Introduction to Quantitative Methods for Development

- Tutorial 1 -

Exercise 1

Assume the simple regression case with two variables y and x.

(a) Write down the population regression function.
(b) Write down the sample regression function.
(c) Derive the sample parameters $\hat{\beta}_0$ and $\hat{\beta}_1$.

Exercise 2

You are interested how wage changes with age.

The Table 1 below shows the characteristics of 10 randomly sampled workers from a South Indian town.

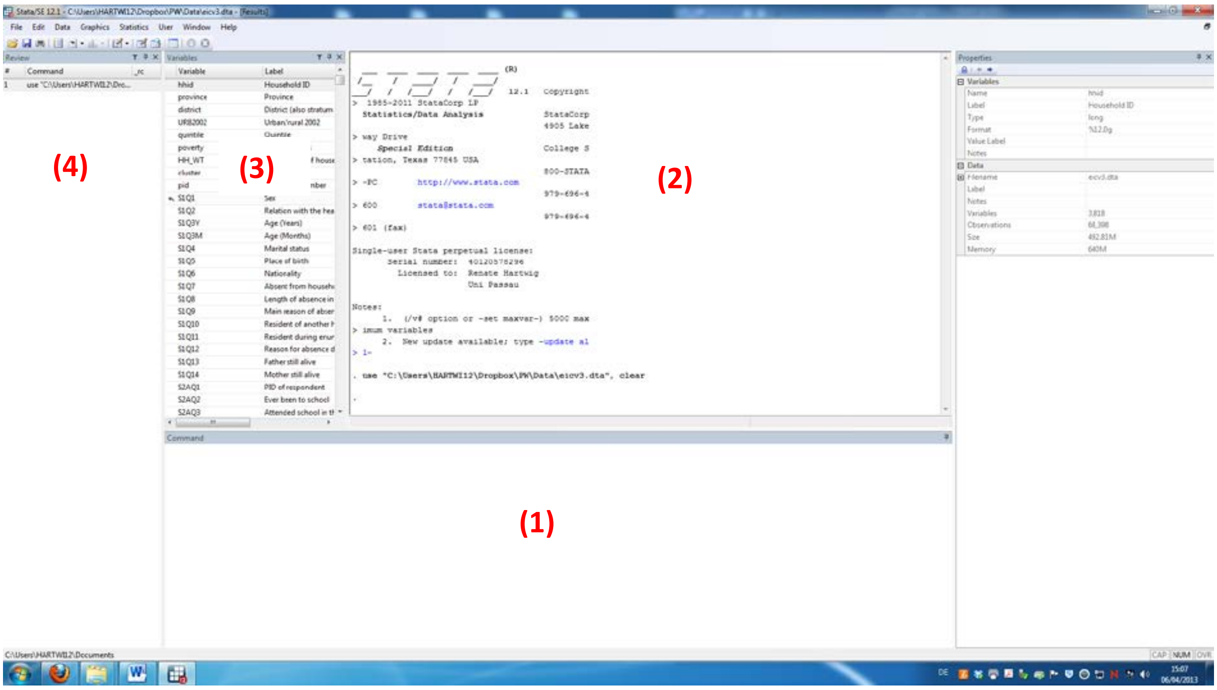
Table 1: Worker characteristics

#	Wage (in thousand rupees)	Gender	Education (in years)	Age (in years)
1	120	Male	1	57
2	224	Male	4	48
3	132	Male	1	38
4	75	Male	3	27
5	111	Female	3	23
6	127	Male	3	22
7	30	Male	1	18
8	24	Male	1	12
9	119	Male	1	38
10	75	Male	1	55

(a) Write down the regression function you would estimate if you are interested in investigating the relationship between wage and age.
(b) Assume the simple regression case, i.e. only considering wage and age and calculate the intercept and slope parameter with the data at hand.
(c) Write down the results in regression format.
(d) How do you interpret the coefficients?
(e) Do you think $\hat{\beta}_1$is an unbiased estimate? Explain why/why not.

Introduction to Stata

1. The Stata Environment


Stata/SE 12.1 - C:\Users\HARTWI12\Dropbox\PW\Data\eicv3.dta - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
Command
(4)
(3)
(2)
(1)
C:\Users\HARTWI12\Documents
CAP NUM OVR

#	Command	_rc
1	use "C:\Users\HARTWI12\Dro...	

Variable	Label
hhid	Household ID
province	Province
district	District (also stratum
URB2002	Urban/rural 2002
quintile	Quintile
poverty	
HH_WT	f house
cluster	
pid	nber
S1Q1	Sex
S1Q2	Relation with the hea
S1Q3Y	Age (Years)
S1Q3M	Age (Months)
S1Q4	Marital status
S1Q5	Place of birth
S1Q6	Nationality
S1Q7	Absent from househ
S1Q8	Length of absence in
S1Q9	Main reason of abser
S1Q10	Resident of another h
S1Q11	Resident during enur
S1Q12	Reason for absence d
S1Q13	Father still alive
S1Q14	Mother still alive
S2AQ1	PID of respondent
S2AQ2	Ever been to school
S2AQ3	Attended school in th

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   12.1   Copyright
>  1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lake
> way Drive
     Special Edition                  College S
> tation, Texas 77845 USA
                                      800-STATA
> -PC        http://www.stata.com
                                      979-696-4
> 600        stata@stata.com
                                      979-696-4
> 601 (fax)

