Introduction to Econometrics

Robert Asiimwe

MSc. Agricultural and Applied Economics (MUK)

Course objectives

- Provide theoretical and practical knowledge of econometric techniques and their applications to real world problems
- Facilitate practical estimation of some common econometric models in Stata
- Provide a foundation for skills required to perform advanced economic analysis
- Provide practical skills in Stata data analysis software
- Provide practical socio-economic data management skills

Motivation … (1)

Why do we do research?

- We want to find answers to a question in a formal and structured way
- We have our own question and we think answering these questions would benefit others
- We then write research papers to convince others that we found the answers to these questions

But we also need to show them / describe how we found the answers to these questions

Research can then be done theoretically or empirical or both

1. Theoretical
   - make reasonable assumptions about economic agents involved and their circumstances
   - Using these and considering any restrictions or incentives , the research can predict theoretically the economic outcome
   -

2. Empirical
   - depends on the research setup, it could be observational or experimental study
   - Observational studies; the researcher observes and draws conclusions on the interventions
   - Experimental; the researcher designs an experiment to actively intervene in data production

3. Combine both theoretical and empirical research
   - The researcher builds on specific assumptions to predict the outcome
   - Then uses the data to test whether their predictions holds true in reality

This and the following insights will heavily lean on empirical research

Motivation …(2)

Steps in writing an empirical paper

1. Identify an interesting and specific research question (information gap): should be formulated and motivated very clearly

2. Learn about the key variables, how they are expected to be related to one another (theoretical background)

3. Explore published research
   - ✓ Familiarize with existing knowledge about
     - ✓ Available data and the gaps
     - ✓ Empirical techniques and their limitations
     - ✓ How key variables have been measured (and critiques or appraisals)
     - ✓ Potential pitfalls while studying related areas

   Establish what we know and what we do not know

4. Collect the data, analyze and report your results in a structured way
   - ✓ At analysis stage is where econometrics comes into the picture

   Note: Not every socio-economic question will be answered with econometrics tools

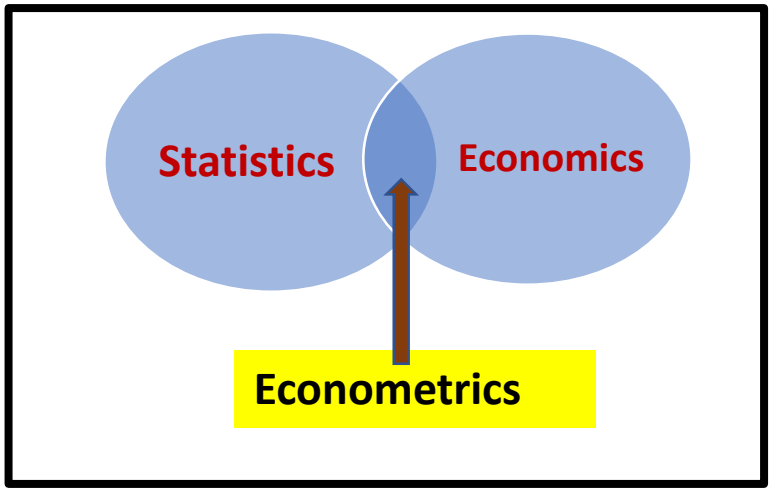
5. Conclusions and policy implications

Example

- Has promotion of (re)-afforestation through provision of free tree seedlings increased farmer willingness to pay for tree seedlings?
- What does theory tell us?
  - Theory of planned behavior
  - From the environmental behavior change campaigns
- What do we expect?
  - Exposure to the benefits of trees should increase WTP for tree seedlings
- Review of literature
  - Measurement issues; how WTP is measured
  - How has WTP been modelled in the past (technology adoption/choice models, contingent valuation models, etc.)
  - Was there any prior WTP for seedlings in the study area or anywhere?
- Econometrics; contingent valuation models
- Conclusions and policy recommendations (so what??)

Econometrics ….. definition

- **Econometrics** is the use of statistical methods for:
  - Estimating economic relationships
  - Testing economic theories
  - Forecasting future scenarios
  - Evaluating policies and programs



- Econometrics is **statistics applied to economic data**
- **Econometrics** uses of real world data to understand real world problems

- **Econometrics is a means to and an end, not an end in itself.**

Why econometrics

- Rare for economics and many other areas (without labs!) to have experiments
- Need to use nonexperimental, or observational data, to make inference
- Allow formal economic theory to be tested on real world data
- Allow application of economic (and other related) theories into practice
- Where theory may be ambiguous as to the effects of some policy change – econometrics can be used to define case specific effects
- Facilitate evaluation of social development programs and related subjects (impact assessment)

Econometrics analysis …. Steps

1. Define the research questions (information gap)
2. Specify the economic model or other conceptual/theoretical framework
3. Specify the econometric model (Operationalize #2)
   Detailed difference between 2 and 3 coming shortly
4. Specify Hypothesis to be tested (from #3)
5. Collect, clean and summarize data
6. Estimate the parameters of the econometric model
7. Make forecasts or predictions
8. Using the model for control or policy purposes

Economic vs Econometric model

1. Economic model

- A stated relationship derived from economic theory or other conceptual/theoretical framework

Example Theoretically quantity of beef demanded ($qbeef$) is driven by; price of beef ($pbeef$), price of substitutes and complements ($pother$), income ($Income$) tastes and preferences-proxy(s) of (Z)

