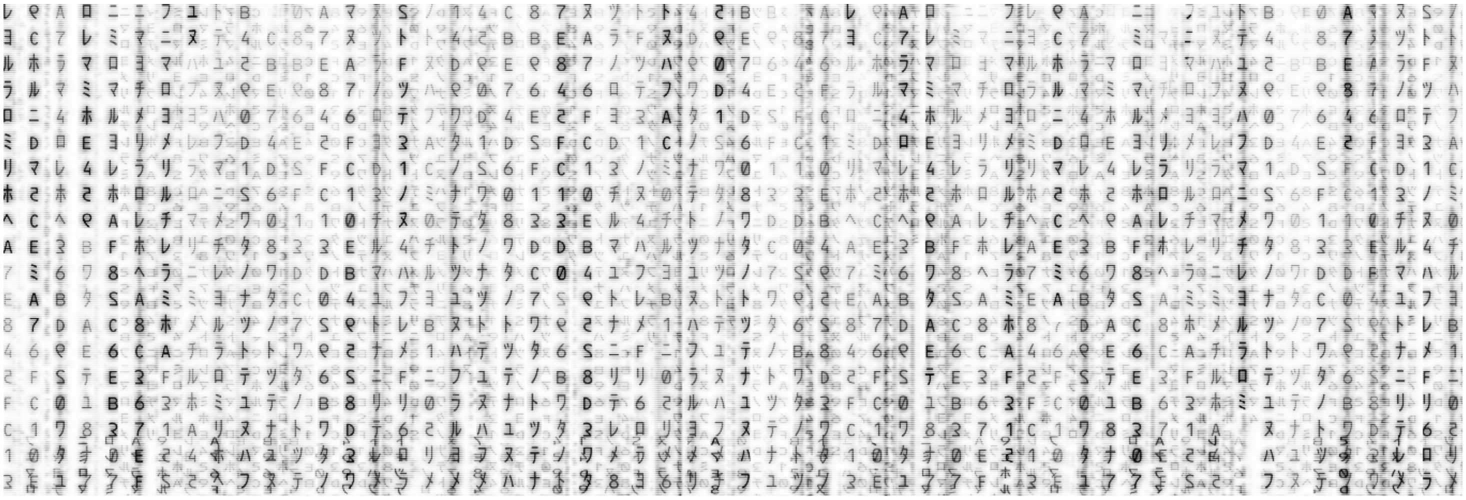
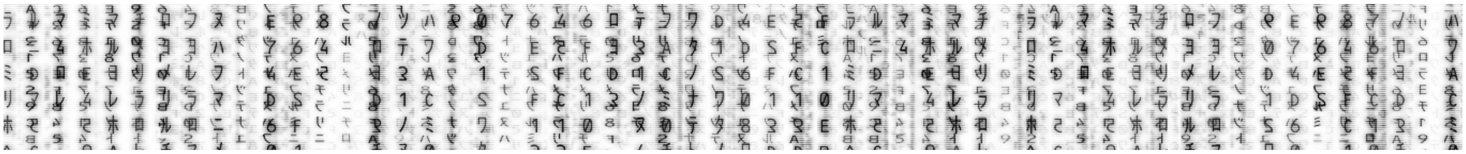
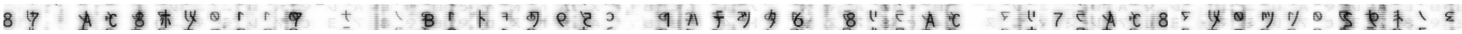


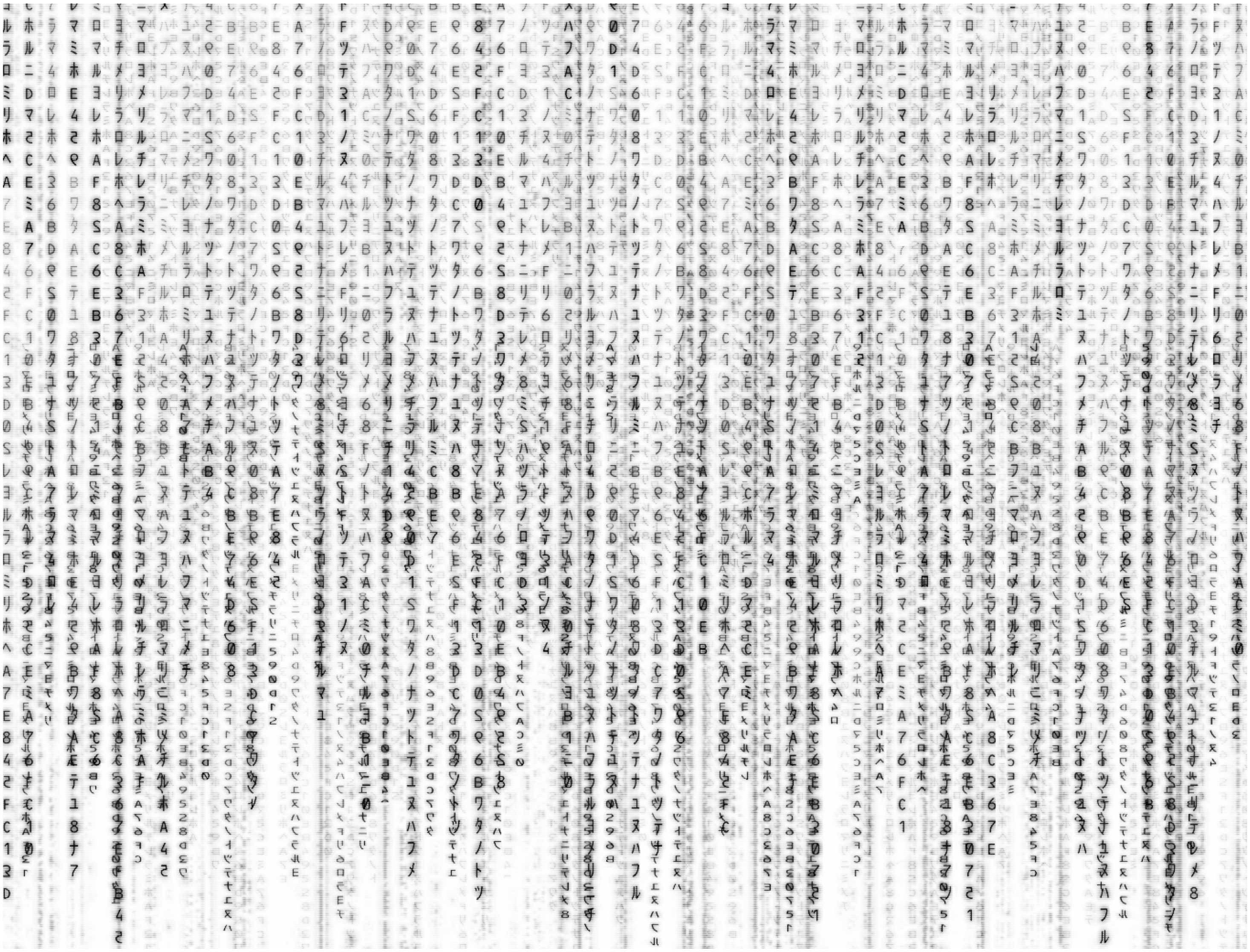
Applied Econometrics using Stata



Ricardo Perez-Truglia



Harvard University



Applied Econometrics using Stata[*]

Ricardo Nicolás Pérez Truglia[γ]

Department of Economics

Harvard University

Extremely early draft: March 2009

Diclaimer: Chapters 3, 4 and 6 are very incomplete and contain some paragraphs in Spanish.

Comments welcomed!

[*] Thanks to my Professors: Walter Sosa Escudero, Sebastian Galiani and Dale Jorgenson. These notes were originally written to be used in classes taught at Universidad de San Andres (with Martin Rossi). They are still extremely preliminary.

[γ] rtruglia@fas.harvard.edu | Harvard University, Department of Economics: Littauer Center, 1805 Cambridge Street, Cambridge, MA (02138).

Preface

The guides use Stata 9.1. They should work without changes in Stata 10 (or simply entering "version 9" at the beginning of the code), and with minor modifications they should work in older versions. We try not to include Stata output in the document in order to emphasize that you should execute the Stata commands while you are reading the Notes.

It is very important to solve the Problem Sets by your own. They are the most important pedagogical tools in this document. The objective of the Notes and the Problem Sets is to reproduce the Tables of published papers, so the reader can experience by herself the “cooking” of applied econometrics. In general Montecarlo experiments are not used for teaching econometrics. We think that is the main innovation of this document, since Montecarlo experiments are a very practical way to incorporate the intuition behind many econometric results.

You can find the databases in the website: ricardotruglia.com.ar. Most of the databases are from third parties, downloaded from the author’s original websites. We do not cover deeply the econometric theory, but we give a brief introduction, specially for those models that are not basic (e.g. ordered logit, program evaluation, kernel estimation, etc.).

If you feel that you need to refresh some of the econometric theory, we recommend Jeffrey Wooldridge’s “Econometric Analysis of Cross Section and Panel Data,” which covers almost every single topic in the notes (or Wooldridge’s “Introductory Econometrics” for the beggineers or for non-economists). It is the standard textbook in undergraduate and graduate courses in econometrics. Whenever a topic is not completely covered by Wooldridge’s book, we will give further references.

Stata makes applied econometrics extremely easy. However, this comes to a price: it is extremely easy to run a regression without understanding the true meaning of what you are doing. You should avoid the temptation to use a model without understanding it, and go to your favorite textbook first.

Please cite as: Perez Truglia, Ricardo Nicolas (2009), “Applied Econometrics using Stata.” Mimeo, Harvard University.

Index

Chapter 1

First steps in Stata

Introduction

The Stata screen is divided in 4 parts. In "review" you can see the last commands that have been executed. In "variables" you can see all the variables in the current database. In "results" you can see the commands' output. Finally, in the window "command" you can enter the commands.

Data input and output

Stata has its own data format with default extension ".dta". Reading and saving a Stata file are straightforward. If the filename is "file.dta" (and if it is located in the Stata's directory), the commands are:

```
. use file
```

```
. save file
```

Notice that if you don't specify an extension, Stata will assume that it is ".dta". If the file is not stored in the current directory (e.g. it is in the folder "c:\user\data"), then the complete path must be specified:

```
. use c:\user\data\file
```

Nevertheless, the easiest way to work is keeping all the files for a particular project in one directory, and then changing the "default" directory of Stata to that folder. For instance, if you are working in the folder "c:\user\data":

```
. cd c:\user\data
```

```
. use file
```

Insheet (importing from MS Excel)

There are two simple ways to transform an Excel database (or similar formats) into a Stata database. For instance, please create the following table in MS Excel:

name	account	money
John Doe	1001	55
Tom Twain	1002	182
Tim Besley	1003	-10
Louis Lane	1004	23

Save it as a text file (tab delimited or comma delimited) by selecting "File" and choosing "Save As", under the name "bank.txt" in the Stata folder. Notice that saving as "txt" (only text) you will eliminate all the information on formats, formulaes, etc. Then import the data to Stata using the command "insheet":

```
. insheet using bank.txt
```

To see whether the datas was imported succesfully, open the data browser:

```
. browse
```

Another (easier) way to import the data is by selecting the cells in MS Excel, and then select "Edit… Copy". First use the command "clear" if there is a database in memory. Then execute:

```
. edit
```

And a spreadsheet will appear (similar to what appeared with the command "browse", but now you can modify the data). Then click the right button on the editor and press "Paste".

The command "save" let us save the data in Stata format (".dta"). The option "replace" replaces the old version if the file already exists:

```
. save prueba.dta, replace
```

The command "clear" clears the current database:

```
. clear
```

Notice that if you want to open the database but there is one database in memory, Stata won’t proceed. The option "clear" can be used simultaneously with the command "use" to open a new database:

```
. use sales, clear
```

By default Stata dedicates 1mb of memory for loading the database. However, most databases demands more memory than just 1mb. If that is the case (before opening the file, and without a database in memory) we need to indicate how much space is needed (using the command "set mem"). For instance, if we need 5 megabytes:

```
. set mem 5m
```

Preserve and Restore

As you may have noticed, there is no "Undo and Redo" in Stata. But you can use "preserve" to "save provisionally" a database. Then, if you want to "undo", you can execute "restore" and go back to the previous state.

```
. preserve
```

```
. drop
```

Ups! You dropped all the observations. However, you can go back to the point where you executed "preserve":

```
. restore
```

Log-File

The log-files are useful to keep record of everything that appears in the "results" window. A log-file records both the history of commands and the history of outputs.

```
. log using test, replace
```

The option "replace" replace the existing file. When the session is finished, you must close the log-file:

```
. log close
```

You can open a log file using the Notepad or using the option "File… Log… View".

Do-File

The do-files are extremely useful in Stata. A do-file is an unformatted text file (ASCII) that contains a sequence of Stata commands. Stata interprets them exactly as if they were entered in the command window. Then you can save code lines and time when you want to repeat an entire piece of code with a minor modification.

Let's begin with a classis of programming: making Stata say “Hello”. The corresponding command is:

```
. display "Hello"
```

Then, if you want to create a do-file with the above code, simple open the do-file editor (you can even use the Notepad) and enter:

```
. display "Hello"
```

Remember always to add an "Enter" at the end! Then save the file as "hello.do" in the Stata fólder. If you want to execute the do-file, you must use the commmand "do":

```
. do hello
```

At the beginning you might find awkard that there is no "Undo and Redo" in Stata. But if you learn to use Stata properly, you will find that there is no need for such commands. You should plan your work in advance. For example, you know that you need to “prepare” the database and then run the regression. Then you write a do-file beginning with “use database” and followed by hundreds or even thousands of commands. Supposse that you finished the last regression and a colleague tells you that (say) you should have defined a variable in a different way. Then you simple go to the do-file, modify exactly that line and then press “run” and Stata will automatically generate all the Tables of your papers with the proper modifications. If this is your first time programming, you may find this procedure too complicated. I assure you that the advantages of automatization will be overwhelminghly beneficious for your work in the future; you only need to pay a small entrance cost at first.