Single-user Stata perpetual license:
       Serial number:  40120578296
         Licensed to:  Renate Hartwig
                       Uni Passau

Notes:
      1.  (/v# option or -set maxvar-) 5000 max
> imum variables
      2.  New update available; type -update al
> l-

. use "C:\Users\HARTWI12\Dropbox\PW\Data\eicv3.dta", clear

.
```

Variables	
Name	hhid
Label	Household ID
Type	long
Format	%12.0g
Value Label	
Notes	
Data	
Filename	eicv3.dta
Label	
Notes	
Variables	3,818
Observations	68,398
Size	482.81M
Memory	640M

The 4 main windows are:

(1) The Command Window: where you enter/type commands
(2) The Results Window: where the results of a command are displayed
(3) The Variables Window: where the list of variables in your dataset are displayed
(4) The Review Window: which shows the trail of commands recently executed

2. Data Entry

There are several ways to read data into Stata's data editor. The data editor is a spreadsheet that looks very much like Excel.

For small datasets in Excel format the easiest way to load your data into Stata is to copy and paste it into the data editor:

- Arrange your data in Excel (variables across and observations down)
- Copy your data (incl. variable names) and go back to Stata
- Click on the data editor icon (8th Icon from the left) and paste (Ctrl+V)
- Save your file in Stata format (File-> Save as…)

If your data is already in Stata format (i.e. with the extension **.dta**), you have two options:

- Use the command window and type **use file location** (e.g. c:\India.dta)
- Click File->Open->(choose directory)

Data saved in other statistical packages like SPSS, SAS etc. can be converted into .dta-files using, for example, the software 'StatTransfer'.

3. Data Management

Once you load and save your data in Stata, the next step is to explore the nature of your data and start the analysis. If the data has been collected through a survey, reading the associated questionnaire in parallel obviously helps a lot to understand the content. Beyond that, however, there are still a few preparations to be made.

3.1 Log-File

Stata allows you to keep a running record of your statistical results – i.e. a separate record of the results of your commands. This is done in a log file. You can open a log file by:

- Clicking on the log-file icon (4th icon on the left)
- Clicking File->Log->Begin

A dialog box appears inviting you to type a file name and select a save-as type. Save your log-files in **.log** (not .smcl) as this allows you to edit your results later on.

If you want to temporarily stop recording your results (maybe because you are not sure you need to keep them), just type **log off** in the command window. Whenever you want to resume recording type **log on** in the command window. To close a log file type **log close**.

3.2 Do-File

Instead of typing each command individually in the command window, Stata allows you to execute a number of commands at once from a do file. A do-file works like a programme file. Do-files can be opened by:

- Clicking on the do-file icon (7th icon on the left)

An advantage of a do file is that you can repeat your sequence of commands as often as you want (or modified versions of it). It also allows you to share an analysis with others.

4. Basic Commands

In the following we use data from a labour force survey covering 261 workers in an industrial town in southern India in 1990. The data is contained in the file India.dta

4.1 Data Description

To see the content of your data click Data->Describe data->Describe data in memory

Alternatively you can also just type **describe** in the command window. This command displays the number of variables and observations in the data as well as the list of all variable names and labels.

In the following we will work with the commands directly instead of clicking through the menu (you can explore this option in your own time).

4.2 Summary Statistics

By summary statistics we refer to the mean, median, standard deviation, percentiles etc. It is important to differentiate between two groups of summary statistics: *mean based* or *order based* summary statistics.

Mean based summary statistics include the mean, variance, standard deviation, skewness, and kurtosis. The order based summary statistics include the median, the first and third quartiles, the inter quartile range (IQR). These statistics provide us with measures of central tendency (a value to which all observations converge), measures of spread (how observations are scattered around the central value) and also the shape of the distribution.

```
summarize age
sum age wage
sum wage, detail
```

Another useful command to get summary statistics is **tabstat**.

```
tabstat age
tabstat age, stats (n mean var sd cv p25 p50 p75 IQR)
tabstat age wage, stats (n mean var sd cv p25 p50 p75 IQR)
```

The preceding summary statistics are useful when we have quantitative variables. There is however a class of variables whose numeric values do not carry the usual meaning and property. We call such variables categorical variables. Categorical variables do not measure a concept as quantitative variables; they are rather count variables in the sense that we use them to count individual members with certain characteristics. In our data set *gender* and education are examples of categorical variables.

For categorical variables we do not use the usual measures of central tendency and spread, like the mean and standard deviation, simply because the number in categorical variables represent class/quality. Instead we aim to count the number of observations (frequency or relative frequency) in each category of a categorical variable.

```
tab education
tab education gender
tab education gender, row
tab education gender, column
tab education gender, sum(wage)
```

Stata commands can be qualified so as to operate over a specific range/portion of the data. You have seen that **tab education** produces the frequency table of *education* for all individuals in your sample. Suppose you want to see the frequency table of *education* only for women then you can use **if** as a qualifier:

```
tab education if gender==1
```

Similarly, you may be interested to see the age, gender and education levels of the top 10 wage earners in your sample:

```
sort wage
list wage in -10/l

list wage in 1/10
sum wage if age<30
list wage if gender==1 & wage<30
```

4.3 Creating New Variables

To create new variables we use the **generate** (short **gen**) command.