Economic model

$$qbeef = f(pbeef, pother, income, Z)$$

1. Econometric model

- An equation relating the dependent variable to a set of independent variables and unobserved disturbances

- Unknown population parameters determine the ceteris paribus effect of each explanatory variable in an econometric model

Econometric model

$$qbeef = \beta_0 + \beta_1 pbeef + \beta_2 pother + \beta_3 income + \beta_4 Z + u$$

Causality in Econometrics

- Most econometric models estimate the direction and/or magnitude of relationship between a dependent variable and a set of independent variables
- Simply establishing a relationship between variables is rarely sufficient to considered a relationship as causal
- If we've truly controlled for enough other variables, then the estimated ceteris peribus effect can often be considered to be causal

Example

- A model of technology adoption implies higher income should lead to higher adoption of improved technologies
- In the simplest case, this implies and equation like

$$adoption_i = \beta_0 + \beta_1 Income_i + \mu_i$$

- The estimated β_1 is the return to Income, but can we say increased income causes technology adoption?
- The error term μ_i includes other factors affecting adoption, we want to control for as much as possible
- But some things are still unobserved which can be problematic

Econometric modeling and model specification

Econometric modeling

- Application of appropriate econometric models to real world data in order to understand real world economic problems

- A model is an abstraction of a phenomena into its simplest and measurable facets

Components of the model

- The equation(s) –variables and functional form
- A priori restrictions on parameters
- Stochastic assumptions (assumptions about the error term)
  - Random nature of human behavior
  - Omitted variables that influence the independent
  - Non-linearity
  - Errors in variable measurement
  - Others

Types/Families of Econometrics models

- Continuous depended variable models
  - Simple bivariate linear regression model
  - Multivariate linear regression model
- Limited dependent variable models
  - Binary response models (LPM, logit, probit)
  - Ordered response models (ordered probit/logit)
  - Multinomial response models (multinomial probit/logit
  - Count data models (Poisson, negative binomial)
  - Censored data models (Tobit)

Introduction to Stata

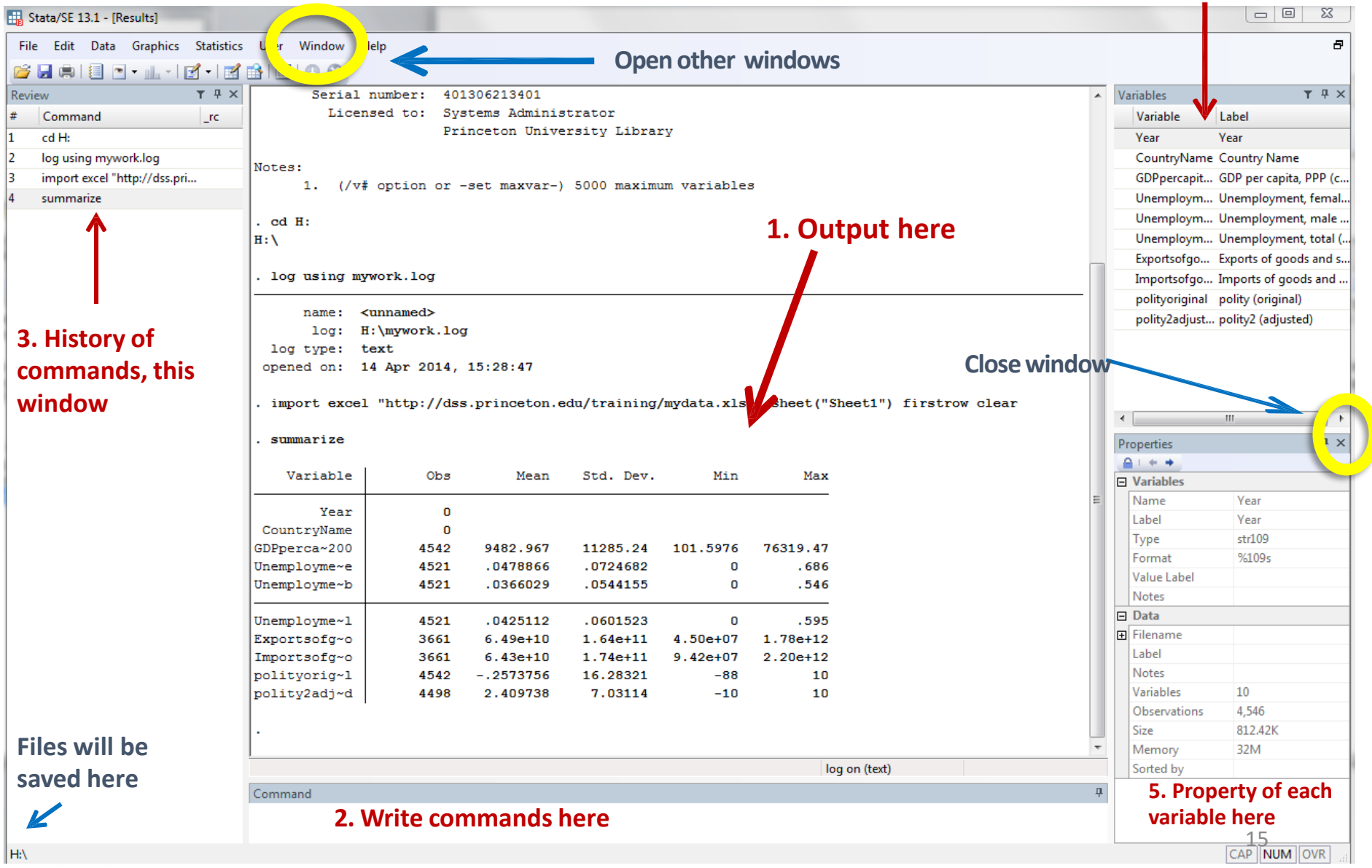
What is Stata?