You will understand the advantages of using do-files as soon as you begin working with them for the Problem Sets. Sometimes you want other people to understand or use your code, or maybe you know that you may need to use the code again some months (or even years) in the future. For that sake it is very useful to include comments in the do-files describing what you are doing in each lines. To insert an entire line of comments you must use an asterisk at the beginning of the line:

```
. * This is a comment, write whatever you want, just begin it with an asterisk
```

The "/* text */" comment delimiter has the advantage that it may be used in the middle of a line. What appears inside /* */ is ignored by Stata (it cannot be used in the command window; it will only work in a do-file). The "//" comment indicator may be used either at the beginning or at the end of a line:

```
. describe */ text /* var1-var10
```

```
. describe var1-var10 // text
```

Exercise 1.1: After finishing the first Chapter, open the do-file editor and create a do-file called "week1" to reproduce all the commands.

Help and search

As in every programming environment, the command "help" has a special importance, as it gives detailed information about the commands under consideration. For instance, write the following and press Enter:

```
. help summarize
```

A window with information on the command "summarize" will appear. If you want to do something, but you do not know which the right command is, you should use "search". For instance, if you want to find the mean of a variable, you may enter the following:

```
. search mean
```

Then just choose the appropriate option.

Commands

As indicated by the help command, there is a generic command structure for the mayority of the Stata commands.

```
[by varlist:] command [varlist] [=exp] [if exp] [in range]
[weight] [using filename] [, options]
```

For any given command, some of these components may not be available. In the help-file you may find links to information on each of the components:

[by varlist:] instructs Stata to repeat the command for each combination of values in the list of variables varlist. For instance, "by location" would repeat the command for the set of observations with each value for the variable "location".

[command] is the name of the command and can be abbreviated. For instance, the command "summarize" can be abbreviated as "sum" and the command "regress" can be abbreviated "reg".

[varlist] is the list of variables to which the command applies. There are some shortcuts. For example, instead of writing "var1, var2,…, var9" you can write "var*", or "var1-var9". Alternatively, if you are interested in listing the variables from "var1john" through "var9john" you can use simply "var?john" (as in the old DOS).

[=exp] is an expression.

[if exp] restricts the command to that subset of the observations that satisfies the logical expression "exp". For instance, "height>170" restricts the command to those observation with "height" greater than 170.

[in range] restricts the command to those observations whose indices lie in a particular range. The range of indices is specified using the syntax f/l (for “first to last”) where f and/or l may be replaced by numerical values if required, so that 5/12 means “fifth to twelfth” and f/10 means “first to tenth”. Negative numbers are used to count from the end, for example: list var in -10/l lists the last 10 observations.

[weight] allows the weighting of observations.

[using filename] specifies the filename to be used.

[options] are specific to the commands.

Version

Some commands change a lot from version to version. If you want to execute a code from a previous version (e.g. Versión 7.0), you can do so by using the command "version" at the beginning of the code:

```
. version 7.0
```

Data Management

Variables

In Stata there are two types of variables: string and numeric. Subsequently, each variable can be stored in a number of storage types (byte, int, long, float, and double for numeric variables and str1 to str80 for string variables).

If you have a string variable and you need to generate numeric codes, you may find useful the command "encode". For instance, consider a variable "name" that takes the value "Nick" for the observations belonging to Nick, "John" for the observations belonging to John, and so forth. Then you may find useful the following:

```
. encode name, gen(code)
```

A numeric code-variable will be generated (e.g. it takes the value 1 for the observations belonging to Nick, the value 2 for the observations belonging to John, and so forth).

Missing values in numeric variables are represented by dots. Some databases have other special characters for missing values, or maybe particular numbers (9, 66, 99, etc.). Missing value codes may be converted to missing values using the command "mvdecode". For instance, if the variable "gender" takes the value 9 if missing value, then enter:

```
. mvdecode gender, mv(9)
```

It will replace by dots all values of variable gender that equal 9.

Let's work

In the website (ricardotruglia.com.ar) you will find some databases that will be used throrought this notes. You should download all the databases and put them in the Stata folder (or whatever folder you want to work in).

Now we will use the database "russia.dta". It is a compilation of health, economic and welfare variables from the Russian Longitudinal Monitoring Survey for 2600+

individuals in 2000 (information at the official website: www.epc.unc.edu/projects/rlms). At the end of the first Problem Set you may find data definitions.

```
. use russia.dta, clear
```

Describe, list, browse and edit

In order to know basic information on the variables allocated in memmory, use:

```
. describe
```

You can see the names, storage types, display formats, labels (e.g. "For how many years" for the variable yr) and value labels (e.g. "Male" if gender==1 and "Female" if gender==0). If you want to see information for only some variables, simply list them:

```
. describe totexpr gender
```

You can list some observations. For instance, the following code list the gender and total expenditure for the first 10 observations:

```
. list gender totexpr in 1/10
```

The sequence "Ctrl+Alt+Break" interrupts the execution of a command (you can also use the red button with a white cross in the top toolbox).

If you want to see the database in a spreadsheet-fashion, you may try:

```
. browse
```

As you might notice, you cannot edit the information. But if you use the command "edit" you will be able to do so:

```
. edit
```

In a while we will study some commands meant to describe the data. But in the meantime you can try the command "summarize"(or "sum"):

```
. summarize
```

It gives means, standard deviations, maximum and minimum values for each variable. As always, if you want to see the information for a particular group of variables, then you simply have to list them:

```
. summarize gender belief
```

Sort

Use "sort" if you need to sort the observations according to one or more variables (numerically and/or alphabetically). Some commands require the data to be sorted first (such as "by" and "merge"). For instance, you can sort the individuals by their real expenditures and real income, in that order:

```
. sort totexpr totincm_r
```

Drop and keep

The command "drop" eliminates either variables or observations. For instance:

```
. drop belief
```

That drops the variable "belief". If you want to delete observations, you can enter:

```
. drop if gender==1
```

That will delete observations for males. Alternatively:

```
. drop in 1/10
```

That will delete the first ten observations. The command "keep" is just the opposite: it deletes everything but what you list. For instance, "keep if gender==1" would delete every observation for females.

Rename and label

The command "rename" change the names of variables:

```
. rename gender sex
```

The command "label" allows us to label variables, where we can indicate more precisely their descriptions:

```
. label variable ortho "=1 if individual professes Orthodoxy Religión, =0 otherwise"
```

Gen and replace

If you want to generate a dummy variable, you may find the following piece of code very useful:

```
. gen tall=(height>190)
```

If you put a condition in parenthesis, then the parenthesis takes the value 1 if the condition is true, and 0 otherwise. Then the above code line assign a 1 to the variable "tall" if "height">190cm, and zero otherwise. Logical expressions have numerical values, which can be immensely useful. But what if “height” were missing? Stata treats numeric missing values as higher than any other numeric value, so missing would certainly qualify as greater than 190, and any observation with “height” missing would be assigned 1 for this new variable.

We should then type:

```
. drop tall
. gen tall=(height>190) if height != .
```

That will assign 1 if height were greater than 190, 0 if height were not greater than 190, and missing if height were missing. You should be extremely careful with this kind of details. Checking and double-checking your database is about honesty, a key ingredient for either scientific or professional work.

The logical operators are "&" (and) and "|" (or), and the conditions may involve "==" (equal), "~=" or "!=" (different), ">" (greater), and ">=" (equal or greater). You can create variables as complex as you want:

```
. gen tall1 =((height>190 & gender==1) | (height>180 & gender==0)) if height != .
```

The command "replace" replaces values of existing variables. For instance, the following line of code put a 0 in "tall" if the individual is obese:

```
. replace tall=0 if obese==1
```

Simple transformations

You can generate variables using mathematical and statistical functions: cos, ln, log10, sqrt, max, min, floor, round, sum, etc. For instance, generate a variable containing the natural logarithm of the household total expenditure:

```
. gen lntotexpr=ln(totexpr)
```

Extended Generate

The command "egen" (extended generate) is useful when you need to create a variable that is the mean, meidan, standard deviations, etc. of an existing variable. For instance, I can create a variable that takes the mean life satisfaction over the entire sample:

```
. egen mean_satlif = sum(satlif)
```

Or I can create variable that takes the mean life satisfaction over their geographical sites:

```
. egen site_mean_satlif =mean(satlif), by(site)
```

The command "egen" have other useful option. We can use the option "group" to group the population according to different combinations of some variables. For instance, I would like to identify the individuals according to if they "smokes" and if they are "obese". As both categories are binary, the command "group" will generate four possible categories (people that smoke and are obese, people that don't smoke and are obese, people that smoke and are not obese, and people that don't smoke and are not obese):

```
. egen so=group(smokes obese)
```

Let's see the results:

```
. browse smokes obese so
```

Label values

We can put labels to the different values of "so":

```
. label values so solabel
```

```
. label define solabel 1 "Smokes:NO Obese:NO" 2 "Smokes:NO Obese:YES" 3 "Smokes:YES Obese:NO" 4 "Smokes:YES Obese:YES"
```

You may see the results using the command "tabulate":

```
. tab so
```

If you want to modify the value labels, you need to drop the old label and create a new one:

```
. label drop solabel
```

```
. label values so solabel
```

```
. label define solabel 1 "Skinny and clean" 2 "Clean but fatty" 3 "Skinny smoker " 4 "Fatty smoker"
```

Let's see the results again:

```
. tab so
```

By and bysort

The option "by" indicates that the command must be run for many groups of variables. Some commands use by as "by xx: command", and other command use it as in "command, by xx". Before running "by" the data must be sorted by the variable after the "by":

```
. sort geo
```

```
. by geo: count if gender==1
```

The command "count if gender==1" counts the number of observations for men. Then, adding "by geo" make Stata count the number of men inside each geographical area (for geo==1, geo==2 and geo==3). If you find uncomfortable to sort the data first, you can use "bysort" directly:

```
. bysort geo: count if gender==1
```

Append

Using "append" you can add observations to database using another database (i.e. you can "append" one to another). In the database "week1_2" there are 100 observations on individuals from the Russian Household Survey, and in the database "week1_3" there are 100 additional observations. You can browse both databases:

```
. use week1_2, clear
```

```
. browse
```

```
. use week1_3, clear
```

```
. browse
```

We want to "paste" one database below the other. You must open the first database:

```
. use week1_2, clear
```

And then using "append" you add the observations of the second database:

```
. append using week1_3
```

Finally, you can see the results:

```
. browse
```

Collapse

The command "collapse" generates a "smaller" database contaning the means, sums, standard deviations, etc. of the original dataset.

```
. use russia.dta, clear
```

We can take the means of life satisfaction ("satlif") and economic satisfaction ("satecc") within geographical sites ("site"):

```
. collapse (mean) satlif satecc, by(site round)
```

You can now see the final product:

```
. describe
```

```
. browse
```

Instead of having thousands of observations on several variables now we have one only observation per geographical site (the mean) for only two variables (life satisfaction and economic satisfaction).

If one variable has more missing values than the other, then the means, standard deviations, etc. will be based in different sample sizes. If you want to avoid that, you can use the option "cw" (casewise deletion), which forces Stata to eliminate any observations without data for every variable involved in "collapse".

Reshape

Suppose we have j=1…J variables on each individual i=1…I. This information may be viewed in two ways. Firstly, each variable j may be represented by a variable xj and the individual identifier may be a variable "individual_id". However, we may need one single response vector containing the responses for all variables for all subjects. These two "data shapes" are called wide and long, respectively.

For instance, the database "week1_3" has a wide shape:

```
. use week1_3, clear
```

name	money1990	money1991	money1992
John Doe	10	12	15
Tom Twain	3	7	1
Tim Besley	25	20	18
Louis Lane	14	14	11

Use the following code to transform the database into a long shape:

```
. reshape long money, i(name) j(month)
```

```
. browse
```

You can also reshape it again back to wide:

name	year	money
John Doe	1990	10
John Doe	1991	12
John Doe	1992	15
Louis Lane	1990	14
Louis Lane	1991	14
Louis Lane	1992	11
Tim Besley	1990	25
Tim Besley	1991	20
Tim Besley	1992	18
Tom Twain	1990	3
Tom Twain	1991	7
Tom Twain	1992	1

```
. reshape wide money, i(name) j(year)
```

Merge

If you want to join two files for the same individuals but with different sets of variables, then you must use the command "merge":

```
. merge using filename
```

Instead of adding new observations (as in "append"), you add new variables joining corresponding observations from the dataset currently in memory (called the master dataset) with those from Stata-format datasets stored as "filename" (called the using dataset). As always, if filename is specified without an extension, then ".dta" is assumed.

A "one-to-one merge" simply "pastes" both datasets side by side. If you want Stata to match each observation in the master dataset with the corresponding observation in the using dataset (e.g. observations for the same individual), you must perform a "match merge". You should have a key variable (e.g. the individuals' names). After merging the datasets, Stata will create a variable called "_merge" to indicate the result of the merging for each observation: 1 for the observations from the master dataset that did not match with the using dataset; 2 for the observations from the using dataset that did not match with the master dataset; 3 for the successful matches. Enter "help merge" for further details.

Exercise 1.2: Generate a do-file to carry out the following: using the dataset "russia.dta", generate an id-variable for the observations (individuals); divide it in two separate databases, each one with different sets of variables; merge those datasets back together.

Descriptive Analysis

We will see some commands to describe the database, which will involve creating complex tables and figures. Even if you want to do “hardcore” econometric, you should learn them for various reasons. In the first place, you will need to give an introduction and show the motivation of your paper/report. In the Problem Sets we will ask you to reproduce a lot of tables and figures from published papers, which we think is a very nice training.

It is also important to show descriptive statistics for letting the reader have a quantitative idea of the economic meaning of the parameters estimated by the econometric model. A coefficient of 0.1 has a different meaning if the standard deviation of the variable under consideration is 10 or 100. You cannot possible interpret all the output from your regressions, but you can provide as much information as possible to allow the readers to interpret the econometric results by themselves.

Using descriptive statistics you can also check the internal consistency of the data. This is a task as tedious as important. Can an individual be 453 years old? Can an individual consume a negative number of cars? Are you using a variable filled mostly by missing values that is “killing” the sample size of your regressions?

Finally, you can also provide evidence in favor of the external validity of the model. For example: are the individuals in the database representative of the entire population? Can you compare your data to the data in other studies? In other countries? An econometric application without external validity is dead letter. You don’t want to spend a lot of time and effort in a beautiful application and don’t pay attention to this aspect.