```
gen doublewage=wage*2
gen wage2=wage*wage
gen wage_2=wage^2
gen lnwage=ln(wage)

rename doublewage dblwage

replace wage2=wage/1000
```

In your data you have information on the age of the individual. Suppose you want to create a categorical variable with age groups.

```
gen agecat=1 if age<20
replace agecat=2 if age>=20 & age <40
replace agecat=3 if age>=40

tab agecat

label variable agecat “age groups 20 20-40 40+”
```

To delete variables from your data:

```
drop wage dblwage

keep wage dblwage
```

Note the difference!

5. Help and Find-it Function

5.1 Help Function

Stata has an online-based help-function. For example, if you are looking for more details on the commands and how they are specified type **help** followed by the command you are looking for into the command window.

e.g. **help sum**

5.2 Find-it Function

The find-it Function works similar to the help function and is particularly useful to look for programme packages that are not included in the standard Stata version.

e.g. **findit ivreg**

References:
Kohler, U., and F. Kreuter (2005). *Data Analysis using Stata*. Texas: Stata Press.

Exercise:

- What is the average wage for males only? What is average female wage?
- How much of the wage can be explained by gender and education?
- Show the distribution of the education variable graphically.

Also, for further exercise and practise see Khandker *et al.* Chapter 11 (Introduction to Stata).

Introduction to Quantitative Methods for Development

- Tutorial 2 -

Exercise 1

Fill in the blank. In a regression line, the ________ the standard error of the estimate is, the more accurate the predictions are.

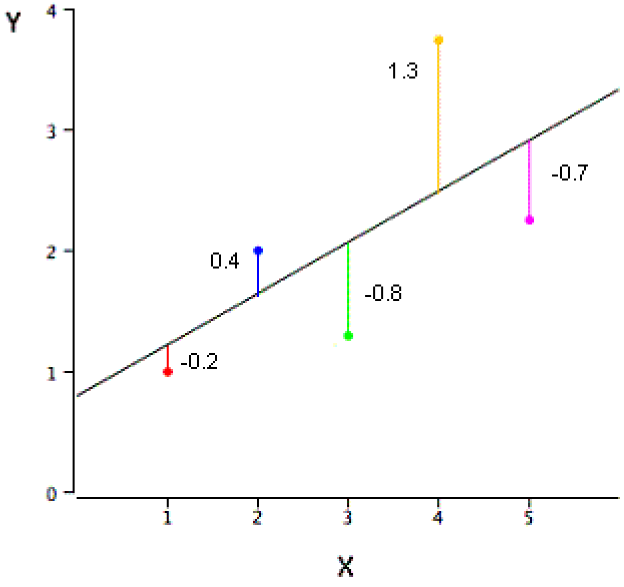
(a) Larger

(b) Smaller

(c) The standard error of the estimate is not related to the accuracy of the predictions.

Exercise 2

The graph below represents a regression line predicting Y from X. This graph shows the error of prediction for each of the actual Y values. Use this information to compute the standard error of the estimate in this sample.



Exercise 3

We want to verify if income has a negative influence on infant mortality. The following model has been used to explain the deaths of children under 5 years per 1000 livebirths (deathun5).

$$deathun5 = \beta_1 + \beta_2\, gnipc + \beta_3 ilitrate + u$$

where *gnipc* is the gross national income per capita and *ilitrate* is the adult illiteracy rate in percentage. With a sample of 130 countries, the following estimation has been obtained:

$$\underset{(5.93)}{deathun5 = 27.91} \; \underset{(0.00028)}{-0.000826\, gnipc} + \underset{(0.183)}{2.043\, ilitrate} + u$$

The numbers in brackets below the estimates are the standard errors. You are asked to determine if there is a negative relationship

(a) What are the hypotheses that you are going to test?