It is a multi-purpose statistical package to help you explore, summarize and analyze datasets.

Features	SPSS	SAS	Stata	JMP (SAS)	R	Python (Pandas)
Learning curve	Gradual	Pretty steep	Gradual	Gradual	Pretty steep	Steep
User interface	Point-and-click	Programming	Programming/ point-and-click	Point-and-click	Programming	Programming
Data manipulation	Strong	Very strong	Strong	Strong	Very strong	Strong
Data analysis	Very strong	Very strong	Very strong	Strong	Very strong	Strong
Graphics	Good	Good	Very good	Very good	Excellent	Good
Cost	Expensive (perpetual, cost only with new version). Student disc.	Expensive (yearly renewal) Free student version, 2014	Affordable (perpetual, cost only with new version). Student disc.	Expensive (yearly renewal) Student disc.	Open source (free)	Open source (free)
Released	1968	1972	1985	1989	1995	2008

Understanding the Stata Environment

Basic STATA windows. There are 5 basic windows when STATA is started:



Stata organisation and window system

- Basic STATA windows. There are 5 basic windows when STATA is started:

1. The Results Window

- Show commands executed and results of commands
- Shows commands executed from the MENU
- Shows logs (errors, warnings etc.)

2. The Command Window

- This is where commands are entered interactively
- Do-files executed here

3. The Review Window

- Shows list of commands executed from the Command Window AND Menu

4. The Variables Window

- Shows variables and properties of variables of the active data set

5. The Properties Window

- Shows details on the composition of the variable selected and the dataset as a whole

There are other STATA windows that are activated for a number of procedures and activities. For example: Graphs, Data editor, Viewer, Variable Manager, and Do-file Editor

Stata setup: Working directory

To see the working directory, type:

```
pwd
```

```
. pwd
C:\Program Files (x86)\Stata
```

To change the working directory to avoid typing the whole path when calling and saving files, type:

```
cd "D:\Econometrics course\Day 1"
```

```
. cd "D:\Econometrics course\Day 1"
D:\Econometrics course\Day 1
```

Use quotes “” if the new directory has spaces

Stata setup: Log file

A ***log file*** is sort of Stata's built-in tape recorder and where you can:
1) retrieve the output of your work and 2) keep a record of your work.

To create a ***log file***, in the command line type:

```
log using mylog.log
```

This will create the file 'mylog.log' in your working directory. You can read it using any word processor (notepad, word, etc.) or Stata.

To close a log file type:

```
log close
```

To add more output to an existing log file add the option append, type

```
log using mylog.log, append
```

To replace a log file add the option replace, type:

```
log using mylog.log, replace
```

Note that the option `replace` will delete the contents of the previous version of the log.

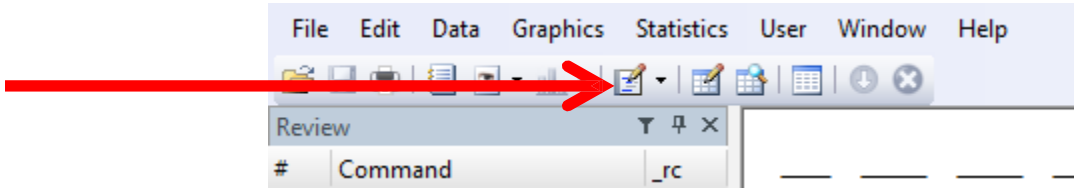
Stata setup: do-file

do-files contain Stata commands to run specific procedures. It is highly recommended to use do-files to store your commands so you do not have to type them again should you need to re-do your work.

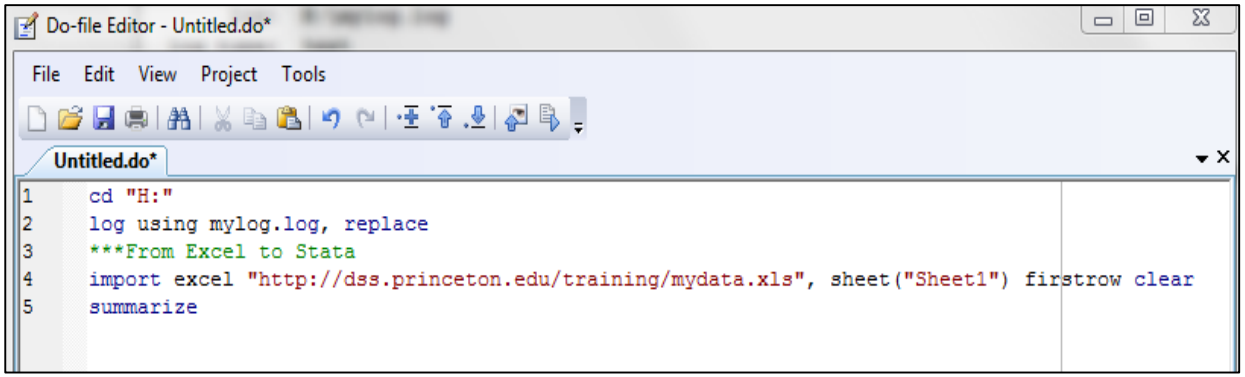
You can use use Stata's 'do-file editor' to write a dofile. To open the dofile editor, in the command window type:

```
doedit
```