Summarize and tabstat

The command "summarize" shows basic descriptive statistics:

```
. summarize totexpr
```

The option "detail" adds different percentiles to the table:

```
. summarize totexpr, detail
```

The option "pweight" weights the observations by using the inverse of the probability of entering the sample:

```
. summarize totexpr [pweight=inwgt], detail
```

The command "tabstat" shows specific statistics:

```
. tabstat hswrk, stats(mean range)
```

Some arguments for "stats(·)" are: mean, count, sum, max, min, range (max-min), sd, var, etc. The option "by(varname)" makes "tabstat" build a table with descriptive statistics for each value of "varname":

```
. tabstat totexpr, stats(mean range) by(gender)
```

The command "ci" makes an confidence interval for the mean of a variable at a given statistical level of confidence:

```
. ci totexpr, level(95)
```

Tabulate, tab1 and table

The command "tabulate" produces frequency tables for numeric variables:

```
. tabulate econrk
```

Using "tab2" you can produce two-way tables of frequency counts, along with various measures of association:

```
. tabulate econrk powrnk
```

Some option for "tabulate" are cell (shows percentages for each cell), column (show percentages by columns), missing (includes missing values as a separate value).

You can create a set of dummies starting from a discrete variable. For example, the variable "belief" takes the values from 1 to 5. Enter:

```
. tabulate belief, gen(belief)
```

That generates 5 dummies: belief1 (that takes value 1 if belief=1 and 0 otherwise) to belief5 (that takes value 1 if belief=5 and 0 otherwise).

You can also produce three-way table of frequency, using the option "by":

```
. bysort gender: tabulate econrk powrnk
```

The command "table" is to "tabulate" what "tabstats" is to "summarize". You can ask for specific information for the tables of frequency:

```
. table econrk, contents(freq)
```

```
. table econrk powrnk, contents(freq)
```

```
. table econrk powrnk gender, contents(freq)
```

And you can also create "super-tables":

```
. table econrk powrnk gender, contents(freq) by(smokes)
```

Some option for "contents(·)" are: "freq" (frequency); "mean varname" (mean of varname), "sd varname" (standard deviation of varname), and so on.

Assert

We can detect errors using the assert command. For instace, we know that gender can only take values 0 or 1 (besided missing values):

```
. assert gender==0| gender==1| gender==.
```

Scatterplot

You graph a "scatterplot" entering "graph twoway scatter" followed by the y-variable and the x-variable, respectively:

```
. graph twoway scatter htself height, by(gender)
```

If you enter more than one x-variable, then Stata will generate in the same graph two scaterplots (in different colors) with the two combinations between the y-variable and both x-variables:

```
. graph twoway scatter waistc htself height, by(gender)
```

There are many options common to all graphs (i.e. options for "graph twoway"). For instance, the option "title(·)" creates a title for the graph, and there are many additional labels to be defined:

```
. graph twoway scatter htself height, title(This is the title) b2(This is bottom 2) l1(This is left 1) l2(This is left 2) t1(This is top 1) t2(This is top 2) r1(This is right 1) r2(This is right 2)
```

Enter "search axis" or "search title" for further details. You will discover much of those options when starting to work with graphs seriously. Enter "help graph twoway" to see other two-way graphs available.

Histogram

Use the command "hist":

```
. hist height, title(Heights) by(gender)
```

Nonetheless, the histograms are very sensitive to the parameters used for their construction. Thus, you are strongly encouraged to (carefully) provide them by yourselves: "bin(#)" (number of bins), "width(#)" (width of bins) and "start(#)" (lower limit of first bin). The option "discrete" indicates to Stata that the variable under consideration in categorical (it is very important to do so):

```
. hist belief, title(Trust in God) discrete
```

If you didn't, your graph would look like this:

```
. hist belief, title(Trust in God)
```

The option "normal" adds a normal density to the graph, and the option "freq" shows the frequencies instead of the percentages:

```
. hist height, width(2) start(140) norm freq
```

More and more

```
. graph box height, by(site)
```

In a vertical box plot, the y axis is numerical (height), and the x axis is categorical (site). If you want an horizontal box plot, you should use “hbox” instead. The box itself contains the middle 50% of the data. The upper edge of the box indicates the 75th percentile of the data set, and the lower hinge indicates the 25th percentile. The range of the middle two quartiles is known as the inter-quartile range. The ends of the vertical lines indicate the minimum and maximum data values, unless outliers are present in which case they extend to a maximum of 1.5 times the inter-quartile range. The points outside the ends are outliers or suspected outliers.

The diagnostic plots allows you to check visually whether the data is distributed with a particular distribution. For example, “symplot” help you see whether the data is symmetric:

```
. symplot height if gender==1
```

If the observations form a line, then the data is approximately symmetric. And “qnorm” plots the quantiles of the variable against the quantiles of the normal distribution (Q-Q plot):

```
. qnorm height if gender==1
```

If the observations form a line (as this is the case), then the data is approximately normal. You can use the same principle (Q-Q plot) to see check whether you data fits any arbitrary distribution using “qqplot”.

Using “help” and “search” you should be able to find all the graphical applications in traditional softwares like MS Excel. For instance, the classic pie chart:

```
. graph pie totexpr, over(site)
```

Some Programming in Stata

Foreach and forvalues

Stata has the typical commands from standard programming. For instance, sometimes you have to repeat the same piece of code several times, which usually would imply a lot of typing. The command "for" is very useful in those ocations. Suppose you want to run several mean-tests for smokes. You may write the test several times:

```
. ttest smokes==0.1

. ttest smokes==0.15

. ttest smokes==0.2

. ttest smokes==0.25

. ttest smokes==0.3

. ttest smokes==0.35

. ttest smokes==0.4

. ttest smokes==0.45

. ttest smokes==0.5
```

Or you can use the "for" syntax:

```
. forvalues k = 0.1(0.05)0.5 {
. ttest smokes==`k'
. }
```

The `k' in the "for" syntax represents the values of the parameter within the given range (from 0.1 to 0.5, taking 0.05-steps). You can also use the "for" syntax for a list of names (using "foreach"):

```
. foreach file in file01.dta file02.dta file03.dta {
.         use `file', clear
.         sum smokes
. }
```

Ado-files

We can make Stata execute a list of command using do-files. For instance, we made Stata say hello running the do-file "hello.do". Alternatively, you can create a program to run a code. For instance, let's create a program called "hello.ado" to make Stata say hello:

```
. program define hello
          1. display “Hello”
          2. end
```

Nothing happened, because we have to run the new program first:

```
. hello
```

Some Macros

The "macros" stores many useful values. For instance, the macro "_n" stores the number of the current observation. For instance, the following line of code lists the first nine observations:

```
. list totexpr if _n<10
```

On the other hand, the macro "_N" stores the total number of observations. For instance, the following line of code shows the last observation within each geographical site:

```
. bysort site: list totexpr if _n==_N
```

You can create a variable based on the macros. For instance:

```
. clear
. set obs 100
```

```
. generate index = _n
```

```
. browse
```

Generate a variable "x" with random numbers distributed uniformly between zero and one:

```
. gen x=uniform()
```

Then, you can generate lagged values:

```
. generate xlag = x[_n-1]
```

```
. browse
```

It may be interesting to obtain lagged values in a panel of individuals. For instance, see the following database:

```
. use week1_5, clear
```

```
. browse
```

There are 5 (already sorted) observations for 5 different individuals. You must generate the lagged values for "x" entering:

```
. bysort N: generate xlag = x[_n-1]
```

```
. browse
```

Please convince yourself that you cannot use "generate xlag = x[_n-1]".

Return and ereturn

After "summarize" and many other commands, the results are stored in "macros". You can retrieve them entering "return":

```
. sum smokes
```

```
. return list
```

For example, the mean for "smokes" has been stored in "r(mean)". We can use that information in the future:

```
. display r(mean)
```

```
. gen mean_smokes = r(mean)
```

For some commands you must enter "eretunr list" (estimation commands) or "sreturn list" (commands that assist in parsing).

Quietly, capture and more off

The capture prefix makes Stata continue running the do-file even if the command throws an error. On the other hand, the prefix quietly eliminates all output but error messages. For instance, if you want to retrieve the mean for smokes, you may enter:

```
. quietly: sum smokes

. quietly: return list

. display display "The mean for smokes is "r(mean)
```

Exercise 1.3: Try each one of the following in a do-file, and comment the results:

```
. capture: blablabla // This is wrong

. capture: sum smokes // This is right

. quietly: blablabla

. quietly: sum smokes
```

Set more

The command "set more off" causes all the output to scroll past automatically instead of waiting for the user to scroll through it manually. The command "set more on" reverses it.

Problem Set #1

Cruces et al. (2009) shows that people have a very biased perception of their own wealth relative to the general population. Ravallion et al. (1999) introduced this discussion using a very illustrative table (Table 1) that compares the objective and subjective measures of relative income. The objective of the first part of the problem set is to reproduce a naïve version of that Table:

1. Generate a variable que that assing to each individual a number from 1 (lowest) to 9 (highest) according to their position in the distribution of expenditures in the population. Present a Table comparing the the objective ranking with the subjective ranking given by the variable “economic rank”. Comment briefly the results. Then generate one table for each of the 3 different geographical regions given by variable “geo”. Do you think this second exercise is more reasonable? Would you use expenditures instead of income? Reproduce one of the above tables using income and comment.

2. Make tables comparing subjective economic rank, power rank and respect rank. Present the results on a separate basis by gender. Comment if you find any remarkable difference.

3. Generate a variable about the "over-reporting" of relative income (i.e. subjective rank minus objective rank). Generate a variable measuring the "over-reporting" of height. Present a graph and report the correlation. Test if the correlation is significant.

4. Read the description of the variables in the dataset carefully. Play with the data using the commands we have seen so far. Present a creative graph or table showing a pattern that you found interesting.

Some *Data Definitions* you might find useful:

Satisfaction with Life: "To What extent are you satisfied with your life in general at the present time? [Fully satisfied 5] [Rather satisfied 4] […] [Not at all satisfied 1]".

Satisfaction with Economic Condition: "Tell me, please, how satisfied are you with your economic conditions at the present time? [Fully satisfied 5] […] [Not at all satisfied 1]".

Health Self-Evaluation: derived "Tell me, please, how would you evaluate your health? It is: [Very good 5] [Good 4] [Average 3] [Bad 2] [Very bad 1]".

Hospitalized last 3 months: "Have you been in the hospital in the last three months? [No 0] [Yes 1]".

Economic Rank Ladder (econrk): "Now, please, imagine a 9-step ladder where on the bottom, the first step, stand the poorest people, and on the highest step, the ninth, stand the rich. On which step are you today? [Lowest 1] […] [Highest 9]".

Power Rank Ladder (powrnk): "And now, please, imagine a 9-step ladder where on the bottom, the first step, stand people who are completely without rights, and on the highest step, the ninth, stand those who have a lot of power. On which step are you? [Lowest 1] […] [Highest 9]".

Respect Rank Ladder: "And now, another 9-step ladder where on the lowest step are people who are absolutely not respected, and on the highest step stand those who are very respected. On which step of this ladder are you? [Lowest 1] […] [Highest 9]".

Chapter 2
Least Squares

A very brief introduction to OLS

We have two different notations, sometimes called vector and matrix notations. The vector notation is the following: $y_i = \beta X_i + \varepsilon_i \quad \forall i = 1,2,...,N$. The scalar y_i is the *i*-th observation of the dependent variable (or equivalently, the value of the dependent variable for the *i*-th individual). The vector x_i is the *Kx1* vector of *i*-th observation of the k's independent variable

The linear model is: $y = X\beta + \varepsilon$. The vector y is *Nx1* and is obtained by "stacking" the N observation of the dependent variable. The matrix X is *KxN* and is composed by the K "independent variables," also known as "covariates" or "explanatory variables". The constant (if there is one) is a variable that always take the value 1. Finally, betha is the Kx1 vector of parameters, one for each independent variable. Notice that it is linear in parameters, but not necessarily linear in variables. For example, we will frequently include variables like age^squared, log(income), etc.

If betha is the estimator, then the residuals are e=y-Xb, and then the sum of squared residuals is:

To be completed.

We want to minimize this sum:

To be completed.

Solving for b in the FOC we get: b=… To solve for b we need a "rank condition": X'X has to be nonsingular (or equivalently, its inverse must exist). One necessary condition for this is that X is full rank: that is, there cannot be perfect multicollinearity among the explanatory variables. One tipical case is dummies: To be completed.

Command Regress

Let's begin with an example. We will use the database "russia.dta" we used in the first chapter and in the first Problem Set (where you may find data definitions).

```
. use russia, clear
```

Suppose that you want to explain the health Self-Evaluation indexes ("evalhl", where greater values corresponds to people that declare themselves as healthier) using the following variables:

```
. reg evalhl monage obese smokes
```

Notice that Stata automatically adds a constant. If you wanted to exclude it, you would have to enter "nocons" as option:

```
. reg evalhl monage obese smokes, nocons
```

As any other command in Stata, "regress" can be applied to a subset of the observations. Suppose you want to run two separate regressions, one for males and the other for females, respectively:

```
. reg evalhl monage obese smokes if gender==1
```

```
. reg evalhl monage obese smokes if gender==0
```

There are many options for "regress" (enter "help regress"). Possibly the most important is the option "robust", which uses a robust estimate for the variances-and-covariances matrix:

```
. reg evalhl monage obese smokes if gender==0 & round==9, robust
```

Besides "robust", there are other options regarding the estimation of standard errors: "cluster()" (adjusts standard errors for intra-group correlation) and "bootstrap" (Bootstrap estimation of standard errors). They will be covered in future Chapters.

You will sometimes find very tedious to copy and paste the list of control variables in each regression you run (you may run hundreds of regressions for a paper). An easy and elegant way to save some time and space is to define at the beginning of the code the list(s) of control variables:

```
. global controls monage obese smokes
```

And then simply enter “$controls” each time that you need the list of control variables:

```
. reg evalhl $controls
```

Testing

After each estimation Stata automatically provides t-tests (for linear regressions) or z-tests (for nonlinear models) of the null hypothesis whether the coefficients are zero. Notwithstanding, other hypotheses on the coefficients can be tested.

