# Introduction to Sata

or how to run the first estimations and live to tell it

**Lucas van der Velde**

Warsaw School of Economics
FAME|GRAPE

October 2018

# What is Stata?

- Stata is a simple tool to conduct complex statistical analysis.
- First release dates back to 1985!
  - ▶ We go for a more recent version: Stata 15, released in 2017.
- Stata is proprietary software (of StataCorp)
  - ▶ Most commands have been tested + official help channels
  - ▶ Frequent updates...
  - ▶ ... but it might take longer to incorporate + recent advances
    (e.g. machine learning, web scrapping are still underdeveloped)

# Elements of the main screen

1. Command window $\rightarrow$ where we can type commands
2. Results window (main) $\rightarrow$ where commands and results are displayed
3. Review window $\rightarrow$ where a list of all commands is displayed. They can be exeuted from there
4. Variables window $\rightarrow$ where variables are displayed.

$\Rightarrow$ How to personalize?

# Opening our first database

- Point & Click: File\Example databases ... → Auto
- Command line: `sysuse auto`

To open the browser

- Point & Click: data\Data Editor\ Data Editor (browse)
- Command line: `browse`

To obtain help : type `help`

# Data types

- Numerical: just numbers (e.g. price)
  - ▶ Different level of precision (byte to Double)
  - ▶ Most operations in Stata require numerical variables

- String: It incorporates letters (e.g. names, models)
  - ▶ We create and refer to them using " " (quotation marks)
  - ▶ Distinct set of operations

- Numerical +: Numerical variable with labels
  - ▶ Label "masks" true value
  - ▶ Facilitates reading and working with data
  - ▶ Often categorical data (e.g. foreign)

# The Do-file

- Do files are routines that we write and can be executed

- Do-files allow replicating results
    $\rightarrow$ Mistakes are less costly. (no ctrl+Z in Stata)

- It is possible to add comments to clarify the code
    - ▶ * At the beggining of the line
    - ▶ // Anywhere
    - ▶ /* ... */ anywhere, only comments the ...

## Describing the database

**Database characteristics**

- describe
- codebook
- labelbook

**Descriptive statistics**

- summarize *varnames*
- tabulate *varnames* [max. 2]
- tabstat *varnames* , by (*varname*)

## Data manipulation

**Modifying the database**

- gen *varname* = 2+3 , $x^2$, etc.
- egen *varnames* = f(*varname*)
- keep, drop

**Conditional statements**

- if, in, inlist, inrange
- boolean operators  ==, != , >,< , &, |

# A Probit model

**Question** Can we distinguish the origin (foreign or american) of the car based on cars' characteristics?

probit *dep.var  indep.vars*  , *options*

# A Probit model

**Question** Can we distinguish the origin (foreign or american) of the car based on cars' characteristics?

probit *dep.var indep.vars* , *options*

In our example
probit foreign price mpg weight length

# Saving and printing results from estimation

**Storing results from 2 models**

```
probit foreign price mpg weight length
estimates store prob1

probit foreign price mpg weight length rep78 headroom
estimates store prob2
```

# Saving and printing results from estimation

**Storing results from 2 models**

```
probit foreign price mpg weight length
estimates store prob1

probit foreign price mpg weight length rep78 headroom
estimates store prob2
```

**Print results from 2 models**

```
estimates store prob1 prob2
* or
outreg2 [prob1 prob2] using myprobs, excel replace
* which saves output in an excel file (xml)
```

# Thank you for your attention

**Lucas van der Velde**

Contact:    lvandervelde@grape.org.pl
            lvelde@sgh.waw.edu.pl