Or click on the icon here:



You can write the commands, to run them select the line(s), and click on the last icon in the do-file window



Stata setup: Opening/saving Stata files (*.dta)

To open files already in Stata with extension *dta, run Stata and you can either:

- Go to file->open in the menu or
- Type use "C:\my data\mydatafile.dta"

If your working directory is already set to e.g "C:\my data", just type;

```
use mydatafile
```

To save a data file from Stata

- Go to file->Save as or just type

```
save, replace
```

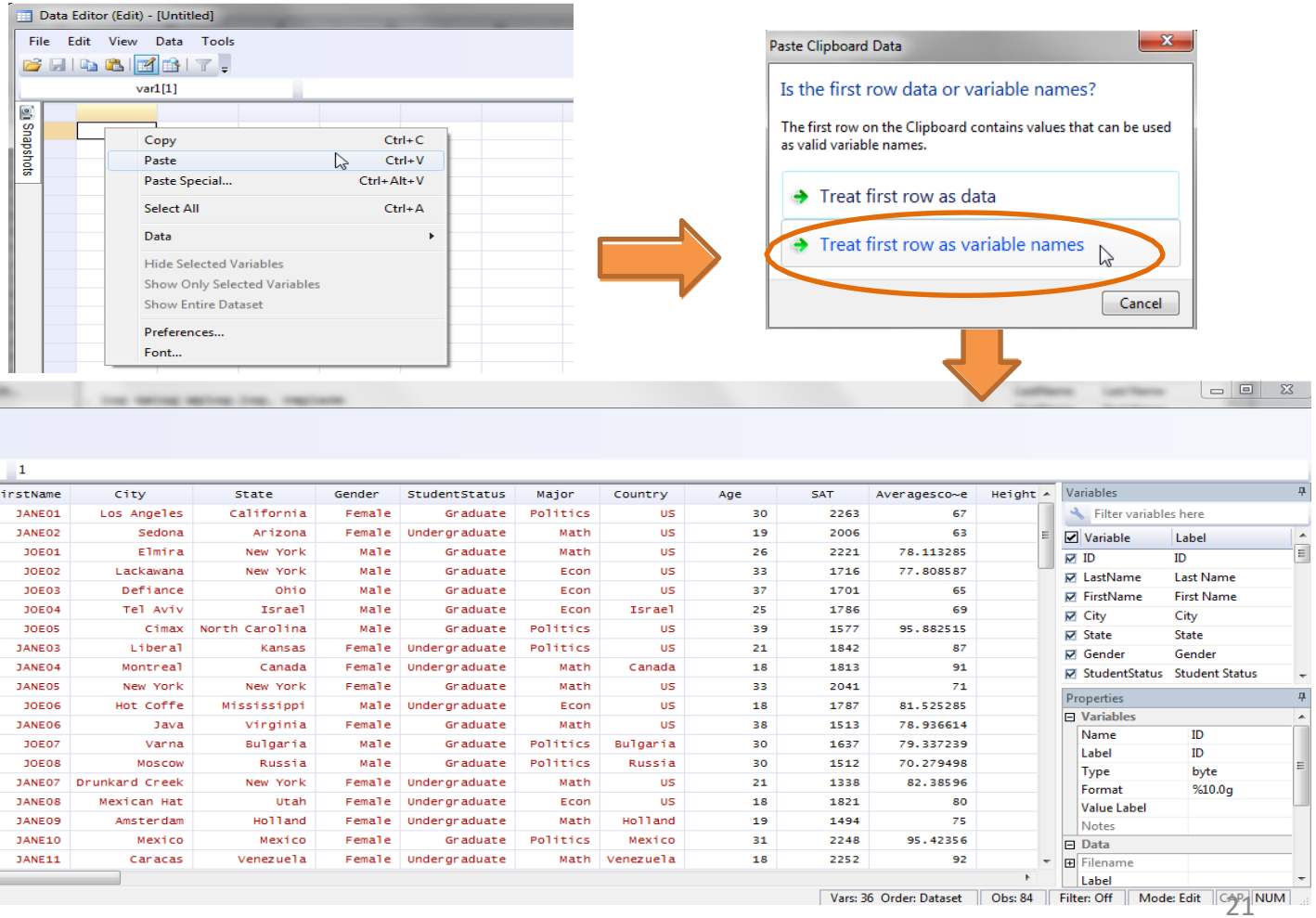
If the dataset is new or just imported from other format'

- Go to file –> save as or just type:

```
save mydatafile /*Pick a name for your file*/
```