For instance, using the command "test" (after the regression was run) you can test whether the effect of being obese equals -0.05:

```
. reg evalhl monage obese smokes gender, robust
```

```
. test obese=-0.05
```

We can test hypotheses that involve more than one variable. For instance, we can test if the coefficients on "smokes" and "obese" are both null, or if the sum of the coefficients on "obese" and "gender" equals 1:

```
. test smokes obese
```

```
. test obese + gender == 1
```

The next section will cover these subjects extensively.

Interpreting coefficients

To be completed.

Predicted values and residuals

After every estimation command (e.g. reg, logit, probit) some "predicted" values (fitted values, residuals, etc.) can be stored in a new variable using the command "predict". In our example:

```
. reg evalhl monage obese smokes gender, robust
```

```
. predict evalhl_hat
```

Enter "browse evalhl yhat" to appreciate the fit of the model. Alternately, the residual can be obtained:

```
. predict res, residual
```

Notice that the residual is simply:

```
. gen res_alternative = evalhl - evalhl_hat
```

The relationship between the actual values, the predicted values and the residuals can be shown in a scatter plot:

```
. graph twoway scatter evalhl_hat res evalhl
```

Goodness of fit

To be completed.

Saturated explanatory variables

To be completed.

Ommited variable bias

To be completed.

Simultaneous causality

To be completed.

Measurement Error

To be completed.

Bad covariates

To be completed.

Multicollinearity

To be completed.

Partial regression plot

The command "avplot" graphs the relationship between the dependent variable and one of the explanatory variables conditional on the rest of the regressors. It is very useful to identify outliers. For instance, we will collapse the database to obtain the means for some variables within each geographical site (enter "tab site" to see the geographical distribution of people). We will use "preserve and restore" to avoid loading the database again:

```
. preserve
```

```
. collapse (mean) satlif totexpr satecc powrnk, by(site)
```

We run a regression of life satisfaction on three variables:

```
. reg satlif totexpr satecc powrnk, robust
```

Finally, using the command "avplot", we show the partial relationship between "satecc" and "satlif", identifying each geographical site:

```
. avplot satecc, mlabel(site)
```

```
. restore
```

"By hand" regression

Since we will use mathematic elements, we need to maximize the room reserved for those objects:

```
. set matsize 800
```

This is the maximum for Stata 9.1. If you are using Stata 10 you can set matsize up to 11,000. Recall the OLS estimator:

$$\beta = (X'X)^{-1} X'Y$$

Where *X* is *N* (number of observations) by *K* (number of covariates, including the constant) and *Y* is *N* by 1. As Stata does not allow for more than 800 rows or columns for matrices, it would be impossible to work directly with *X* or *Y* (as they have +2600 rows). But there is a rather simple trick: the command "mat accum". It executes an intermediate step, *X'X*, which creates a "small" matrix (*K* by *K*). First we have to eliminate the observations with missing values in the variables that will be included in the model (can you tell why?):

```
. drop if evalhl==. | monage==. | obese==. | smokes==.
```

Then we run the regression in the "traditional" way:

```
. reg evalhl monage obese smokes
```

Let's begin with calculating *X'X* and storing it in the matrix "XpX":

```
. mat accum XpX = monage obese smokes
```

You can see the result entering:

```
. mat list XpX
```

Calculate *X'Y* and store it in the matrix "XpY":

```
. mat vecaccum YpX = evalhl monage obese smokes
```

Then transpose it (using an aphostrophe) to obtain *X'Y*:

```
. mat XpY = YpX'
```

```
. mat list XpY
```

Finally, we can get β using the above formula:

```
. matrix beta = invsym(XpX)*XpY
```

```
. mat list beta
```

Standard errors “by hand”

To be completed.

Regressions' output

You need to show the regressions in tables similar to those utilized in most economics papers (e.g. one column per specification, and standard errors in parentheses). Use then the command "outreg". As it is not a "default" command in Stata, you need to install it first:

```
. search outreg
```

Select "sg97.3" and then press "click here to install".

We run a regression:

```
. reg evalhl monage obese smokes satlif totexpr
```

And then we save the output in the file "regresion.out":

```
. outreg using regresion, replace
```

The option "replace" indicates to Stata that if the file "regresion.out" exists, then it must be replaced. Go to the Stata folder and open "regresion.out" (using MS Excel or MS Word).

By default "outreg" shown in parentheses the t-values and puts an asterisk if the coefficient is significative at the 5%, and two asterisks if it is significative at the 1%. We can ask for the standard errors in parenthesis, and one asterisk if the coefficient is significative at the 10%, two asterisks if it is significative at the 5%, and three asterisks if it is significative at the 1% (options "se" and "3aster" respectively):

```
. outreg using regresion, se 3aster replace
```

The command can also show various regressions in the same table (as columns). We must add regression outputs using the option "append" instead of "replace". Let's run three regressions: one with only "monage" and "obese" as explanatory variables; another with only "smokes", "satlif" y "totexpr"; and finally a regression with all them.

```
. reg evalhl monage obese
. outreg using regresion, se 3aster replace

. reg evalhl smokes satlif totexpr

. outreg using regresion, se 3aster append

. reg evalhl monage obese smokes satlif totexpr

. outreg using regresion, se 3aster append
```

Open using MS Excel the file "regresion.out" and see the results. If you do not want to show the coefficient for a set of dummy variables, you will find the command "areg" very useful.

Frisch-Waugh Theorem

This is an extremely useful result. We will illustrate it with an example to help you incorporate the intuition more easily. Consider the partitioned model:

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

Where both X_1 and X_2 may contain more than one variable. The Frisch-Waugh theorem says:

$$\hat{\beta}_2^{OLS} = \left(\tilde{X}_2'\tilde{X}_2\right)^{-1}\tilde{X}_2'y$$

Which is an OLS regression of y on $\tilde{X}_2$, where $\tilde{X}_2$ are the residuals from the OLS regression of X_2 on X_1:

$$\tilde{X}_2 = X_2 - X_1\hat{\theta} = \left(I - X_1\left(X_1'X_1\right)^{-1}X_1'\right)X_2$$

Intuitively, $\tilde{X}_2$ is the variability in X_2 that cannot be explained by X_1.

One way to use the Theorem is to obtain $\hat{\beta}_2^{OLS}$ in two steps: i. Regress X_2 on X_1 and save the residuals $\tilde{X}_2$; ii. Regress y on $\tilde{X}_2$, and that coefficient is $\hat{\beta}_2^{OLS}$.

Let's calculate the coefficient on "monage" in this fashion. First, we have to make sure that the observations used in the two steps are the same than those in the regular one-step regression. For this sake we will eliminate observations with missing values in the variables of interest:

```
. reg evalhl monage obese smokes alclmo gender belief

. drop if evalhl==. | monage==. | obese==. | smokes==. | alclmo==. | gender==. | belief==.

. reg monage obese smokes alclmo gender belief

. predict res, residuals
. reg evalhl res, nocons
```

Notice that the coefficient for “monage” is exactly equal. The same is not true for the standard errors: among other things, Stata is not taking into consideration the degrees-of-freedom adjustment corresponding to the first-stage variables. Even though you will not use it at the beginning, this theorem will prove very useful in the future.

In the simple bivariate case where the regression vector includes only the single regressor, xi, and a constant, the OLS coefficient is given by:

To be completed.

Creating Random Data and Random Samples

Now we will create an artificial database to grasp better the basic properties of OLS. Before starting this exercise, you must learn how to create random samples. To "generate" observations from nothing, you must use "set obs":

```
. set obs 100
```

If you want to generate a variable "number" containing random numbers distributed uniformly between 0 and 1, enter:

```
. gen number=uniform()
```

The function "uniform()" returns uniformly distributed pseudorandom numbers on the interval [0,1). Entering either "browse" or "list" you will be able to see the generated observations. If you want to generate uniformly distributed pseudorandom numbers on the interval [a,b), you must enter "gen var=uniform()*(b-a)+a". For instance, to generate numbers on the interval [3,7):

```
. gen number_1=uniform()*4+3
```

Check the desired properties (max, min and mean) entering "sum number_1". If you want to generate integers, then you must multiply the numbers by a multiple of ten and then truncate them towards zero (using "int"). For instance, to create integers between 0 and 9:

```
. gen number_2=int(uniform()*10)
```

Use "list" to see the results. Alternatively, if you want to generate random numbers with different distributions, you have to evaluate the inverse density function at "uniform()". For instance, using "invnormal(uniform())" returns normally distributed random numbers with mean 0 and standard deviation 1. Sum and multiply accordingly in order to obtain different means and deviations.

Furthermore, if you want to repeat the same "succession" of pseudo-random numbers each time that you execute a do-file, then you must initialize the "seed":

```
. set seed 339487731
```

For instance, generate ten random numbers twice:

```
. clear
. set obs 10
. gen n=uniform()
. gen n1= uniform()
. list
```

And then repeat the process using the same seed:

```
. clear
. set obs 10
. set seed 11223344
. gen n=uniform()
. set seed 11223344
. gen n1= uniform()
. list
```

To be completed.

A “fictional” example

We will create values for some variables, using the "actual" values of the linear parameters involved. Then we will try to retrieve those parameters using OLS, what will let us experiment with some basic properties.

Let's generate i.i.d. data on wages, education, intelligence, two explanatory variables uncorrelated with education and intelligence but correlated with wages (a and b), and finally a variable (c) totally uncorrelated with all the former variables.

```
. clear
. set obs 100
```

The variable intelligence will be the IQ of the individuals. IQs have approximately a normal distribution centered in 100 with a standard deviation of 20:

```
. gen intelligence=int(invnormal(uniform())*20+100)
```

Notice that we have truncated the decimal part of the numbers. Since more intelligent people is expected to study more (see the original model of Spence on the signaling purpose of education), the years of education will be equal to the intelligence (over 10) plus a normally distributed noise with mean 0 and deviation 2. Finally, we will keep only the integer part of the numbers:

```
. gen education=int(intelligence/10+invnormal(uniform())*2)
```

I will stop repeating "enter browse to see the results". Feel free to do so whenever you want. Variable a (b) will be normally distributed with mean 10 (5) and standard deviation 2 (1). Variable "c" will be normally distributed with mean 15 and standard deviation 3.

```
. gen a=int(invnormal(uniform())*2+10)
```

```
. gen b=int(invnormal(uniform())*1+5)
```

```
. gen c=int(invnormal(uniform())*3+15)
```

Finally, the unobserved error term "u" will be normally distributed with mean 7 and standard deviation 1:

```
. gen u=int(invnormal(uniform())*1+7)
```

Wages will be the result of "intelligence" multiplied by 3, plus variables "a" and "b" multiplied by 1 and 2 respectively, plus the unobserved error term "u":

$$wage_i = 3\cdot intelligence_i + 1\cdot a_i + 2\cdot b_i + u_i$$

```
. gen wage=3*intelligence+a+2*b+u
```

Let’s estimate the "true" model (a.k.a. data generating process). Remember that the command for OLS is "reg" followed by the dependent variable and then the list of explanatory variables. We will include the option "robust", which indicates the use of robust variance estimates:

```
. reg wage intelligence a b, robust
```

The estimated coefficients are near the true values. Notice that "education" does not "affect" wages. Then, if we included "education" and "intelligence" in the regression, then the former should not appear with a significative coefficient:

```
. reg wage education intelligence, robust
```

Notwithstanding, education is correlated with intelligence. Thus, if we forgot to include "intelligence" then the coefficient on "education" would be different from zero at a reasonable level of confidence:

```
. reg wage education, robust
```

The reason is that in the last equation "intelligence" is in the error term (because it "determines" wages but it is not included in the regression), and "intelligence" is correlated with "education". Thus, the orthogonality condition is not satisfied, and the estimator is not biased nor consistent.

Let's see that the exclusion of "a" and "b" does not violate the exogeneity condition. Since "intelligence" is not correlated with "a" and "b", its coefficient should remain consistent and unbiased:

```
. reg wage intelligence, robust
```

Nonetheless, including "a" and "b" should decrease the standard deviation of the coefficient on "intelligence", because there is less variability in the error term:

```
. reg wage intelligence a b, robust
```

Conversely, due to their independence, including "a" and "b" but excluding "intelligence" should not affect the consistence of the coefficients on the former:

```
. reg wage a b, robust
```

Finally, let's see the effect of including an "irrelevant" variable ("c") in the "true" equation:

```
. reg wage intelligence a b c, robust

. reg wage intelligence a b, robust
```

Compared to the "right" regression, the loss of one degree-of-freedom is irrelevant.

Taking advantages of do-files and macros

We can create a do-file including all the previous exercise:

```
. clear

. set obs 100

. gen intelligence=int(invnormal(uniform())*20+100)

. gen education=int(intelligence/10+invnormal(uniform())*2)

. gen a=int(invnormal(uniform())*2+10)

. gen b=int(invnormal(uniform())*1+5)

. gen c=int(invnormal(uniform())*3+15)

. gen u=int(invnormal(uniform())*1+7)

. gen wage=3*intelligence+a+2*b+u
. reg wage intelligence a b, robust
. (…)
```

You can also set a seed in order to keep your results "tractable".

Conditional Expectation Function

To be completed.

Appendix: Linear algebra review

To be completed.

Problem Set #2

1. If you want to learn applied econometrics, the best way is by working and experimenting on your own. In the Russian database you will find a great variety of variables. Read them carefully and try to come up with a model. For instance, you might want to run a “happiness regression” (satisfaction with life on explanatory variables such as gender, age, height, etc.). Once you decided what variable you want to explain and what covariates you are interested in, run different specifications and report the following things:

a. Report your preferred specifications using the command outreg, and format the tables in Excel in Journal-fashion.

b. Analyze qualitatively (sign, statistical significance) and quantitatively each coefficiente. If it is plausible, try with a log-log specification (to get elasticities) or some non-linear variables (e.g. age and age squared), and present the marginal effects carefully.

You should present all the information in good-looking tables and figures. Before running any regression, take a couple of lines to motivate the regression and provide some descriptive statistics. You should run over 100 regressions, but you should present at most 5 specifications of the model. The real challenge of working both in the private market and in academia is to produce concise and well-presented information. Never (and I mean never) present a table or figure without some accompanying lines commenting it.