(b) Assuming a 1% significance level do you accept or reject the null and what do you in consequence conclude about the relationship?

(c) What does the statistical significance mean?

Practise with Stata

The file DHS2000.dta contains data on 11,926 children from the Demographic and Health Survey conducted in Malawi in 2000. The file DHS2005.dta contains information of 10,914 children from the Demographic and Health Survey in Malawi conducted in 2005.

Your tasks:

1. Load the data into Stata

2. Create a log file

3. Create a do file

4. Explore your data

    a. Of what age are the children in your data?

5. Combine both data sets into a single file named Malawi.dta

6. Create the following variables for analysis

    a. d_rural: a dummy indicating whether the child lives in a rural or urban environment.
    b. d_twin: a dummy indicating whether the child is a twin or not.
    c. d_male: a dummy indicating whether the child is male or not.
    d. d_dead: a dummy indicating whether the child has died or not.
    e. d_femalehead: a dummy indicating whether the head of the household is female or not.
    f. d_married: a dummy indicating whether the mother is married.
    g. mom_age_birth: a variable indicating the mother’s age at the birth of the child.
    h. childage: a variable with the current age of the child in years.
    i. d_2005 a dummy indicating whether the child has been surveyed in 2004/5 or not.

7. Compare the characteristics of the children in 2000 to those in 2005. Are there significant differences between them?

8. Show your results in a table.

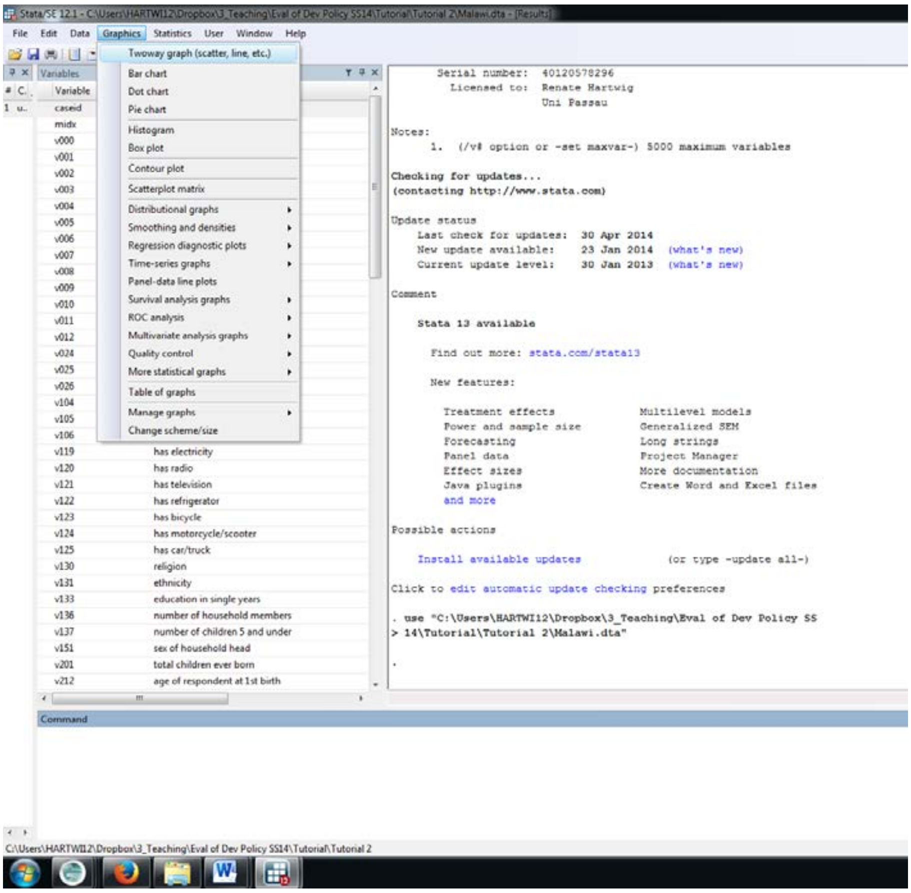
Introduction to Quantitative Methods for Development

- Tutorial 3 -

Graphics with Stata

Load the Malawi dataset (Malawi.dta) that you have been using previously.

As a starting point, all basic graphs can be produced via the Menu 'Graphics'

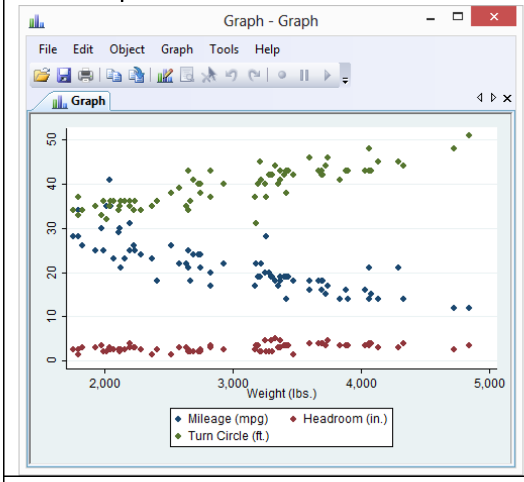


Variable	
caseid	
midx	
v000	
v001	
v002	
v003	
v004	
v005	
v006	
v007	
v008	
v009	
v010	
v011	
v012	
v024	
v025	
v026	
v104	
v105	
v106	
v119	has electricity
v120	has radio
v121	has television
v122	has refrigerator
v123	has bicycle
v124	has motorcycle/scooter
v125	has car/truck
v130	religion
v131	ethnicity
v133	education in single years
v136	number of household members
v137	number of children 5 and under
v151	sex of household head
v201	total children ever born
v212	age of respondent at 1st birth

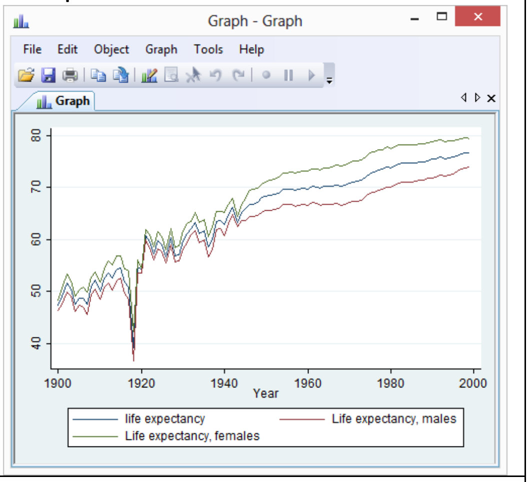
```
          Serial number:  40120578296
            Licensed to:  Renate Hartwig
                          Uni Passau

Notes:
      1.  (/v# option or -set maxvar-) 5000 maximum variables

Checking for updates...
(contacting http://www.stata.com)

Update status
    Last check for updates:  30 Apr 2014
    New update available:    23 Jan 2014  (what's new)
    Current update level:    30 Jan 2013  (what's new)