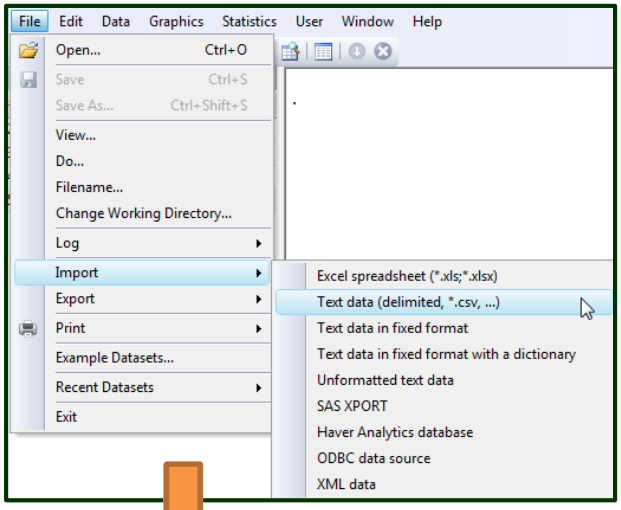
Importing from Excel to Stata using copy-and paste

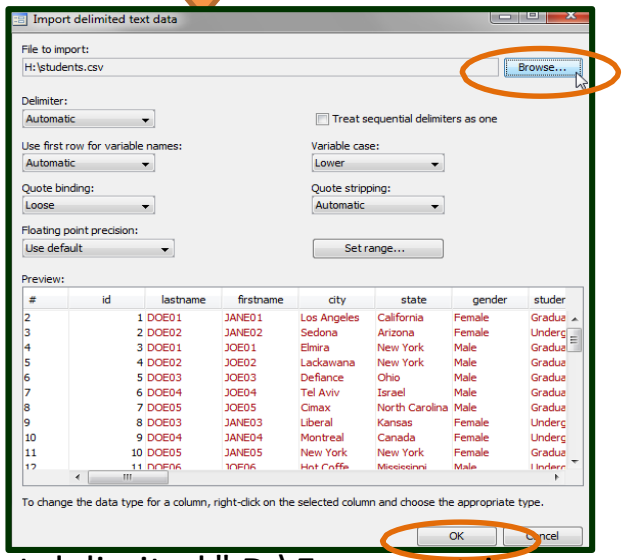
In Excel, select and copy the data you want. Then, in Stata type edit in the command line to open the data editor. Point the cursor to the first cell, then right-click, select '***Paste***'.