2. Repeat the simulation exercise including minor modifications to see the following points. If you want, you can come up with a model of your own:

a. That an increase in the sample size implies a decrease in the standard errors.

b. What happens if you increase the variance of "u"?

c. In the real world "u" is by definition something we cannot measure nor observe. We estimate the coefficients using that "u" is orthogonal to the included regressors. If we estimated "u" (as the residual of the regression), we would find that it is exactly orthogonal to the regressors. But in this fictional world you know "u", and then you can test the orthogonality condition. Do it for all the possible specifications.

d. Include a measurement error in the variable intelligence (e.g. a normally distributed noise). Then observe that the estimated coefficient is downward biased. Now create a measure of “confidence” as a stochastic and increasing function of the variable intelligence. Show that even though IQ does not appear in the data generating process (only intelligence), when you increase the measumerement error in intelligence the variable “confidence” will appear as significant in the regression. People don’t usually pay attention to this effect, although it is probably a very common source of biases in econometric estimations.

Chapter 3
Testing

Brief Review of Asymptotic Theory

Small Sample

To be completed.

Large Sample

To be completed.

Review of Hypothesis Testing

To be completed.

Exercise 3.1: Run a regression and pick one variable. Using the coefficient and the standard error compute the t-statistic. Then compute "by hand" the p-value and the confidence interval using both the t distribution and the normal distribution. Verify that Stata uses the former.

The linear model with Maximum Likelihood

To be completed.

Trinity of tests

To be completed.

Simultaneous Tests

To be completed.

Nested Tests

To be completed.

An example: Growth regressions

Mankiw et al. (1992) have fitted cross section regressions to data for a NONOIL sample of 98 countries for the period 1960-1985 using the Solow model. The corresponding database is mrw.dta, and was obtained from the paper's appendix. The model used in Table IV is:

$$\ln(GDP_{1985}) - \ln(GDP_{1960}) = \beta_1 + \beta_2 \ln(GDP_{1960}) + \beta_3 \ln\left(\frac{I}{GDP}\right) + \beta_4 \ln(n + g + \delta) + \eta$$

For simplicity, assume $g + \delta = 0.05$. First, we want to give a joint LR test of the following two hypotheses: the Solow restriction ($\beta_3 + \beta_4 = 0$) and no conditional convergence ($\beta_2 = 0$). But we want to obtain the LR statistic "by hand". Furthermore, we want to perform the test both under the "small sample" assumptions and under the "large sample" assumptions. For this example and the ones that follow we will always use a 5% level of significance.

Run the growth regression:

```
. use mrw, clear

. gen lngdp1960=ln(gdp1960)

 . gen lngrowth = ln(gdp1985)- ln(gdp1960)

 . gen lns=ln(invest)

 . gen lnngd=ln(workpop+.05)

. reg lngrowth lngdp1960 lns lnngd if sn==1
```

Recall that the LR-statistic is given by:

$$LR = n \ln \frac{e'e_R}{e'e_U}$$

The numerator and the denominator are the sums of squared residuals from the restricted and unrestricted models, respectively. Under the large-sample assumptions this is distributed chi-squared with *q* degrees of freedom (the number of independent linear hypotheses involved). Under the small-sample conditions we know the distribution of:

$$F = \frac{n-k}{q}\left[\exp(LR)^{\frac{1}{n}} - 1\right]$$

This is distributed *F(q,n-k)*. In practice you should probably use the small-sample distribution in order to be more conservative.

So, let’s find all those expressions by hand. Store the sum of squared residuals from both models:

```
. reg lngrowth lngdp1960 lns lnngd if sn==1

. ereturn list // we need some auxiliary variables

. local k=e(df_m)+1 // number of variables; +1 because of the constant

. local n=e(N) // number of observations

. local rss_unrestricted=e(rss)

. gen solow=lns-lnngd

. reg lngrowth solow if sn==1

. local rss_restricted=e(rss)
```

Compute the statistics, the critical values and the p-values:

```
. local LR=`n'*ln(`rss_restricted'/`rss_unrestricted')

. local F=((`n'-`k')/2)*(exp(`LR')^(1/`n')-1). local Chi_critical = invchi2(2,0.95)

. local Fcritical = invF(2,`n'-`k',0.95)

. display "Statistics: LR=`LR', F(2,`n'-`k')=`F'"

. display "Critical Values: invchi2(2,0.95)=`Chi_critical', invF(2,`n'-`k',0.95)=`Fcritical'"

. display "P-values: Large-sample: " chi2tail(2,`LR') " | Small-sample: " Ftail(2,`n'-`k',`F')
```

Notice that the results are the same for large- and small-sample conditions: the statistics are greater than the critical values, so reject the joint hypothesis of no conditional convergence and the Solow restriction.

Hay una forma mas facil de computar el Asymptotic LR test:

```
. reg lngrowth lngdp1960 lns lnngd if sn==1
. est store unrestricted
```

```
. reg lngrowth solow if sn==1
```

```
. est store restricted
```

```
. lrtest unrestricted restricted
```

Vemos que efectivamente llega al mismo LR-statistic que lo que hicimos mas arriba. Lo denotamos “Asymptotic” LR porque utilize la distribucion (menos conservadora) Chi_squared. Quizas no parece muy importante ahora, pero “lrtest” provides an important alternative to Wald testing for models fitted by maximum likelihood.

El Wald test hubiera sido mucho mas facil, pues es el que Stata computa automaticamente:

```
. reg lngrowth lngdp1960 lns lnngd if sn==1
```

```
. test (lngdp1960==0) (lns+lnngd==0)
```

Recordemos que los dos tests (junto al LM) son equivalent asymptotically, por lo que no deberiamos sorprendernos de que den resultados sumamente similares.

Simultaneous Testing

Now we want to give simultaneous tests of the Solow restriction and the hypothesis of no conditional convergence using the S-method. We simply need to calculate the two t-statistics associated with each hypothesis and compare it to a single t critical value computed using the S-method, which is given by $S = \sqrt{qF_{1-\alpha}(q, n-k)}$.

```
. display "Scheffe critical value= " sqrt(2*invF(2,`n’-`k’,0.95))
```

The (absolute value of the) t-statistics for the no-conditional-convergence hypothesis can be obtained directly from the regression output, or taking the square root of the value provided by the command “test”:

```
. reg lngrowth lngdp1960 lns lnngd if sn==1
```

```
. test lngdp1960==0
```

The t-statictic is then 2.82, which is greater than the S critical value and then we reject the hypothesis of no conditional convergence. For the Sollow restriction, we can use the square root of the value provided by “test”:

```
. test lns+lnngd==0
```

Or we can construct the t-statistic by hand. Recall the formula:

$$t = \frac{a'b - a'\beta}{s_{a'b}} = \frac{b_3 + b_4}{\sqrt{a'\hat{V}a}}$$

Where $\hat{V} = \hat{s}^2 \left(X'X\right)^{-1}$ is the covariance matrix from OLS. We can calculate it by hand like in the previous Chapter, or we can simply save the matrix computed by Stata, e(V), in “V” after running the regression:

```
. matrix V = e(V)
```

If you want to see a matrix simply enter “matlist”:

```
. matlist V
```

Notice that the coefficients for “lns” and “lnngd” are the second and third, so $a' = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}$:

```
. local b_sum = _b[lns] + _b[lnngd]
```

```
. matrix input ap=(0,1,1,0)
```

```
. matrix a=ap'
```

```
. matrix aux=ap*V*a // we do this intermediate step because we cannot “mix” scalars and matrices in the same calculation
```

We can finally calculate the t-statistic:

```
. display “t-statistic= ” `b_sum'/sqrt(aux[1,1])
```

Which is exactly the same as the square root of the F-statistic given by “test”. Since the statistic is greater than the S critical value, we reject the Solow hypothesis.

Notice that with a non-simultaneous test we would use a t-statistic of 1.99:

```
. display invttail(`n’-`k’, 0.975)
```

Scheffe t-statistic is greater than the non-simultaneous t-statistic. In this case we would have rejected both hypotheses anyway. However, in some cases not accounting for simultaneous testing may lead to <u>serious</u> over-rejection of the null.

Unfortunately, it is not a practice in economic to perform simultaneous or nested testing.

The p-values corrected by Bonferroni’s method can be obtained directly from the “test” command:

```
. test (lngdp1960==0) (lns+lnngd==0), m(b)
```

You should notices that correcting the p-values consists simply of multiplying each (non-simulataneous) p-value by the number of tests being performed simultaneously (two in this case). Here it should be very clear to you that the results with simultaneous and non-simultaneous testing can be quite different.

Nested Testing

We want to give a “nested” sequence of LR tests for (i) the Solow restriction and (ii) the hypothesis of no conditional convergence. The first test compares the unrestricted model with respect to the model with the Solow restriction (from now on, the Model S). Run the unrestricted regression:

```
. reg lngrowth lngdp1960 lns lnngd if sn==1

. local rss_unrestricted=e(rss)

. local k=e(df_m)+1

. local n=e(N)
```

Now save the sum of squared residuals from Model S:

```
. reg lngrowth lngdp1960 solow if sn==1

. local rss_S=e(rss)
```

Compute the *LR-statistic*, F-statistic, and the *F-critical* using a 2.5% level (because of Hogg’s Theorem):

```
. local LR=`n’*ln(`rss_S’/`rss_unrestricted’)

. local F=(`n’-`k’)*(exp(`LR’)^(1/`n’)-1)

. local Fcritical = invF(1,`n’-`k’,1-0.025)

. display "F=`F’, Fcritical=`Fcritical’"
```

We see that the F-statistic is greater than the critical value, and then we reject the Solow restriction. Therefore, the nested test ends right here. If we had accepted, we would have proceeded to step two: compare the Model S with respect to the model with both the Solow restriction and the no conditional convergence restriction (Model N). You can find the corresponding piece of code below, but please notice that it is not correct to proceed, since you have rejected in the first stage:

```
. reg lngrowth solow if sn==1

. local rss_N=e(rss)

. local LR=`n’*ln(`rss_N’/`rss_S’)

. local F=(`n’-`k’+1)*(exp(`LR’)^(1/`n’)-1)

. local Fcritical = invF(1,`n’-`k’+1,1-0.025)

. display "F=`F’, Fcritical=`Fcritical’"
```

The idea of Nested testing is that from the Economic Theory the validity of one hypothesis may be conditioned on a second hypothesis. If that is the case, you should account for such hierarchy in the testing procedure. This idea has a lot of potential, but unfortunately most economists do not rely on nested testing and the procedure is almost absent in the applied literature.

Heteroskedasticity-robust standard errors

The variance matrix of the error term is given by: $\mathrm{Var}(\varepsilon \mid \mathrm{X}) = \Omega$, where ε is the error term. In the case of conditional homocedaskicity: $\Omega = \sigma^2 \cdot I$. But that is a very restrictive case, and we should consider a more general one. Recall that the assumption about homokedasticity is not used when we prove that OLS coefficients are unbiased and consistent. However, we did use the assumption to obtain consistent standard errors (and for the Gauss-Markov result).

Consider now the following case:

$$\Omega = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix}$$

Notice that all covariances are zero, but the variances can differ irrestictly. We would need to estimate *n* standard errors plus *k* coefficients, but that would be impossible because we only have *n* observations. However, Eicker (1967) and White (1980) found a wonderful result, which we will show below. Some authors also give credit to Huber (1967), or to other combinations of the former. The covariance matrix of the OLS estimator is:

$$\begin{aligned}
&\mathrm{Var}(\mathrm{b}_{\mathrm{OLS}} \mid \mathrm{X}) = V\left[(X'X)^{-1} X'\varepsilon \mid X\right] \\
&= (X'X)^{-1} X'V(\varepsilon \mid X) X (X'X)^{-1} \\
&= (X'X)^{-1} X'\Omega X (X'X)^{-1}
\end{aligned}$$

The brilliant idea is that you don't need to estimate Ω, but it is sufficient to consistently estimate $X\Omega X$. Using the analogy principle, you can use $X'WX$ as an estimator of $X'\Omega X$, where:

$$W = \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$$

Where e_i are the OLS residuals. To demonstrate the Eicker-White result in a simple way, we consider the case of a single independent variable (*k=1*). First we need to notice that we assume that $Plim X'\Omega X$ exists and is finite. Notice that:

$$X'WX \;\; = \sum_{i=1}^{n} e_i^2 x_i^2$$

For *k=1* we have: $e_i = y_i - x_i b_{OLS} = \varepsilon_i - (b - \beta) x_i$. Then:

$$e_i^2 = \varepsilon_i^2 + (b_{OLS} - \beta)^2 x_i^2 - 2(b_{OLS} - \beta) x_i \varepsilon_i$$

Multiply by x_i^2 and average:

$$X'WX = \sum_{i=1}^{n} e_i^2 x_i^2 = \sum_{i=1}^{n} \varepsilon_i^2 x_i^2 - 2(b_{OLS} - \beta)\sum_{i=1}^{n} \varepsilon_i x_i^3 + (b_{OLS} - \beta)^2 \sum_{i=1}^{n} x_i^4$$

We need the probability limits of the fourth moments to exist and be finite. If that is the case, the last two terms vanish, since b_{OLS} is a consistent estimator of β. Then we obtain:

$$Plim X'WX = Plim \sum_{i=1}^{n} e_i^2 x_i^2 = Plim \sum_{i=1}^{n} \omega_i^2 x_i^2 = Plim X'\Omega X$$

In summary, a consistent estimator of the asymptotic robust covariance matrix is:

$$\hat{A}Var(b_{OLS}) = (X'X)^{-1} X'WX (X'X)^{-1}$$

This is what Stata calculates when you specify the option "robust". You should be able to reproduce the estimates by hand. The option "robust" should be the default option, since homocedasticity is a particular case. Unless you have a strong reason to argue that the model is conditionally homokedastic, you should always use robust standard errors (not doing so would lead to over-rejection of the null hypotheses that the coefficients are zero).

Exercise 3.2: Experiment with some regressions: using the homocedastic standard errors, are you over-rejecting or under-rejecting the null hypotheses?

Similarly, when you are using panel data you should (almost) always use clustered-robust standard errors, what we discuss in the Chapter on Panel Data. We will cover further topics in the Chapter “Robust Inference” where we deal with bootstrapping and two-stages models.