Comment

    Stata 13 available

      Find out more: stata.com/stata13

      New features:

        Treatment effects            Multilevel models
        Power and sample size        Generalized SEM
        Forecasting                  Long strings
        Panel data                   Project Manager
        Effect sizes                 More documentation
        Java plugins                 Create Word and Excel files
        and more

Possible actions

    Install available updates             (or type -update all-)

Click to edit automatic update checking preferences

. use "C:\Users\HARTWI12\Dropbox\3_Teaching\Eval of Dev Policy SS
> 14\Tutorial\Tutorial 2\Malawi.dta"

.
```

Examples include:

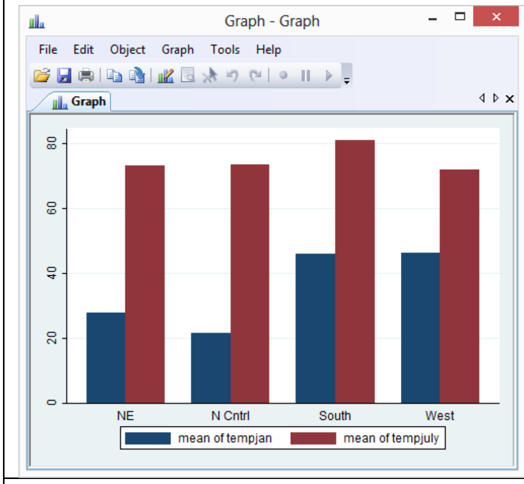
Scatter plots



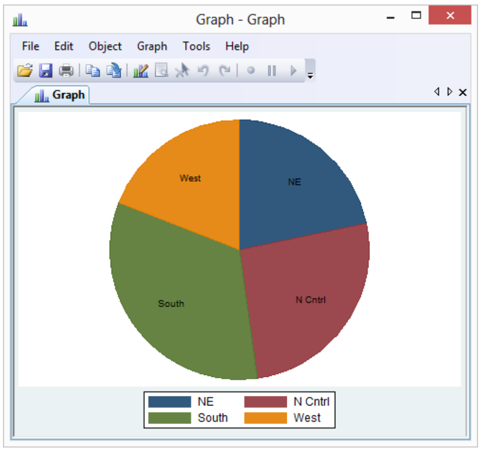
Line plots



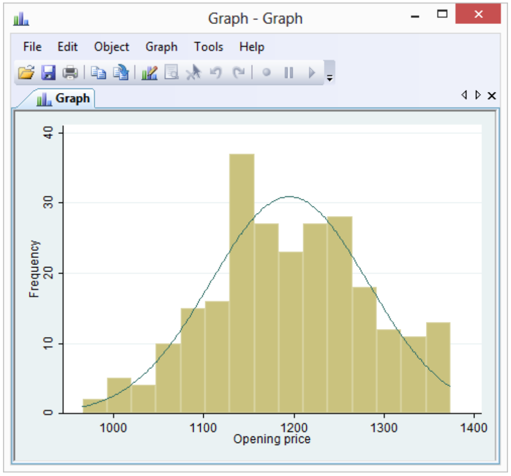
Bar charts



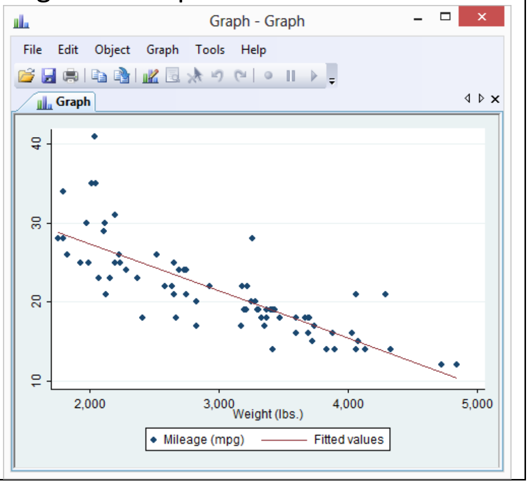
Pie charts



Distribution plots



Regression it plots



To begin wit we want to look at some distributional features of the variables in our data. For now let us concentrate on the age distribution of the children in our dataset.

1) What kind of graphs could be useful to look at the distribution?

2) Produce a histogram entitled ‘Age Distribution (DHS 2000 & 2005)’. The x-axis should display the age (in months) and be labelled accordingly.

3) You do have a strong colour preference and what the bars to be displayed in red.

4) Save the graph.

Child stunting is a major issue in Malawi. We therefore now want to graphically explore how serious this problem is for children, at different ages, from different backgrounds and how it evolved over time.

5) Investigate if there are any potential differences in stunting for boys and girls

6) Investigate the relationship between stunting and age.

7) Do children from wealthier backgrounds fare better? Is there a considerable difference?

8) How has stunting evolved over time?

Analysis with Stata

Following the graphical exploration in the following we want to identify and analyse the factors that influence stunting in children. In order to identify the determinants of stunting we will use standard OLS regressions.

For now we concentrate on the observations from 2000

1) In your data set which variables do you think have an effect on stunting?
    a. Group them in appropriate categories (i.e. Household characteristics, Location etc…).
    b. Identify the variables which are most useful.

2) How do you interpret your regression results?

3) Are the same determinants also relevant in 2005? Any changes? How do you interpret those?

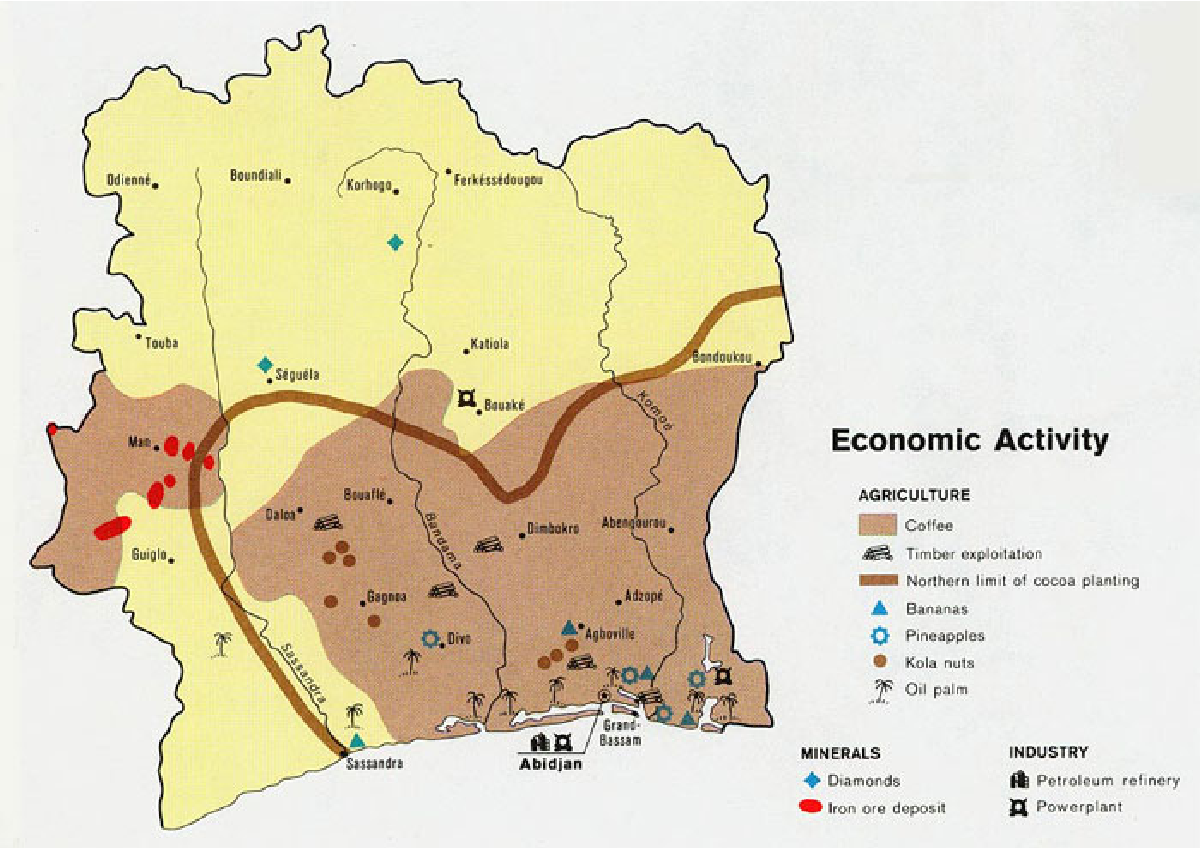
4) Present you regression output in a table.

Finally we are approaching the main objective of our analysis. Can you identify an effect of the food crisis?

Introduction to Quantitative Methods for Development

- Tutorial 4 -

Today you are taking a trip to Ivory Coast, more precisely the cocoa belt.



In 2012 you have spent a few months in Cote d'Ivoire to collect data on cocoa farming for your master thesis. In your master thesis you want to investigate if farmers are using fertilizer, what are the constraints to fertilizer use and what are the potential benefits. During your stay, you interviewed 340 farmers across 12 villages spread through the cocoa on their farming practises and cocoa production. You got some funds for your fieldtrip so you could hire a research assistant that helped you coding the survey. You are back to Passau enjoying the autumn sun and your research assistant just sent you the data. He named the file Cocoa.dta. Your research assistant is also a very good one and he has also already coded a few variables for you (but not all, because he got bored of listening to the same songs over and over again while coding…).

So, now it's your turn to finish the job and get cracking with your analysis.

A few things you want to think about:

- How do you go about it?
- What is the regression specification you might want to run?
- Do you have all the variables you need? What do they look like?