Importing data from Excel to Stata using the menu

From *.csv using the menu

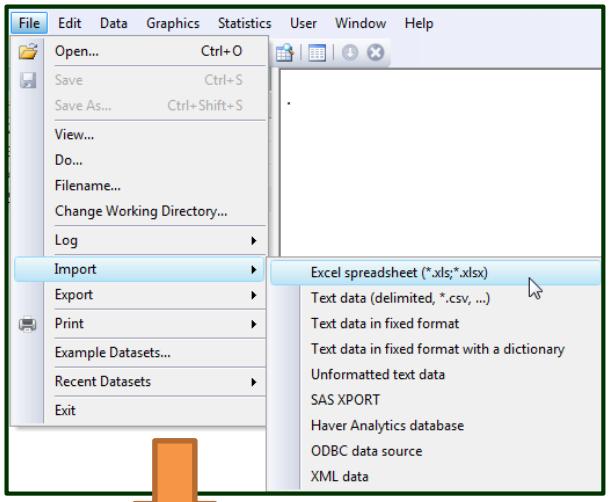


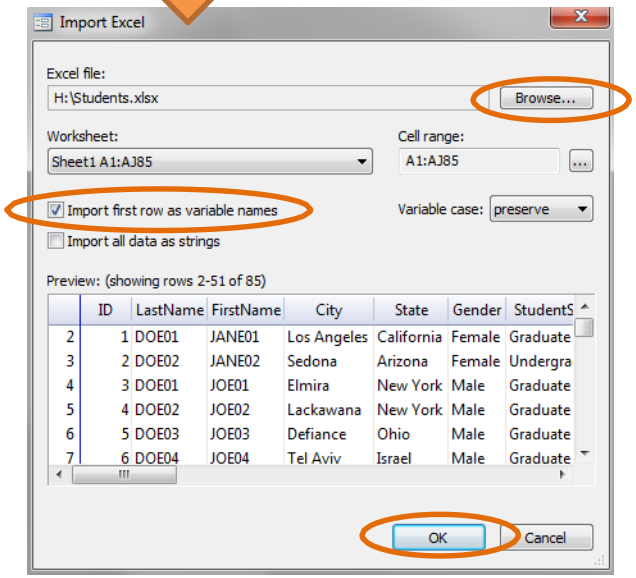


import delimited " D:\Econometrics course\Day 1\ WTP", clear

insheet using WTP, clear

From *.xls(x) using the menu





import excel "D:\Econometrics course\Day 1", sheet("WTP") firstrow clear

Stata variables and color-coded system

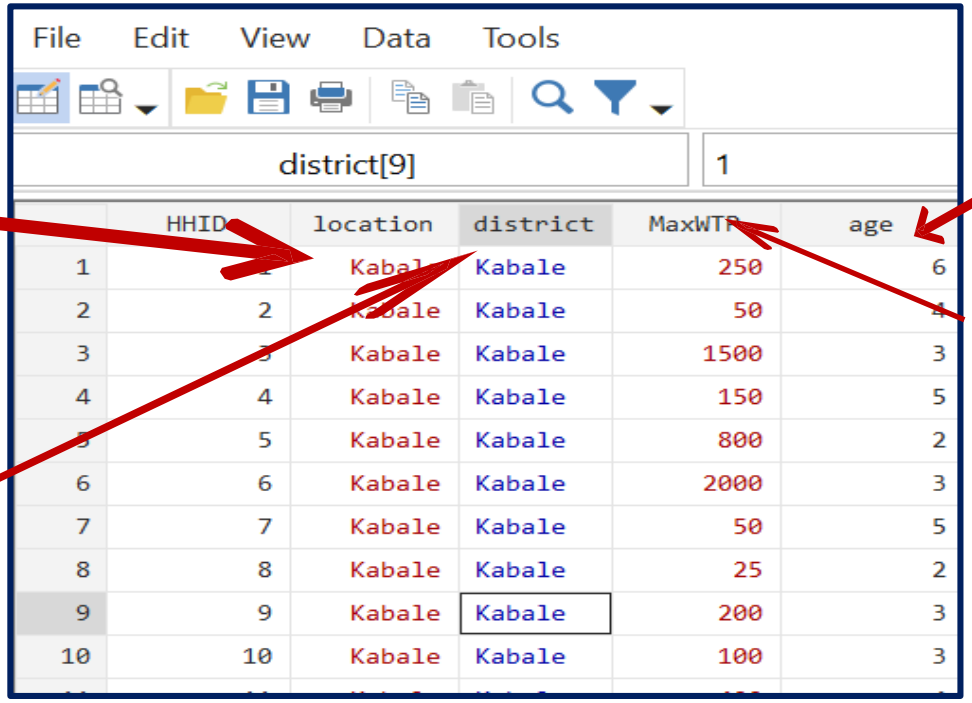
An important step is to make sure variables are in their expected format. To view imported data, type;

`browse` or `edit`

Stata has a color-coded system for each type. **Black** is for numbers, red is for text or string and blue is for labeled variables.

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

For **district** a value 1 has the label "Kabale". It is still a numeric variable



	HHID	location	district	MaxWTP	age
1	1	Kabale	Kabale	250	6
2	2	Kabale	Kabale	50	4
3	3	Kabale	Kabale	1500	3
4	4	Kabale	Kabale	150	5
5	5	Kabale	Kabale	800	2
6	6	Kabale	Kabale	2000	3
7	7	Kabale	Kabale	50	5
8	8	Kabale	Kabale	25	2
9	9	Kabale	Kabale	200	3
10	10	Kabale	Kabale	100	3

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

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

Data types and measurement scales

Qualitative (Nominal or ordinal scales)

- Information that cannot be measured with number
  - e.g. gender, gender- male or female, seed type-improved, hybrid, local
- Can be coded with numbers 0, 1,2,3 but these are simply codes for identification purposes and ease of analysis
- Data analysis terms applicable are
  - **Indicator/binary/dummy variables** –special type of categorical variable i.e. divides data into 2 groups e.g. questions
  - **Discrete/categorical variable**– has limited number of values that form categories or groups

Quantitative (Ratio or interval scales)

- Information that can be measured with numbers
- Discrete: take on integers within a given range, e.g household size
- Continuous: take on any value (including decimals) within a defined range e.g weight, distance in kms, yield, etc

Stata command: `describe`

To get a general description of the dataset and the format for each variable type:

`describe` or `des`

```
. des

Contains data from practice1.dta
  obs:           524
 vars:            16                          25 Jun 2021 06:58
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
HHID            float   %9.0g
location        str7    %9s                   District household is located
district        float   %10.0gc    location   District household is located
MaxWTP          str4    %9s                   Maximum willingness to pay per seedling received (UGX)
age             byte    %10.0gc               age of the respondent
gender          byte    %10.0gc               Dummy variable if respondent was male (=1) or female (=0)
Mstatus         byte    %10.0gc               Dummy variable if respondent was married (=1) or otherwise =0
HeadHH          byte    %10.0gc               Dummy variable if respondent was head (=1) or otherwise = 0
Cunder10        byte    %10.0gc               Number of children in family under the age of 10
hsize           byte    %10.0gc               Size of the family/household
EStatus         byte    %10.0gc               Access to agro-forestry information =1 if received and 0 otherwise
education       byte    %8.0g                 Household head's years of completed education
Income          double  %10.0gc               Household income in thousand ('000) Ugx (per month)
free_seed       byte    %10.0gc               Ever got Free tree seedlings =1; otherwise =0
paid_seed       byte    %10.0gc               Ever paid for tree seedlings =1; otherwise =0
MaxWTA          double  %10.0g                Maximum willingness to accpet compensation per seedling sold (UGX)
-------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

Type `help describe` **for more information**

Stata command: `tabulate`

To get a general description of the dataset and the format for each variable type:

`tabulate` or `tab`

```
. tabulate gender

       Dummy
variable if
 respondent
    was male
     (=1) or
female (=0) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        131       27.12       27.12
          1 |        352       72.88      100.00
------------+-----------------------------------
      Total |        483      100.00

. tabulate gender district

      Dummy
   variable
         if
respondent
   was male
     (=1) or
     female |  District household is located
       (=0) |    Kabale   Kanungu    Kisoro |     Total
------------+---------------------------------+----------
          0 |        41        56        34 |       131
          1 |       157       117        78 |       352
------------+---------------------------------+----------
      Total |       198       173       112 |       483
```

Type `help tabulate` **for more information**

Stata commands: `summarize` **and** `mean`

To calculate and display a variety of univariate summary statistics, type: `summarize varlist`

To produce estimates of means, along with standard errors, type: `mean varlist`

```
. summarize age

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |        469    38.17271    8.788956        19         72