Lo bueno de esta correccion es que es muy facil corregir los Wald, LR and LM statistics. Por ejemplo, en el caso del Wald:

$$\begin{aligned} W^{robust} &= (b_{OLS}-\beta)' A\left[A'\hat{A}Var(b_{OLS})A\right]^{-1} A'(b_{OLS}-\beta) \\ &= (b_{OLS}-\beta)' A\left[A'(X'X)^{-1}X'WX(X'X)^{-1}A\right]^{-1} A'(b_{OLS}-\beta) \end{aligned}$$

Which (as always) has a chi-squared asymptotic distribution.

Tests of heteroskedasticity

The command "imtest" performs the White's test, and the command "hettest" performs the Breusch-Pagan's test. In both tests the null hypothesis is whether the variance of residuals is homogeneous.

Nevertheless, you should be very careful. These tests are pretty sensitive to the assumptions of the model (for instance, they suppose normality for the error term).

Normality of the residuals

The multiple regression model does not require normality of the residuals to make inference with large samples (depending on the version of the limit central theorem used, the key assumption is the error term being i.i.d.). Nonetheless, normality is needed to make inference in small samples (there is not any “explicit” convention on what sample size is “small”).

After running a regression, we can predict the residuals:

```
. reg waistc monage height hipsiz gender
```

```
. predict res, residual
```

As a first step, we can graph the nonparametric estimate of the density function of the residuals using Kernels:

```
. kdensity res, norm
```

The option "norm" graphs a Gaussian bell with the sample mean and variance. There is an entire Chapter dedicated to Kernel estimation. We could also use the commands "qnorm" and "pnorm" to evaluate graphically the normality of the residuals. Some tests for normality, such as the Shapiro-Wilk and Shapiro-Francia ("swilk"), can also be useful.

Problem Set #3

To be completed.

Chapter 4
Instrumental Variables

Instrumental Variables

To be completed.

Two-stages least squares

To be completed.

GIVE (Generalized Instrumental Variables Estimator):

$$b_{2SLS} = \left[X'Z(Z'Z)^{-1}Z'X \right]^{-1} X'Z(Z'Z)^{-1}Z'Y$$

Call $X^* = Z(Z'Z)^{-1}Z'X$ to the fitted values from an OLS regression of X on Z. Notice that $X'Z(Z'Z)^{-1}Z'X = X'X^* = X^{*'}X^*$ and $X'Z(Z'Z)^{-1}Z'y = X^{*'}Y$. Thus, the IV estimator can also be written:

$$b_{2SLS} = (X^{*'}X)^{-1}X^{*'}Y = (X^{*'}X^*)^{-1}X^{*'}Y$$

The last expression is an OLS regression of Y on X^* (the latter being the residual from a previous regression). This explains why this estimator is also called two-stage estimator (2SLS). Later we will estimate the two stages by hand.

An example

To be completed.

Two-stages "by hand"

To be completed.

Hausman Test

To be completed.

Exercise 4.1: If you reject the null, which regression should you run? Why? What if you do not reject the null hypothesis? If I already knew for sure that the instruments are valid, why should I care for this?

Test of Overidentifying Restrictions

If the model is overidentifyed (the number el número de condiciones de momentos es mayor al número de coeficientes a ser estimados), then we can test whether some of the moment conditions are invalid (although the test will not indicate which conditions are invalid). There are two well-known tests: Sargan test and J-test.

The Sargan test

To be completed.

Weak Instruments

To be completed.

The bias of 2SLS

To be completed.

Local Average Treatment Effect

To be completed.

Binary endogenous variable

To be completed.

Ivreg2

You can download the command "ivreg2" (with the commands "overid", "ivendog", "ivhettest" associated), which provides extensions to Stata's official "ivreg". It supports the same command syntax as official "ivreg" and supports (almost) all of its options. Among the improvements you may find some very useful as the enhanced Kleibergen-Paap and Cragg-Donald tests for weak instruments.

The post-estimation command "overid" computes versions of Sargan's (1958) and Basmann's (1960) tests of overidentifying restrictions for an overidentified equation. You may find details in the help file. The Durbin-Wu-Hausman test for endogeneity ("ivendog") is numerically equivalent to the standard Hausman test obtained using "hausman" with the "sigmamore" option.

Problem Set #4

A convincing analysis of the causal link between education and wages needs an exogenous source of variation in the former. This Problem Set is based on Card (1993)

“Using Geographic Variation in College Proximity to Estimate the Return to Schooling”. It explores the use of college proximity as an instrument for schooling. Read the paper carefully before starting to answer the questions. The database "schooling.dat" is ready to be used.

1. Show descriptive statistics of the relevant variables (as in Table 1 from the paper).

2. Reproduce all the least squares regressions in Table 2.

3. Reproduce all the regressions in Table 4.

4. Can you test the instrument's exogeneity? What about education's exogeneity? If you knew that the instrument is exogenous, would you be able to test education's exogeneity? Explain and perform the test. Is it useful? You can use, for instance, the basic specification.

5. Take the baseline IV specification and add "Grew up near 2-yr College" and "Grew up near 4-yr public College" as instruments. Perform an over-identifying restrictions test (Sargan-test or J-test). What is the conclusion?

6. Suppose that you consider only education as endogenous (i.e. replace experience by age directly). Perform the weak instrument test suggested by Stock and Watson.

Chapter 5

Panel Data (Static)

Introduction

To be completed.

An example

We will begin using the database ("crime.dta") from Rafael Di Tella and Ernesto Schargrodsky "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack" (American Economic Review, 2004). You may want to read at least the introduction of the paper first. El paper utilize la metodologia de differences-in-differences explicada en el Chapter anterior. In the Problem Set you will be asked to explain the estimation design, present some descriptive statistics and replicate much of the regressions.

```
. set mem 5m
```

```
. use crime, clear
```

The syntax for panel data regressions is "xtreg depvar indepvars, i(unit) model", where "unit" must be replaced by the variable that identifies the different units in the panel, and "model" must be replaced by "fe" (fixed effects model), "re" (random effects model), or "be" (between effects model), among others. The default option is "re".

In Di Tella et al. (2004) the block constitutes the unit of observation for their study. They obtained information about each auto theft in 876 blocks for the nine-month period starting April 1, 1994 and ending December 31, 1994. The difference-in-difference model is:

$$CT_{it} = \alpha_0 SB_{it} + \alpha_1 OB_{it} + \alpha_2 TB_{it} + M_t + F_i + \varepsilon_{it}$$

Where CT_{it} is the number of car thefts in block i for month t (variable ""); SB_{it} is a dummy variable that equals 1 for the months after the terrorist attack (August, September, October, November, and December) if there is a protected institution in the block, and 0 otherwise (variable "instp"); OB_{it} is a dummy variable that equals 1 after the terrorist attack if the block is one block away from the nearest protected institution, and 0 otherwise (variable "inst1p"); TB_{it} is a dummy variable that equals 1 after the terrorist attack if the block is two blocks away from the nearest protected institution, and 0 otherwise; M_t is a month fixed effect (dummy variables "month5- month12"); F_i is a block fixed effect (variable "blockid"); and ε_{it} is the error term.

Let's run different estimates (you should compare the fixed effects estimate to the second column from Table 3). First we have to eliminate the observations after the terrorist attack and before the police intervention (months 72 and 73):

```
. drop if month==72 | month==73
```

```
. xtreg cartheft instp inst1p month5-month12, fe i(blockid) robust
```

```
. xtreg cartheft instp inst1p month5-month12, re i(blockid) robust
```

```
. xtreg cartheft instp inst1p month5-month12, be i(blockid)
```

Exercise 5.1: Why did the between estimator dropped the monthly effects? In the fixed effects model, can you include dummy variables for neighborhood or streets? Why didn't we include a dummy "month4"?

The prefix "xt" is common to every panel data model. For example, the IV estimator with panel data is "xtivreg", and the logit model with panel data is "xtlogit".

Hausman test

There is a test useful to choose between the fixed effects estimates, which can be consistent if the individual effects are correlated with the included variables, and the random effects estimates, which are consistent and efficient if the individual effects are correlated with the included variables, but inconsistent otherwise.

Recall that the RE model can be seen as a quasidemeaned model:

$$\sigma T y_i = \sigma T X_i \beta + \sigma T v_i$$

The idea of the test is to augment the quasidemeaned model by adding the demeaned data:

$$\sigma T y_i = \sigma T X_i \beta + M X_i \gamma + \upsilon_i$$

If RE is true then the quasi-demeaned data should suffice. The null hypothesis of RE is then $\gamma = 0$. This test can be carried out for all the independent variables by means of an F-ratio for the ordinary least squares estimator for the augmented model. The test statistic *qF* is asymptotically chi-squared.

If the null hypothesis is accepted, the random effects panel data model can be employed. If we reject the null, the fixed effects model must be used instead.

Let’s give an example. First estimate your model using fixed effects and save the estimates:

```
. xtreg cartheft instp inst1p month5-month12, fe i(blockid) robust
```

```
. est store fixed
```

Then run the random effects estimation:

```
. xtreg cartheft instp inst1p month5-month12, re i(blockid) robust
```

Finally, enter:

```
. hausman fixed
```

The null hypothesis is that the differences in the coefficients are not systematic. As we cannot reject such hypothesis, we could use the random effects estimates to improve efficiency.

Exercise 5.2: perform the Hausman test "by hand".

Test of unobserved effects

To be completed.

Fixed effects estimates "by hand"

There are many ways to obtain the fixed effects estimates. For instance, we could have included a set of dummy variable identifying each unit, what is known as the LSDV (Least Squares Dummy Variables) estimator. You can use the command "areg":

```
. areg cartheft instp inst1p month5-month12, absorb(blockid) robust
```

The option "absorb" generates a dummy variable for each value of "blockid" (i.e. a dummy for each block), but it does not show their estimates (as that would involve estimating and showing 876 coefficients that are not interesting by themselves). Notice that the estimates are the same than those obtained using "xtreg".

The most common strategy (and the one that Stata uses) is the within transformation. consider the following model:

$$y_{it} = \beta X_{it} + \alpha_i + \varepsilon_{it}$$

It cannot be consistently estimated by OLS when α_i is correlated with ε_{it}. The within transformation proposes to take out individual means in order to obtain the following:

$$\left(y_{it} - \bar{y}_i\right) = \beta\left(x_{it} - \bar{x}_i\right) + \left(\varepsilon_{it} - \bar{\varepsilon}_i\right)$$

Where over-bar denotes the mean value for a given *i*:

$$\bar{x}_i = \frac{\sum_{t=1}^{T} x_{it}}{T}$$

Such transformation got rid of α_i, so POLS is now consistent. We want to replicate such transformation. First generate the deviations from the mean using the command "egen". For instance, for the dependent variable you should enter:

```
. bysort blockid: egen mean_cartheft= mean(cartheft)

. gen dm_cartheft= cartheft - mean_cartheft
```

And for each independent variable X you should enter:

```
. bysort blockid: egen mean_X= mean(X)

. gen dm_X= X - mean_X
```

First we should eliminate observations with missing values in any of the variables included in the model. However, there are no missing values in this database. In order to save code lines and time, we will use the "foreach" syntax:

```
. foreach var of varlist cartheft instp inst1p month5-month12 {

. bysort blockid: egen mean_`var'=mean(`var')

. gen dm_`var'= `var' - mean_`var'

. }
```

Then we run the regression:

```
. reg dm_cartheft dm_instp dm_inst1p dm_month*, nocons robust
```

Compare the results with the original estimates using "xtreg":

```
. xtreg cartheft instp inst1p month5-month12, fe i(blockid) robust
```

Exercise 5.3: the standard errors are not “exactly” the same than those showed by the "xtreg" command. Why? (Hint: degrees of freedom adjustment).

Exercise 5.4: A less-known way to obtain fixed effects estimates is running a regression for each unit, and then averaging all the estimates obtained. Carry out such estimation.

Exercise 5.5: Explain intuitively why the within transformation is more efficient than first differences. Show this running some regressions with this database. Furthermore, show that they are numerically identical when *T=2*.

Clustered Standard Errors

One of the common violations to the i.i.d. assumption is that errors within groups are correlated in some unknown way, while errors between groups are not. This is known as clustered errors, and OLS estimates of the variance can be corrected following a Eicker-Huber-White-robust treatment of errors (i.e. making as few assumptions as possible). The (unclustered) heterocedasticity-robust variance estimator is given by:

$$(X'X)^{-1}\left(\sum_{i=1}^{N}(e_i x_i)'(e_i x_i)\right)(X'X)^{-1}$$

And the robust cluster variance estimator is:

$$(X'X)^{-1}\left(\sum_{j=1}^{N_C}u_j'u_j\right)(X'X)^{-1}\text{, where } u_j=\sum_{j=1}^{N_C}e_i x_i$$

Where N_C is total number of clusters, and e_i are the regression residuals. For the sake of simplicity the multipliers (which are close to 1) were ommited from the formulas. This is for the Pooled OLS estimator, but a similar expression arises for the within transformation.

Over-rejection

Ya comentamos que los errores standard hekerocedasticity-robust eran casi siempre greater than the homokedastic one. As a consequence, if you have heterocedasticity and you do not use the robust standard errors you are over-rejecting the null hypothesis that the coefficients are zero.

But when you compare the robust (unclustered) y and the clustered variance estimators, there is not a general result. If the within-cluster correlation is negative, then within the same cluster there will be big negative e_{ij} along with big positive e_{ij}, and small negative e_{ij} with small positive e_{ij}. This would imply that the cluster sums of $e_{ij}x_{ij}$ have less variability than the individual $e_{ij}x_{ij}$, since within each cluster the eij's will be "cancelling" each other. Then the variance of the clustered estimator will be less than the robust (unclustered) estimator. In this case, using cluster-robust standard errors will not only make your inference robust to within-cluster correlation, but it will improve your statistical significance.

You repeat the reasoning for the opposite case (i.e. with positive within-cluster correlation), and where there is no clustered errors. The following is an applied example.

