

# Applied Econometrics with STATA

## Introduction to Data Management

Assistant Professor Alexandros Bechlioulis

Department of Banking and Financial Management,  
University of Piraeus

MSc in Bio-economics



# In brief

- General information
  - File types
  - Variable types
  - Dataset types
- Variable management
  - Basic commands
  - Descriptive statistics
  - Data reporting
- Dataset management
  - Import-Merge-Append datasets
  - Reshape data (long - wide data forms)
  - Collapse (dataset of summary statistics)

# File types

Information may include data, commands, comments and output.

- .dta files
  - It is the data directory.
- .do files
  - It contains STATA commands and comments.
- .smcl files (log files)
  - It is a STATA output file.

# Variable types

Variables may include numbers, letters or other symbols.

- String variables
  - It usually contains not only numbers but also characters.
  - It is a non-measurable variable.
- Numeric variables
  - int or long  
They include integer values.
  - float or double  
They include real values with 8 or 16 digits.

Can we display variable types?

- Use the command "describe".

# Dataset types

A data collection may be used to analyze the evolution of a variable or the relationship between two variables across different objects.

- Time-series
  - It is a data collection obtained by measurements over time.
  - Command "tsset time"
- Panel - Longitudinal
  - They are multi-dimensional data involving measurements over time (firms, sectors, countries).
  - Command "xtset id time"
- Cross-sectional
  - This type includes data collected by observing many objects in a specific time period.

# Load and save data

- "use filename, clear"

To open a dataset, we use the command "use".

- If the dataset is in the current working directory, we give only the filename.
- If the dataset is NOT in the current working directory, we should type the full pathname.
- If the path contains blanks, you should use " " to indicate it.

- "save filename, replace"

The "replace" option alters the old file with the same name.

# Basic commands (1)

- "gen" or "generate"
  - It creates a new variable.
  - Command: `gen var1=`
- "rename"
  - It changes a variable's name.
  - Command: `rename old_name new_name`
- "label"
  - It alters a variable's description.
  - Command: `label variable var_name "description"`
- "order"
  - It puts in a specific order the variables of our dataset.
  - Command: `order var1 var2 var3`
- "sort"
  - It classifies the dataset according to the values of a specific variable.
  - `sort var1 var2`

## Basic commands (2)

- "keep" or "drop"
  - It maintains or deletes a number of variables.
  - Command: "keep var1 var2 var3"
- "destring" or "tostring" or "encode"
  - It converts non-numeric variables to numeric (destring).
  - It converts numeric variables to non-numeric (tostring).
  - It converts non-numeric variables to numeric. We do not use this command to variables that contain only numbers (encode).
  - Command: encode var1, gen(var2)



## Basic commands (3) - Metrics

- "Mean"
  - We obtain the mean value of variable.
  - Command: `egen m_var1=mean(var1)`
- "Median"
  - We obtain the median value of variable.
  - Command: `egen med_var1=median(var1)`
- "Min" or "Max"
  - It creates the minimum and the maximum value of a variable.
  - Command: `egen min_var1=min(var1)`
  - Command: `egen max_var1=max(var1)`
- "Summarize"
  - It produces the cumulative sum of the values of the variable up to the current value.
  - Command: `gen sum_var1=sum(var1)`
- "Total"
  - It produces the total sum of all observations of a variable.
  - Command: `egen tot_var1=total(var1)`

## Basic commands (4) - Group identifiers & Operators

### Group identifiers

- `"_n"`
  - It creates a new variable with consecutive numbers, one for each observation.
  - Command: `gen k1=_n`
- `"_N"`
  - It creates a new variable that summarizes all observations.
  - Command: `gen k1=_N`

### Operators

- `"L." or "F."`
  - We obtain a new variable that includes past or future values.
  - Command: `gen lvar1=L.var1`
  - Command: `gen fvar1=F.var1`
- `"D."`
  - We obtain a new variable that includes the first difference between current and previous period's values.
  - Command: `gen dvar1=D.var1`

## Basic commands (5) - Dummy variables / if / symbols

### Dummy variables

**Definition:** In regression analysis, dummy variable takes values either "0" or "1" to show the presence or not of a categorical effect.

- Command: `ta year, gen(yr)`

### "if"

- We use "if" when we need to specify certain periods of time or groups.

### "Symbols"

- ">": more than, "<": less than
- "=": equal to
- ">=": more than or equal, "<=": less than or equal
- "!=" or "~=": not equal
- "&": and, "|": or

# Descriptive statistics - Data reporting

"sum" or "sum...,detail"

- It displays important information and variables' metrics.

"tabulate"

- It presents frequency values.

"codebook"

- It shows basic information of variables.

"list"

It explores specific observations, variables with common letters :

- **Command:** list var1 var2 in 1/15 (first 15 observations).
- **Command:** list k\* in 1/15 (variable names starting with "k")
- **Command:** list \*s in 1/15 (variable names ending with "s")

# Import - Merge - Append

- "Import"

- To insert data in STATA from a different source (xls, csv), we use this command.
- Command: `import excel file, sheet("capm1") firstrow`

- "Merge"

- It adds variables to the dataset. We always "merge" two datasets.
- It requires that both datasets have at least one common variable (key variable), with the same name.
- It depends on the dataset type (time-series, panel) to use one or more key variables.
- Command: `merge m:m key_variable using file.`

- "Append"

- It adds new observations to the dataset.
- It requires that both datasets have the same variables with same names.
- Command: `append using file.`

# Reshape - Collapse

- "Reshape"

- In panel-data analysis, the appropriate format of the STATA dataset is the following: id, time, variable (long form).
- Many databases use the following format: id, variable, time (wide form).
- To convert datasets from wide to long, all time-variables should have common e.g., initial letters (yr).
- Command: `reshape long yr, i(id) j(year)`

## Reshape notes

- "Collapse"

- Many times you have e.g., monthly data but you need yearly data or firm-level data but you need sector-level data. To do that, you just use "collapse".
- Command: `collapse var1 var2, by(year)`

# References

- Princeton University Notes.  
[https : //dss.princeton.edu/online\\_help/stats\\_packages/stata/](https://dss.princeton.edu/online_help/stats_packages/stata/)