.
. sum age gender hsize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |        469    38.17271    8.788956        19         72
      gender |        483   .7287785    .4450511         0          1
       hsize |        485   3.795876    2.127541         1         14

.
. summarize age, deta

                      age of the respondent
-------------------------------------------------------------
      Percentiles      Smallest
 1%           20            19
 5%           25            19
10%           27            19       Obs                 469
25%           32            20       Sum of Wgt.         469

50%           37                     Mean           38.17271
                      Largest        Std. Dev.      8.788956
75%           44            64
90%           49            66       Variance       77.24575
95%           54            70       Skewness         .4883
99%           62            72       Kurtosis       3.455409
```

Type `help summarize` **for more information**

Stata commands: `generate` **and** `replace`

To create or change contents of variable, use generate or replace

```
generate age2 = age^2
```

```
replace age2 = age^2
```

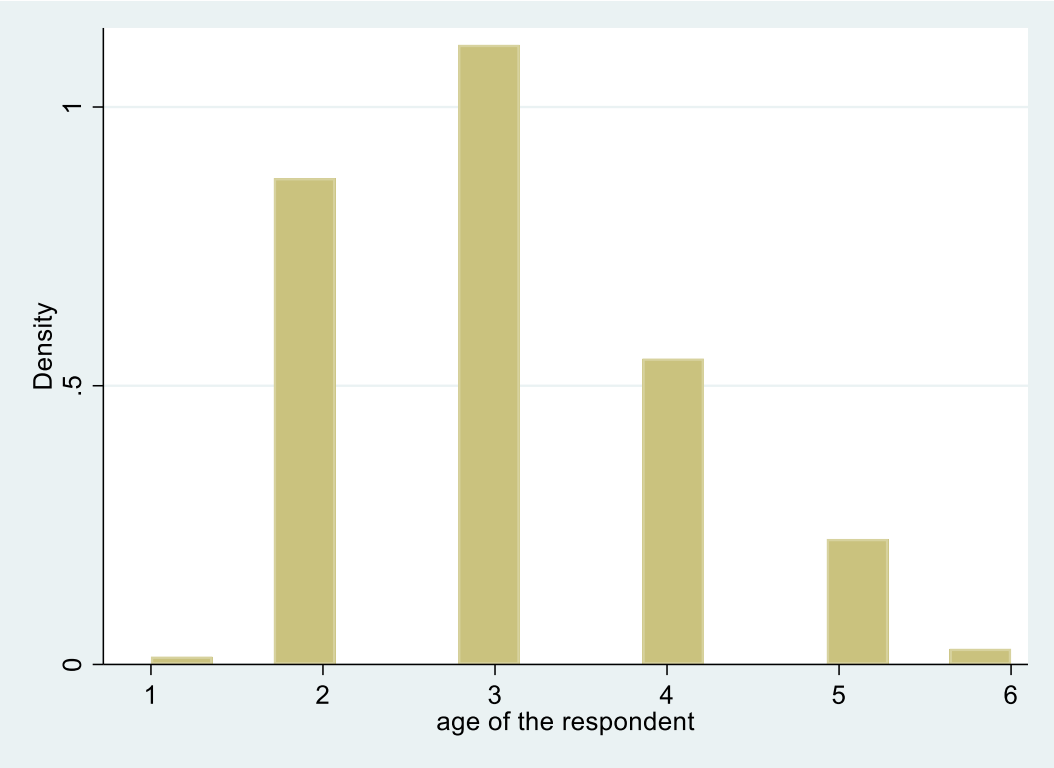
```
. generate age2 = age^2
(55 missing values generated)

. drop age2

. generate age2 =age
(55 missing values generated)

. replace age2 = age^2
(469 real changes made)
```

Type `help generate` **for more information**

Stata common graphing commands

Using the menu

Histograms can be used to visualize continuous and categorical data



Type `help graph` **for more information**

Preliminary data exploration in Stata

- To perform a one-sample t-test

```
ttest MaxWTP==350
```

- To perform a two-sample t-test using groups

```
ttest MaxWTP, by( gender)
```

- To perform a one-way analysis of variance

```
oneway age district
oneway age district, b
```

- To display correlation matrix or covariance matrix

```
correlate MaxWTP age
```

- To display all the pairwise correlation coefficients between the selected variables