An example

Open the database:

```
. use education, clear
```

This is a database on test scores that was generated including within-school correlation in the error term. The estimate of the coefficient on treatment is consistent. Now we can estimate the unclustered and clustered standard errors:

```
. xtreg test treatment female educ income, fe i(school) robust
```

```
. xtreg test treatment female educ income, fe i(school) cluster(school)
```

In this example not accounting for within-cluster correlation would have led to wrong inference at the 10% for the treatment coefficient.

Should I always use clustered standard errors?

If the assumptions are satisfied, and the error term is clustered, you will get consistent standard error estimates if you use the cluster-robust estimator. On the other hand, if the assumptions are satisfied and errors are not clustered, you will get roughly the same estimates as if you had not specified cluster.

Why not always specify cluster? Well, the cluster-robust standard error estimator converges to the true standard error as the number of clusters M approaches infinity, not the number of observations N. Kezdi (2004) shows that 50 clusters (with roughly equal cluster sizes) is often close enough to infinity for accurate inference. Moreover, as long as the number of clusters is large, even in the absence of clustering there is little cost of using clustered-robust estimates.

However, with a small number of clusters or very unbalanced cluster sizes, inference using the cluster-robust estimator may be very incorrect. With finite M, the cluster-robust estimator produces estimates of standard errors that may be substantially biased downward (i.e. leading to over-rejection). See Wooldridge (2003) and Cameron et al. (2006) for further discussions and suggestions on the matter.

Right specification

The within-cluster correlations may disappear with a correctly specified model, and so one should always be alert to that possibility. Consider as an example a model where the dependent variable is cell phone expenditure. If you only included explanatory variables at the individual level (pooled OLS), then you would find serious within-cluster correlation: since many of the calls are between household members, the errors of the individuals within the household would certainly be correlated to each other. However, a great deal of such within-cluster correlation would disappear if you used a fixed-effects model. Furthermore, by adding the right predictors the correlation of residuals could almost disappear, and certainly this would be a better model.

For instance, you can see whether in the above model the difference between the unclustered and clustered standard errors is magnified or not after controlling for fixed effects:

```
. reg test treatment female educ income, robust

. reg test treatment female educ income, cluster(school)

. xtreg test treatment female educ income, fe i(school) robust

. xtreg test treatment female educ income, fe i(school) cluster(school)
```

Nested multilevel clustering and nonlinear models

Beyond the basic one-dimensional case, one may consider a multiple-level clustering. For instance, the error term may be clustered by city and by household, or it may be clustered by household and by year. In the first example the levels of clustering are called nested. To estimate cluster-robust standard errors in the presence of nested multi-level clustering, one can use the “svy” suite of commands. However, specifying clustering solely at the higher level and clustering at the higher and lower level is unlikely to yield significantly different results.

Additionally, in Wooldridge (2006) there is an entire section of nonlinear clustered standard errors, with examples and Stata commands in the Appendix.

Clustered Data

To be completed.

Problem Set #5

The Problem Set is based on Rafael Di Tella and Ernesto Schargrodsky "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack" (American Economic Review, 2004). Read the paper carefully before starting to answer the questions. The database "crime.dat" is ready to be used. However, some variables (such as "Two-Blocks Police") must be generated first.

1. Carefully reproduce Table 2, Table A1 and (the monthly version of) Figure 2 from the paper.

2. Reproduce all the regressions in Table 3.

3. Reproduce all the regressions in Table 7.

4. Criticize the paper in no more than three paragraphs (focus on exogeneity).

Chapter 6
Natural Experiments

Introduction

To be completed.

Differences-in-differences

To be completed.

Bertrand, Duflo and Mullainathan

Bertrand, Duflo and Mullainathan (BDM) present a study of the consequences of using standard errors based on OLS, rather than "cluster" standard errors. BDM first present a survey of DD papers published in major journals. Of the 92 papers in their survey 80 have a "clustering" problem arising from heteroscedasticity and 36 deal with in some way.

The second step in the BDM study is to construct an elaborate sampling experiment based on a wage equation for female workers in the Current Population Survey (CPS) for 1979-1999. Although there are 540,000 individual observations, these are aggregated into 50 states, yielding a total of 50 states times 21 years equals 1050 state-year observations.

states. The dependent variable is average weekly earnings; the independent variables include time and state dummies as well as controls, including employment status, education, and age. BDM introduce a "placebo law" or policy change chosen randomly. If this artificial change in the law occurs in, say 1985, then it continues through 1999.

Note that this policy change does not occur in reality. This should lead to a rejection of the null hypothesis of "no effect" of the policy change in five percent of the stimulations. In Table VIII they compare the rejection rates for OLS standard errors based on the strong form of homoscedasticity with robust standard errors, treating the observations for all time periods for a given state as a cluster. The rejection rates for OLS standard errors are slightly less than ten times the rejection rates for robust standard errors.

Regression discontinuity design

To be completed.

Problem Set #6

Meyer et al. (1995) consider a “natural experiment” involving an increase in workers’ compensation benefits in the states of Kentucky and Michigan. They calculate differences between high- and low-earnings groups before and after the policy change, in a difference-in- differences fashion.

1. Reproduce Table 4.

2. Reproduce Table 6.

To be completed.

Chapter 7
Dynamic Panel Data

Introduction

Cuando trabajabamos con el static panel with fixed effects (i.e. we need to use the within transformation), we needed the “strict exogeneity” condition for consistency:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it} \quad i = 1,2,...,N;\ t = 1,2,...,T$$

$$E[\varepsilon_{it} \mid x_{i1},...,x_{iT},c_i] = 0$$

To see this:

$$\hat{\beta}_{FE} = \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)'(x_{it}-\overline{x}_i)\cdot\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)'(y_{it}-\overline{y}_i)$$

$$plim\ \hat{\beta}_{FE} = \beta + plim\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)'(x_{it}-\overline{x}_i)\cdot plim\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)'(\varepsilon_{it}-\overline{\varepsilon}_i)$$

We need to show that the last term converges to zero. To obtain this result by applying the Law of Large Numbers, we need $E[(x_{it}-\overline{x}_i)'(\varepsilon_{it}-\overline{\varepsilon}_i)] = 0$. That is to say $E[x_{it}\varepsilon_{it}] = 0$ is not sufficient. We need the stronger condition given by strict exogeneity.

But we will show that strict exogeneity rules out various time-series models, among them those with lagged dependent variable as regressor:

$$y_t = \alpha y_{t-1} + x_t\beta + \varepsilon_t \quad t = 1,2,...,T$$

This is called first-order autoregressive model (AR1). See the Chapter on Time Series for further details. Strict exogeneity would require:

$$E[\varepsilon_t \mid y_0, y_1,..., y_{T-1}, x_1, x_2,..., x_T] = 0 \quad \forall t = 1,2,...,T$$

But ε_t is correlated with y_t, ε_{t-1} is correlated with y_{t-1}, and so on. So this condition cannot hold. Notice that we don’t need within transformation nor even panel data to get this results. A good answer for time series is to assume sequential exogeneity:

$$E[\varepsilon_t \mid y_0, y_1,..., y_{t-1}, x_1, x_2,..., x_t] = 0 \quad \forall t = 1,2,...,T$$

In other words, for strict exogeneity we assumed that the error term was orthogonal to all past, present and future values of the explanatory variables, while for sequential exogeneity we only assume that it is orthogonal to the past and present values.

Even though this solves the problem in the time series model, this will not work in the fixed effects model:

$$plim \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)'(\varepsilon_{it}-\bar{\varepsilon}_i)\neq 0$$

Since $\bar{x}_i$ includes all past, present and future values of x_{it}.

In summary, with AR(1) strict exogeneity fails, and with the within transformation the sequential exogeneity condition is not enough. As a consequence, a model with both fixed effects and AR(1) is bound to be inconsistent with either strict or sequential exogeneity.

Podemos obtener una idea mas precisa de la inconsistencia if we focus in the following AR(1) example without additional regressors:

$$y_{i,t}=\gamma y_{i,t-1}+\alpha_i+\varepsilon_{i,t}$$

Consider $|\gamma|<1$. We have observations on individuals *i=1,…,N* for periods *t=1,…,T*. The within estimator for γ is:

$$\hat{\gamma}_{FE}=\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t}-\bar{y}_i\right)\left(y_{i,t-1}-\bar{y}_{i,-1}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t-1}-\bar{y}_{i,-1}\right)^2}$$

with $\bar{y}_i=\frac{1}{T}\sum_{t=1}^{T}y_{i,t}$ and $\bar{y}_{i,-1}=\frac{1}{T}\sum_{t=1}^{T}y_{i,t-1}$

Substituting by the definition of $y_{i,t}$, we can obtain:

$$\hat{\gamma}_{FE}=\gamma+\frac{(1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\varepsilon_{i,t}-\bar{\varepsilon}_i\right)\left(y_{i,t-1}-\bar{y}_{i,t-1}\right)}{(1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t-1}-\bar{y}_{i,t-1}\right)^2}$$

This estimator is biased and inconsistent for $N\to\infty$ and fixed *T*. This was first noticed by Nickell (1981). It can be shown that (Hsiao, 2003, Section 4.2):

$$\underset{N\to\infty}{plim}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\varepsilon_{i,t}-\bar{\varepsilon}_i\right)\left(y_{i,t-1}-\bar{y}_{i,t-1}\right)=-\frac{\sigma_\varepsilon^2}{T^2}\frac{(T-1)-T\gamma+\gamma^T}{(1-\gamma)^2}\neq 0$$

Thus, for fixed *T* we have an inconsistent estimator. One way to solve this inconsistency problem was proposed by Anderson and Hsiao (1981). Take first differences:

$$y_{i,t}-y_{i,t-1}=\gamma\left(y_{i,t-1}-y_{i,t-2}\right)+\left(\varepsilon_{i,t}-\varepsilon_{i,t-1}\right)$$

If we estimated the above model by OLS we would yield inconsistent estimates, since $y_{i,t-1}$ and $\varepsilon_{i,t-1}$ are correlated by definition. However, notice that $y_{i,t-2}$ is correlated with $\left(y_{i,t-1}-y_{i,t-2}\right)$ but not with $\left(\varepsilon_{i,t}-\varepsilon_{i,t-1}\right)$ (unless $\varepsilon_{i,t}$ exhibits autocorrelation). Then, we can use $y_{i,t-2}$ as an instrument.

Anderson and Hsiao also proposed $\left(y_{i,t-2} - y_{i,t-3}\right)$ as an instrument. Furthermore, Arellano and Bond (1991) suggested that the list of instruments can be extended by exploiting additional moment conditions, and then using a GMM framework to obtain estimates.

To be completed.

And there are alternatives, such as Arellano and Bover. To be completed.

Exercise 7.1: What do you think about the orthogonality condition between $y_{i,t-2}$ and $\left(\varepsilon_{i,t} - \varepsilon_{i,t-1}\right)$? (Hint: using the same argument, one could argue to use $y_{i,t-1}$ as an instrument for $y_{i,t}$ in non-dynamic models).

Xtabond

The Arellano-Bond estimator is obtained with the command "xtabond".[1] Nevertheless, we need first to declare the dataset to be panel data using "tsset":

```
. tsset panelvar timevar
```

declaration is also needed for other commands, such "stcox" and "streg" for duration models and the time series models. Once you have done such declaration, you can refer to lagged values using "L.variable" and first difference using "D.variable". You can also utilize the commands "xtsum" and "xttab", similar to "summarize" and "tabulate" but designed specifically for panel data.

To be completed.

Xtabond2

If you are using Stata 7 or a newer version, you can install "xtabond2". If you do not find the command using "search xtabond2," download it using the following command: ssc install xtabond2, all replace. If that does not work, then download the ado- and help-files from "http://ideas.repec.org/c/boc/bocode/s435901.html", and copy them in "StataFolder\ado\base\?\" (where "?" must be replaced by the first letter of each command). This procedure is particularly useful in computers with restrained access to the Web.

The command "xtabond2" can fit two closely related dynamic panel data models. The first is the Arellano-Bond (1991) estimator, as in "xtabond", but using a two-step

[1] This command has been updated on May 2004. You may want to execute "update all" to install official updates to Stata.

finite-sample correction. The second is an augmented version outlined in Arellano and Bover (1995) and fully developed in Blundell and Bond (1998).

A problem with the original Arellano-Bond estimator is that lagged levels are often poor instruments for first differences, especially for variables that are close to a random walk. Arellano and Bover (1995) described how, if the original equations in levels were added to the system, additional moment conditions could be brought to bear to increase efficiency. In these equations, predetermined and endogenous variables in levels are instrumented with suitable lags of their own first differences.

However, if you are using Stata 10, you may use the command "xtdpdsys".

Monte Carlo Experiment

We will implement a Monte-Carlo experiment to evaluate the seriousness of the bias in "short" dynamic panels. Consider the Data Generating Process of the following growth model:

$$GDP_{i,t} = \gamma GDP_{i,t-1} + \beta_1 inst_{i,t} + \beta_2 entr_{i,t} + \alpha_i + \varepsilon_{i,t}$$

Where $GDP_{i,t}$ is the real gross domestic product per capita, $inst_{i,t}$ is an index of the quality of institutions, $entr_{i,t}$ is an index of the entrepreneurship spirit in the country, α_i's are the country fixed effects, and $\varepsilon_{i,t}$ is the error term. The information is collected for *i=1,…,N* individuals, during *t=1,…,T* periods.