You are meeting your thesis supervisor to present the first results. She is going to ask you a bunch of questions:

a) Why have you chosen this specification?
b) Can you think of a better alternative?
c) Why are you not using the log of certain variables?
d) Would it make sense to standardize some variables?
e) How do you interpret your regression results?
f) Have you looked at the residuals of your regressions?

Introduction to Quantitative Methods for Development

- Tutorial 6-

Unfortunately you are still stuck with your Farmers in Cote d'Ivoire.

To compensate for all the stress and frustration you have been eating tons of Mars bars and you are seriously considering entering these kind of competitions now…..

Before running away to Japan, you have to get cracking and finish some unfinished business.

Your supervisor is really nagging you and has been asking

a) If you think heteroscedasticity is a concern in your regression and why?

b) If you tested for heteroskedasticity. i.e.

i. Which test did you use?
ii. What was your null hypothesis?
iii. What was the result of your test?

c) If heteroscedasticity did matter, how did it change your estimation results?

What do you respond?

Introduction to Quantitative Methods for Development

- Tutorial 7-

For a quick review respond to the following 3 questions.

Question 1:
List 7 examples of time series data relevant for development studies?

Question 2:
What is the meaning of a unit root?

Question 3:
What is the Granger test? What does it test for?

Exercise

Road traffic accidents are among the most common reasons for death in developing countries. A number of countries have started to put increasing emphasis on road safety measures and laws. A country that has been quite progressive early on – in the 1980s – was South Africa. You are interested how laws and economic conditions affect driving behaviour. Use traffic2.dta to answer these questions.

a) The variable *prcfat* is the percentage of accidents resulting in at least on fatality. Note that this variable is a percentage, not a proportion. What is the average of this variable over this period?

b) Run a regression of *prcfat* on a linear time trend, 11 monthly dummies (set January as your base month), *wkends*, *unem*, *spdlaw*, and *beltlaw*. Discuss the estimated effects of *unem*, *spdlaw*, and *beltlaw*. Do the signs and magnitudes make sense to you?

c) Test the errors for AR(1) serial correlation.

d) Re-estimate the model accounting to serial correlation.

e) Compute the first order autocorrelations for *unem* and *prcfat*. What do these suggest about possible unit root(s)?

f) Estimate the model in (b) using first differences for *unem* and *prcfat* (Do not difference the month or policy variables.) Compare your results to those in (b).

Introduction to Quantitative Methods for Development

- Tutorial 8 -

Difference-in-Difference (DID)

Difference-in-Difference with cross-section data:

The Malawi.dta contains information on children (0-5 years) surveyed in the Demographic and Health survey in 2000 and 2004/5. In each year the children to be surveyed where randomly selected. Thus, in the data you do not observe the same children twice but you have a different group of children at each point in time. Hence, the Malawi.dta contains cross-section information or given that it combines two sets of cross-section data we also call it pooled cross-section data.

In 2002 Malawi experienced a major food crisis due to crop failure leaving about 2.5 million people at the verge of starvation. Past research shows that particularly the first 2-3 years in life are crucial for later development and that negative health shocks encountered during these early years in childhood can have lasting consequences for outcomes later in live. Against this background you are interested in investigating the impact of the food crisis on child health to see if there are any long-term negative consequences to be expected. The outcome you are interested in is the weight-for-height z-score. The Weight-for-height z-score (WHZ) is a measure of the deficit in tissue and fat mass and is sensitive to temporary food shortages and episodes of illness and therefore a frequently used indicator to identify short-term malnutrition. Usually children with WHZ below -2SD are considered to be malnourished.