```
pwcorr MaxWTP age education hsize
pwcorr MaxWTP age education hsize , sig
pwcorr MaxWTP age education hsize , sig st(0.05)
```

Type `help` *command* **for more details, e.g** `help pwcorr`

Stata tips and tricks

- Stata files can also be opened by dragging them into Stata
- Stata is case sensitive
  - All commands are lowercase
  - Variable names must match exactly
- Use the keyboard shortcut (control d) to execute commands in the do-file
- It is generally unnecessary to save changes to your data set if you used a do-file
  - The do-file should be saved, and can be re-run to replicate what you already did
  - Any saves of a data set should be made with a new file name so as not to change your original file

Survey data management

Commonly used types of data in econometrics

Cross-sectional data

- Data on one or more variables collected at the same point in time e.g. UBoS Livestock census, UBoS household survey for 2018, firms turnover for a given year, operating margins, market shares, individual data at a point in time

Time series

- Data on a set of values that a variable takes at different times.
- Can be collected daily (covid deaths, weather elements stock prices); weekly (exchange rates), monthly (unemployment rate), quarterly, annually …

Panel/Longitudinal data

- Combines elements of both time series and cross-section.
- A given individual (district, country, farm, household) provides data on the same variables for every time point in question
- Example is a data set where a number of firms are randomly selected say in 1990 and traced from that time to 2000

Data sources in econometrics (Primary vs Secondary data)

Primary Data: Collected by researcher directly from the main source

Secondary Data: Has already been collected through primary sources and made readily available for researchers to use for their own research

Clearly answer the following about secondary data

- What was the purpose of collecting the data?
- For what **purpose** do we want to use the data?
- For what purpose(s) are the data fit?
- What does the **process** of producing the data tell us about their fitness for the specified purpose (and for other purposes)?

If you decide to explore using secondary data

- Clearly define what kind of data you need to answer your objectives
- Review available data sets and contact holders for permission
- Review documentation on the data's generation process
- Consult data producers to understand the data collection process and the context(s) in which it was carried out
- Test the data (Missing, outliers, wrong entries, etc)
- Decide to use or not to use it

Survey data collection and mgt practices … (1)

Before Data collection

- Follow all the guidelines you have received about question types and question quality
- First design questions per objective to make sure key variables are not left out
- For key variables, explore options of having 2 or 3 constructs of the same variable
- Arrange your questions logically (e.g. do not start with sensitive questions, some sociodemographic are getting sensitive)
- Find appropriate proxy measures for sensitive questions or latent variables (intuition, literature and key informants/opinion leaders
- Establish a careful balance between Open vs close ended questions
  - Most open ended questions can be answered in community (KII and FGD) questionnaires
- Choose your enumerators carefully; your friend innocently let you down
- Pre-test and pre-test
- Try to capture variables in their most natural/raw form
- Where rating scales are involved, aim at the higher scales e.g. a scale of 5 compared to 3
- Triangulating methods

Survey data collection and mgt practices … (2)

During data collection;

- Before you start, confirm all key variables are well captured
- On the first or second day, read through the data collected
  - Check if your key variables are coming out well
  - Make necessary changes if necessary-document the changes appropriately
  - Do not discard the old format variables
- Timing of data collection is very important
  - How available are the respondents-may affect sampling efficiency
  - Is there seasonal variability in your variables of interest
  - Is it the best timing for your variables of interest
- Try to summarize notes on every location while still in the location
- Maintain properly dated and identifiable field note
- Make friends with the field guides, you may need to call back for more information
- Don’t push the enumerators too hard, they aren’t machines
  - During the pretest, set a realistic daily target and use the first 2 days of data collection to revise this if necessary
  - Data quality checks should be done as often as feasible

Survey data collection and mgt practices … (3)

After data collection

- Secure an original file of your data and hard copies(scans) of field notes
- Code open ended questions early enough
  - Some context specific responses will make more sense
- Generate clear methods to link community, household, and individual level data
  - Confirm village names in the household data match the field notes
  - Check for duplicate entries in the data if CAPI was used or ensure all hard copy questionnaires are uniquely identifiable
- Allocate time to explore your data and check for
  - Outliers
  - Missing data
  - Incorrect entries *o* instead of 0

Missing data … (1)

Goals of statistical analysis with missing data

- Avoid bias
- Maximize use of available data
- Obtain appropriate estimates of uncertainty

If you treat some values as missing; Make sure they exist!!!

- Commonly used;
  - Complete case analysis (listwise deletion);
    - Delete cases missing data on any variable of interest
    - Loses sample size and statistical power hence larger standard errors
  - Available case analysis(pairwise deletion)
    - Calculate means, variances, and covariance matrices based on all available non-missing data
    - Less lose of power due to missing information but no consistent sample size
    - Parameter estimates often different from estimates from a full sample size
  - Unconditional mean imputation
    - Replace missing value of an individual with the overall mean from the available cases
    - Leads to artificial reduction in variability (Centre value replacement)
    - Changes magnitude of correlation between imputed variables and other variables

Missing data…. (2)

Imputation methods in missing data analysis

- Single or deterministic imputation
  - Replaces missing values with predicted scores from a regression
  - Strength: it uses complete information to impute values
  - All imputed values will fall directly on the regression line –decreasing variability
  - Inflates the association between variables (Applied missing data analysis Craig Anders, 2010)

- Stochastic imputation
  - A residual term is added to the regression scores from the regression imputation to restore some of the lost variability
  - Superior from previous methods and produces unbiased coefficient estimates
  - Although standard errors are less biased, they have been changed.

- Multiple imputation:
- Is an iterative form of stochastic imputation
- Instead of imputing one value, the distribution of the observed data is used to estimate multiple values that reflect the uncertainty around the true value

Missing data analysis are difficult because there is no inherently correct methodological procedure

Missing Data … Multiple Imputation

Phases in MI

1. Imputation or fill-in phase: The missing data are filled in with estimated values and a complete data set is created. The process of fill in is repeated m times

2. Analysis phase: Each of the m complete data sets is then analyzed (separately) using statistical methods of interest e.g. (linear regression)

3. Pooling phase: The parameter estimates (e.g. coefficients and standard errors) obtained from each analyzed data set are then combined for inference

Type `help mi` in Stata for more details and estimation procedures

Common misconception of missing data methods: Assumption that the imputed values should represent "real" values.

Actual Purpose: Correctly reproduce the variance/covariance matrix we would have observed had our data not had any missing values