

Using Stata

Applied Econometrics

What is Stata?

- It is a multi-purpose statistical package to help you explore, summarize and analyze datasets.
- A dataset is a collection of several pieces of information called variables (usually arranged by columns). A variable can have one or several values (information for one or several cases).

ID	Age	Year	Course
Person 1	18	Freshman	Econ
Person 2	21	Senior	English
Person 3	22	Senior	Econ

Other Programs

- Stata is most used statistical software on college campuses
- Other programs are available that are useful for specific functions

Features	Stata	SPSS	SAS	R
Learning curve	Steep/gradual	Gradual/flat	Pretty steep	Pretty steep
User interface	Programming/point-and-click	Mostly point-and-click	Programming	Programming
Data manipulation	Very strong	Moderate	Very strong	Very strong
Data analysis	Powerful	Powerful	Powerful/versatile	Powerful/versatile
Graphics	Very good	Very good	Good	Excellent
Cost	Affordable (perpetual licenses, renew only when upgrade)	Expensive (but not need to renew until upgrade, long term licenses)	Expensive (yearly renewal)	Open source (free)

What does Stata look like?

Stata 12/13+ screen

The screenshot shows the Stata 12/13+ interface with several windows open:

- Results window:** Shows the command history and log output. A red arrow points up from the bottom-left towards this window.
- Variables window:** Shows a list of variables in the dataset. A red arrow points down from the top-right towards this window.
- Properties window:** Shows detailed properties for selected variables, such as Name, Label, Type, Format, Value Label, Notes, Data, and Filename.
- Command window:** Shows the current command being typed: "log on (text)". A red arrow points left from the bottom-left towards this window.
- Status bar:** Shows the CAP NUM OVR indicator and the F1/F2/DSS/OTR key.

Annotations:

- History of commands, this window**: Points to the Results window.
- Files will be saved here**: Points to the Command window.
- Write commands here**: Points to the Command window.
- Output here**: Points to the Results window.
- Variables in dataset here**: Points to the Variables window.
- Property of each variable here**: Points to the Properties window.

Variables in dataset (from Variables window):

Variable	Label
Year	Year
CountryName	Country Name
GDPpercapita	GDP per capita, PPP (constant 2005 US\$)
Unemployment_fem	Unemployment, female (%)
Unemployment_mal	Unemployment, male (%)
Unemployment_tot	Unemployment, total (%)
Exportsofgoods	Exports of goods and services
Importsofgoods	Imports of goods and services
polityoriginal	Polity (original)
polity2adj	Polity2 (adjusted)

Properties of Year variable (from Properties window):

Name	Year
Type	str10
Format	%10s
Value Label	
Notes	

Types of Variables

- String versus numeric

Var2 is a string variable even though you see numbers. You can't do any statistical procedure with this variable other than simple frequencies

For var1 a value 2 has the label "Fairly well". It is still a numeric variable

Var3 is a numeric. You can do any statistical procedure with this variable

	var1	var2	var3	var4
1	Fairly well	2	2	Fairly well
2	Very well	1	1	Very well
3	Fairly badly	3	3	Fairly badly
4	Fairly well	2	2	Fairly well
5	Very badly	4	4	Very badly
6	Fairly badly	3	3	Fairly badly
7	Fairly well	2	2	Fairly well

Var4 is clearly a string variable. You can do frequencies and crosstabulations with this but not statistical procedures.

Coding on Stata

- Drop down menus versus do-files
- What do we mean by coding?
 - Telling our computer what to do
- Stata as an Interpreter:
 - reads the source code of the program as written by programmer (us)
 - parses the source code
 - Interprets instructions on the fly
- Sequence of statements that have been crafted to do something

Tutorial Follow Along

1. Using do-files
2. Uploading a file: different formats, using drop-down and using code
3. Getting to know your dataset: number of variables, number of cases, format of variables, browser/editor
4. Looking at specific variables: descriptive statistics
5. Commands: understanding if, and, or, not, etc.
6. Cross-tabs – two categorical variables at once
7. Renaming variables and labels
8. Generate new variables
9. Cleaning your data: Sorting, deleting, keeping variables
10. Simple visualizations: histograms and bar graphs
11. Keeping a log