Given that not all regions of Malawi have been hit equally bad by the crisis at the time and some regions have even been spared we can use this variation in the severity of the shock to construct a difference-in-difference estimator to identify the effects of the food crisis.

The basic estimation equation to be implemented is as follows:

$$y_i = \beta_0 + \beta_1 C_i + \beta_2 M_i + \beta_3 H_i + \eta_n + \delta_1 dT_n + \delta_2 dS_t + \delta_3 dS_t dT_n + u_i,$$

where y represents the outcome of interest, weight-for-height (WHZ). The subscript i indicates that the outcome varies per individual, n by district and t by time period. With T representing the treatment group, the dummy dT captures possible permanent differences between the affected and non-affected regions, the dummy dS indicates the survey year (2000=0, 2004=1) and thus is a time-effect absorbing aggregate factor that would cause changes in the outcome variable over time for all observation units, even in the absence of a shock or intervention. The effect of interest is the so-called treatment effect, δ_3, which is associated with the interaction between belonging to the treated group and the time effect. More precisely, it captures the difference in means within the treatment and control group before and after the crisis and thus provides an estimate of the impact of the food shortage on the respective outcome variable.

In order to increase the precision of the estimate, a number of control variables have been included in the specification, accounting for potential observable differences between the treatment and control group. C is a vector of child characteristics e.g. gender of the child, age etc. M includes information about the mothers, i.e. the years of education. The vector H contains household specific information, i.e. the wealth category, or whether the head of the household is a woman.

- The variable d_2004 is a dummy variable indicating whether the child has been survey in 2000 (=0) or 2004 (=1). The variable d_affected indicates whether the child was living in a region that was hit by the crisis or not. In order to implement the above estimation equation in Stata you have to generate a new variable representing the interaction term. To do that type:

  *gen interaction=d_2004*d_affected*

- Now you can implement the estimation equation using the *reg* command as you have already seem several times now.

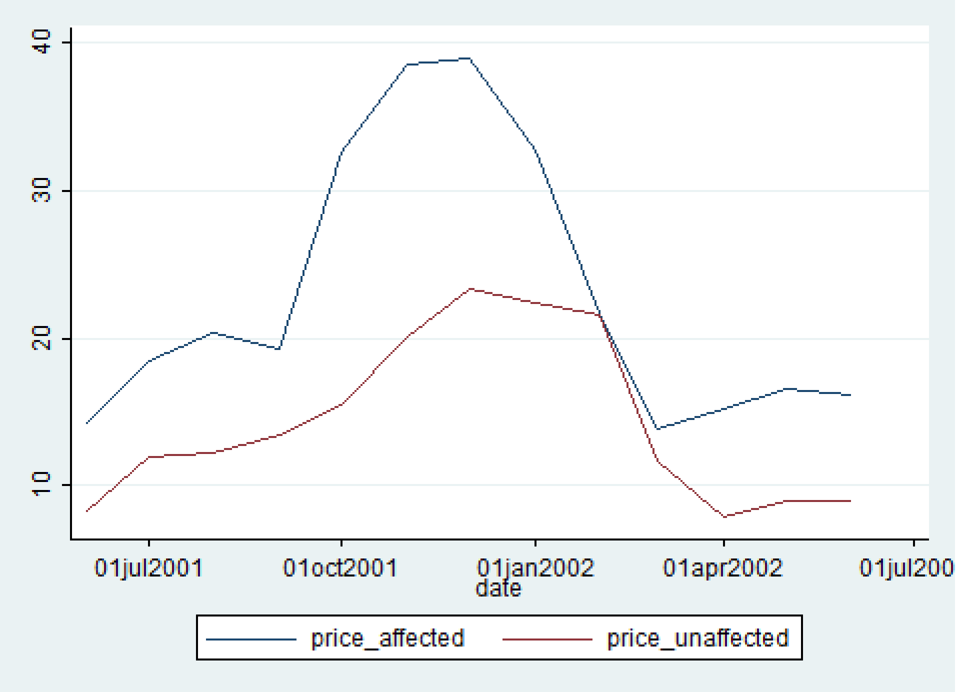
  reg whz male childage_month …. d_2004 d_affected interaction, robust cluster(v001)

- The coefficient on the interaction gives you the impact estimate. How would you interpret the results you got?

- Given that children of different ages or male and female children might have been differently affected by the crisis, try to estimate the effect of the food crisis separately for male and female children and children aged between 6-24 months and above 24 months separately.

For a difference-in-difference estimation to give you an unbiased estimate, one assumption that has to hold is the so called parallel trends assumption. This basically means that prior to the event or intervention, the treatment and control groups should have developed similarly over time. In our example, the affected areas are those in which maize prices between January and March 2002 were particularly high. The data set Prices.dta contains information on the average prices in the affected and unaffected regions.

- Using this data plot a graph which shows the development of the Maize price in the regions over time. Have a go at it using the graphics menu you saw in the previous tutorial. You should get a graph looking like this:



Based on the graph, would you say that the parallel trends assumption holds in our case?

Difference-in-Difference with panel data:

