-way tabulations: `tabulate`
- To count the number of observations: `count`

Operators in Stata

Logical operators: Stata knows abou the logical operators `&` (and), `|` (or), `!=` (not equal). A synonym for `!` is `~`

Relational operators: Relational operators (`>`, `<`, `>=`, `<=`) work as expected, apart from equality: `==` (double equal)

Arithmetic operators: Arithmetic operators also work as expected: `+` (plus), `-` (minus), `*` (times), `/` (divide). The symbol for power is `^`, e.g. 5^2 is typed in as `5^2`

Calculator: The Stata function `display` can be used as a calculator, e.g. `display log(137)*_pi^2`

Modify the data

To rename a variable in the data: `rename`

Example

```
rename var1 score
rename var2 gender
rename var3 chocolate
```

Modify the data

Knowing the operators, we can start changing the data.

To create new variables in the data: `generate`

To replace the content of an existing variable: `replace`

Task

1. To create a binary variable equal to 1 if the person likes dark chocolate, and 0 otherwise:

```
generate darkchoco= 1 if chocolate=="Dark Chocolate"
```

2. Create a binary variable `female` equal to 1 if it is female, and 0 otherwise

3. Create a variable `score_median` that equals 1 if the predicted score is greater and equal to the median score, and 0 otherwise:

Note: the equality is written with `=` (not a condition).

University of Exeter

Subscripting

Stata allows you to refer to a specfic observation (row) with `[ ]`.

Example

To generate variable that takes the value (for all obs.) of the value of the `darkchoco` variable in observation no. 5:

```
generate newvar=darkchoco[5]
```

The current observation can be referred to as: `[_n]`

The total number of observations can be referred to as: `[_N]`

Example

To generate a unique ID in the data: `generate newid= _n`

Example

To generate a constant holding the total number of observations:

```
generate total= _N
```

Subscripting

Example

Create a new variable representing participants' chocolate preference percentages.

1. generates separate constants holding the number of observations like dark chocolate, milk chocolate, or white chocolate
2. divide it by the total number of observations.

Sort order

In many situations it is important that your data is sorted in a certain order, i. e. when using explicit subscripting. Try the following using the introduction survey data:

- Sorts your data in ascending order by gender
- sorts the data by gender, and within gender, by predicted scores.
- We can specify more variables if necessary.

If the sorting criteria do not uniquely identify each observation, the data are sorted randomly within the remaining subgroups.

Example

```
sort newid
```

```
list newid female darkchoco in 1/5
```

Order and subscripting

Note that subscripting depends on the current sort order of the data.

```
sort newid chocolate
display newid[12]
display newid[_N]
```

```
sort chocolate newid
display newid[12]
display newid[_N]
```

The `by:` prefix

Stata offers an easy way to repeat commands for different subgroups.

To find out the average predicted scores by males and females, we can type:

```
by female: summarize score
```

Note that the data have to be sorted by female for this to work. If they are not, we can either issue a `sort female` prior to the command, or we can modify our command to include the sorting:

```
bysort female: summarize score
```

Combining the `by:` prefix with explicit subscripting is a powerful tool for dealing with panel data. Example of usage in cross section:

```
bysort female (score): generate scorerank = _n
```

```
browse newid female score scorerank
```

University of Exeter

Memory

The size of the data you can have in Stata is limited by the amount of memory that is allocated to Stata, which is limited by the amount of RAM in your computer.

To find out about memory usage:

```
help memory
```

University of Exeter

Basic data description - Task

1. Reading the `pen01aL3` data into Stata, it is a random sample from one wave of an employment panel data set.
2. Create a two-way tabulation of the variables `marr` and `ptime`.
3. Find the average log hourly wage of those living in a large city.
4. Count the number of government employees who use a computer for work.

Basic data description - Task

Try these descriptive graphical commands:

- `histogram lnw`
- `graph box lnw`
- `graph hbox lnw, over(female,total)`
- `spikeplot ed`
- `dotplot lnw, over(female)`

University of Exeter