In this case we are evaluating the impact of microcredit in rural Sri Lanka. The dataset SriLanka_9198.dta contains data on 826 households which have been repeatedly sampled in 1991 and 1998. The variable “credit” indicates whether the household has benefited from a microcredit by 1998 or not.

Your task is to evaluate the impact of the microcredit on household food consumption.

1. Have a look at your data and the information you have. Identify which variables will be important in order to apply a difference-in-difference approach analogue to the cross-section example above.

2. Write down the regression function which you would use to estimate the effect of microcredit on food consumption. Which control variables would you include and why?

3. Start by running a basic regression with the outcome variable and the difference in difference operators. Your command should be:

   reg expfd year credit interaction, robust

   - Which coefficient gives you the impact estimate?
   - How do you interpret the regression results?

4. Often, when looking at expenditure as outcome you will see that researchers are not using the absolute values but that they are using log(expenditure).

   - Why would you do that?
   - Generate a new outcome variable called lexpfd which is the natural log of the expenditure variable.

     gen lexpdf=ln(expdf)

   - Rerun the basic regression from above with log-food expenditure as outcome. How does this change the results? How would you interpret the regression results now?

5. Now, include control variables in your regression. How does this change your impact estimate?

6. The microcredit programme has not been implemented using random assignment, instead households are free to choose whether to ask for a loan or not. Given that we have no information on the underlying motivation on why some households asked for credit and others did not, the results that you have obtained so far might be influenced by something called "selection bias".

   In the panel data that you have you do observe the same households at two points in time. This situation allows to control for factors that are actually unobserved but which remain constant overtime. (Basically by subtracting the actual condition of the household from the initial condition all factors that remained stable at the two periods should be netted out). This is done by using a fixed effects model. The basic command would be:

   xtreg lexpfd year credit interaction, fe i(nh)

   xtreg is a panel data command. Have a look at the description of this command using the help function in Stata. The "fe" indicates that you are running a fixed effects model. With the "i(nh)" you tell Stata which variable identifies the household. Remember, the household ID has to be unique for each household interviewed in 1991 and 1998 respectively. Do you remember how to check for duplicates?

7. Now include control variables in your fixed effects regression. Are the results from the fixed-effects regression different to the ones you got before from the pooled regression?

Introduction to Quantitative Methods for Development

- Tutorial 9 -

Logit and Probit models

The dataset insurance.dta contains information on the insurance status and individual characteristics of a random sample of elderly in Mexico collected by the Society of Aging - a think tank in Mexico City in 2012. You have already seen regression results from this data in the lecture. I now want you to:

1. Define a regression specification which allows you to investigate the determinants of insurance enrolment.
   a. Which variables would you include and why?

2. Run an OLS regression with insurance enrolment as dependent variable.

3. Apply a probit model and compare the results.

4. Apply a logit model and compare the results.

5. What do you conclude based on the models that you have estimated?

   a. Which specification, is the most appropriate? Explain.
   b. You are hired by the Ministry of Health of Mexico. Based on your analysis what would you tell the Minister and what policies would you propose?