Suppose *N=50* and *T=5*. Generate variables "year" and "country":

```
. clear

. set obs 250

. gen year=mod(_n-1,5)+1

. gen country=ceil(_n/5)
```

See the results using "browse". Let's generate data on "inst" and "entr":

```
. gen inst = 0.2 + year*0.1 + uniform()/2

. gen entr = 0.4 + year*0.08 + uniform()/3
```

Generate the fixed effects as correlated with both "inst" and "entr":

```
. gen fe_aux=(inst + entr + uniform())/3 if year==1

. bysort country: egen fe = mean(fe_aux)

. drop fe_aux
```

Finally, generate data on "GDP":

```
. gen GDP = 0.2*inst+0.3*entr+fe+uniform()
```

```
. bysort country: replace GDP = GDP+ 0.3*GDP[_n-1] if _n>1
```

```
. drop fe
```

See the results for the first 6 countries:

```
. graph twoway line GDP inst entr year if country<7, by(country)
```

Now declare "tsset":

```
. tsset country year
```

We know that the "true" values of the parameters are: $\gamma = 0.3$, $\beta_1 = 0.2$ and $\beta_2 = 0.3$. Firstly, run a pooled OLS regression:

```
. reg GDP inst entr, robust
```

As you can see, the estimates are seriously biased. Now add the lagged dependent variable as a regressor:

```
. reg GDP L.GDP inst entr, robust
```

The estimates are still very biased. Run a regression including only fixed effects:

```
. xtreg GDP inst entr, fe i(country) robust
```

The estimates are also wrong. Run a regression including both fixed effects and lagged dependent variable, but inconsistent:

```
. xtreg GDP L.GDP inst entr, fe i(country) robust
```

Although considerably less biased than the previous estimates, they are still inconsistent. Finally, run the Arellano-Bond estimator:

```
. xtabond GDP inst entr, robust
```

The bias is even smaller. We can repeat the whole experiment using different values for *N* and *T*. In particular, keep *N* fixed in 100 and vary *T=5, 10, 15, …, 35*. We will show the estimates of the coefficients and their standard values. Run the following do-file:

```
. * Generate data for T=25 and N=50
. clear
. set mat 800
. set obs 1250 // 25*50
. gen year=mod(_n-1,25)+1
. gen country=ceil(_n/25)
. gen inst = 0.2 + year*0.05 + uniform()/2
. gen entr = 0.35 + year*0.08 + uniform()/3
. gen fe_aux=(inst + entr + uniform())/3 if year==1
. bysort country: egen fe = mean(fe_aux)
. drop fe_aux
. gen GDP = 0.2*inst+0.3*entr+fe+uniform()
. bysort country: replace GDP = GDP+ 0.5*GDP[_n-1] if _n>1
. drop fe
. tsset country year
. * Repeat the experiment for different T's
. forvalues T=3(2)25 {
. preserve
. quietly: keep if year<=`T'
. display "T="`T'
. quietly: xtreg GDP L.GDP inst entr, fe i(country) robust
. quietly: ereturn list
. matrix b = e(b)
. matrix se = e(V)
. display "XTREG: gamma=" b[1,1] " (" se[1,1] ")"
. quietly: xtabond GDP inst entr, robust
. quietly: ereturn list
. matrix b = e(b)
. matrix se = e(V)
. display "XTABOND: gamma=" b[1,1] " (" se[1,1] ")"
. restore
. }
```

Run the do-file a couple of times. You will notice that for *T=3* the "xtabond" may yield more inconsistent coefficients than the original "xtreg". For *T>5* the coefficient yielded by "xtabond" is always closer to the real value (0.5) than that yielded by "xtreg". In these simulations the bias becomes relatively insignificant for T>21.

Exercise 7.2: Repeat the experiment changing the error term (e.g. "uniform()*2" instead of "uniform()"). Relate to the formula on inconsistency (Hsiao, 2003).

Exercise 7.3: Repeat the last experiment including autocorrelation in the error term.

Kiefer

Sometimes you also have fixed N.

To be completed.

Problem Set #7

To be completed.

Chapter 8
Binary Logit/Probit

Binary regression

To be completed.

Heteroskedasticity is by no means an excuse to abandon linearity. A second drawback is that the model might predict values below zero and above one, which is absurd from the point of view of the underlying economic model. This problem disappears if we saturare the model, as you are requested to prove in the following exercise.

Exercise 8.1: Prove that in a completely saturated model, the linear probability model does make predictions above 1 or below 0.

But even without saturated models predictions out of bounds may not appear. And if they do, there is a rather obvious solution: replace your prediction betha_hat * Xi by Q(betha_hat * Xi), where the function Q replaces values below zero by zero and values above one by one.

So this is not an strong reason to abandon the beautiful world of linearity. There is a third issue with the LPM that may be strong enough to change to nonlinear models: in the LPM marginal effects are linear in parameters (recall that the linear model is not constrained to be linear in variables).

In some economic models (i.e. from the aprioristic analysis) you may expect to find nonlinearities. For instance, consider a model for the probability of approving an exam. For the most prepared students (those with high probabilities of passing), an additional hour of study will have virtually no effect on their probabilities of approving. The same is valid for the students that have not paid any attention since the first day of class: the impact of an additional hour of study on their probabilities of approving is almost zero.

However, consider a student right in the middle. As she is in the borderline, she might pass the exam just because what she studied in the last hour: the marginal effect of an additional hour of study on her probability of approving will be considerable.

Notice that no podriamos capturar esa linealidad con OLS. Si por ejemplo los efectos se hacen mas fuertes para individuos con mas horas de estudio (i.e. increasing returns to study) entonces si podriamos, introduciendo hours and hours^2. Pero en el ejemplo el efecto de hours of study no depende de una de las x’s, sino que depende de la probabilidad de aprobar misma.

The attractive of the logit/probit model is to capture exactly that effect. Consider a vector x of explanatory variables, and a vector β of parameters. The probability (p) that an event y happens is:

$$p = F(x'\beta)$$

Where $F(\cdot)$ has the following properties:

$$F(-\infty) = 0, F(\infty) = 1, \ f(x) = dF(x)/dx > 0$$

For instance, in the probit model $F(\cdot)$ is the accumulated distribution of a standard normal:

$$F(x'\beta) = \int_{-\infty}^{x'\beta} \frac{1}{\sqrt{2\pi}} e^{\frac{s^2}{2}} ds$$

On the other hand, in the logit model $F(\cdot)$ is the accumulated distribution of a logistic random variable:

$$F(x'\beta) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

Notice that the marginal effects of the explanatory variables are not linear:

$$\frac{\partial p}{\partial x_k} = \beta_k f(x_i'\beta)$$

Furthermore, they depend on xi. Notice also that the sign of the derivative is the sign of the coefficient β_k:

$$\mathrm{sgn}(\partial p/\partial x_k) = \mathrm{sgn}(\beta_k)$$

However, the value of bethak lacks direct quantitative interpretation. The parameter β_k is multiplied by $f(x_i'\beta)$, which is maximum when x'ß = 0 and decreases when $x_i'\beta$ goes to $\pm\infty$ (indeed, the difference between the models logit and probit is the weight of the tails of the logistic and standard normal distribution).

A rule-of-thumb is to evaluate these effects at x'ß = 0 , where g(0) = .4 for probit and g(0) = .25 for logit.

Later we will discuss more about marginal effects, including how to estimate their standard errors.

Latent variable model

<u>To be completed.</u>

The Maximum-Likelihood estimator

We have an i.i.d. sample of binary events and explanatory variables (y_i, x_i), for i =1,…,n,. The random variable y_i follows a Bernoulli distribution with $p_i = P(y_i = 1)$. The likelihood function is then:

$$L(\beta) = \prod_{y_i=1} p_i \prod_{y_i=0} (1-p_i) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$

And the log-likelihood function:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i) \right] \\ &= \sum_{i=1}^{n} \left[y_i \ln(F(x_i'\beta)) + (1-y_i)\ln(1-F(x_i'\beta)) \right] \end{aligned}$$

The score function is:

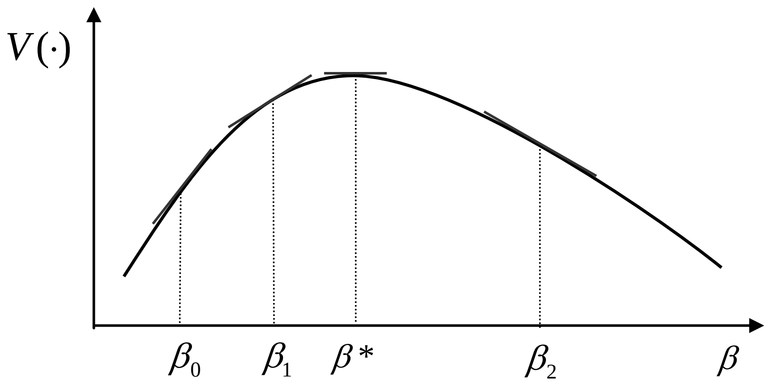
$$s_i(\beta) = \frac{\partial \ell_i}{\partial \beta} = \frac{g(x_i'\beta)[y_i - G(x_i\beta)]}{G(x_i'\beta)[1-G(x_i'\beta)]} x_i'$$

And first order conditions are:

$$\sum_{i=1}^{n} \frac{(y_i - F(x_i'\beta)) f(x_i'\beta) x_{ki}}{F(x_i'\beta)(1-F(x_i'\beta))} = 0 \quad k = 1,\dots,K$$

It is a system of *K* non-linear equations and *K* unknowns. There is a solution for β if the regressors are linearly independent and if there is no a perfect classifier. However, it is impossible to find it explicitly. Macfadden showed that the likelihood problem is globally concave, so standard numerical maximization methods are guaranteed to find the maximum.

For those who have never heard about numeric methods for maximization, you can grasp the intuition from the following example. Suppose that you know that the objective function, $V(\cdot)$, is continuous and globally concave:



You can find the (unique) maxima (β^*) with the following algorithm. Start with an arbitrary value (β_0) and then evaluate the derivative at that point, $V'(\beta_0)$. If the derivative is positive (negative), then move to the right (left). You can use information about the second derivative to decide about the size of the “step”. You may have problems if your step is too long (e.g. go from β_0 to β_2), but those issues are very easy to solve. Eventually you will end up in β^*, where the derivative is zero.

The principle is the same when β is a vector. This simple idea can be extended to more complicated methods (including even genetic algorithms) to find a solution even if there are multiple local maximums, not-well-behaved objective functions, etc.

Mencionamos que una de las buenas cosas de OLS es que es un modelo muy estudiado, y por lo tanto uno puede usar muchos resultados. De los modelos logit and probit se puede decir algo similar, pues hay muchisimos resultados para maximum likelihood estimators. Uno de ellos es el de invariance: if tita_hat is the maximum likelihood estimator of tita0, then the maximum likelihood estimator of h(tita0) is h(tita_hat). Usaremos muchos resultados para maximum likelihood cuando veamos testing.

Although Stata has a simple command to estimate both models, first we will write a Stata routine to solve the maximum likelihood problem “by hand”.

Maximum likelihood estimation in Stata

Recall the Russian database from the first two weeks. We have a variable called "smokes" that indicates whether the individual smokes or not. We have plenty of interesting dependent variables (income-variables, health-variables, etc.), so we will try to estimate a model of the decision of smoking:

```
. use russia.dta
```

The command "ml" indicates to Stata each element of the maximum-likelihood estimation. Its syntax is:

```
. ml model lf name equations
```

Where "lf" is the method (we will discuss this soon), "name" is the name of the program where we will save the log-likelihood function, and "equations" is/are the specification/s (i.e. the list of the independent variable plus the dependent variables).

Then, the first task is to generate a program with the log-likelihood function:

```
. program define probit_ml
.         version 1.0
.         args lnf XB
.         quietly replace `lnf' = $ML_y1*ln(norm(`XB'))+(1-$ML_y1)* ln(1-norm(`XB'))
. end
```

Where "args lnf XB" indicates the value of the likelihood function ("lnf") and $x_i'\beta$ ("XB"). The expression "$ML_y1" is the convention for y_i. The program you write is written in the style required by the method you choose. The methods are "lf", "d0", "d1", and "d2". See "help mlmethod" for further information. However, some global macros are used by all evaluators. For instance, "$ML_y1" (for the first dependent variable), and "$ML_samp" contains 1 if observation is to be used, and 0 otherwise.

Then, the expression "`lnf' = $ML_y1*ln(norm(`XB'))+(1-$ML_y1)* ln(1-norm(`XB'))" is exactly the formulae of the log-likelihood function:

$$l(\beta) = y_i \ln(F(x_i'\beta)) + (1-y_i)\ln(1-F(x_i'\beta))$$

Notice that we used the normal accumulated distribution, and then we are estimating a probit model. Now we can use the command "ml":

```
. ml model lf probit_ml (smokes = gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2)
```

Once you have defined the maximum likelihood problem, you can verify that the log-likelihood evaluator you have written seems to work (strongly recommended if you are facing a problem not covered by any standard command):

```
. ml check
```

And you can also display a description of the current problem:

```
. ml query
```

There are a bunch of additional commands to use after "ml model" (see "help ml"). For instance, "ml init" provides a way to specify initial values and "ml search" searches for (better) initial values.

As you can see, we have included demographic, economic and health variables as regressors. But "ml" does not perform the maximization nor obtain the estimates. To do so, we need to invoke the command "ml maximize":

```
. ml max
```

Finally, we got the estimates. We can obtain a robust variance-covariance matrix or even clustered standard errors adding "robust" and "cluster(·)" as options when invoking "ml model". After invoking "ml max" you can use "ml graph" to graph the log-likelihood values against the iteration number, and "ml display" to redisplay the final results.

You can choose a particular maximization algorithm. If you are estimating a (not necessarily binary) logit model you know that the maximization problem is a concave one, and then you do not have to worry about this. Nonetheless, if you are working with another problem you might face multiple local maximums, sensitiveness to the choice of initial values, etc., so you need to be careful.

The option "technique(nr)" specifies Stata's modified Newton-Raphson (NR) algorithm, "technique(bhhh)" specifies the Berndt-Hall-Hall-Hausman (BHHH) algorithm, "technique(dfp)" specifies Davidon-Fletcher-Powell (DFP) algorithm, and "technique(bfgs)" specifies the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. The default is "technique(nr)". For a detailed explanation of the NR and BHHH algorithms we suggest Kenneth Train's "Discrete Choice Methods with Simulation" (2003), Chapter 8 ("Numerical Maximization").

Probit and logit

As we previously introduced, there is a simple command to run a logit/probit model. We can check that if the above estimates are exactly the same:

```
. probit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2
```

As always, we have the options "robust", "cluster(·)", and so on. Notice that we can only infer qualitative information from the raw coefficients (e.g. there is a positive association between smoking and being male, and it is statistically different from zero at the 1% level of confidence).

The binary logit and probit models are particular cases of more general models. For instance, we have the multinomial logit ("mlogit"), nested logit ("nlogit"), mixed logit ("mlogit"), ordered logit ("ologit"), and so forth. In the last week we will explain those models. The literature is known as "discrete choice", and it is particularly useful to model the consumer behavior. At the end of this section we will develop a logit model with fixed effects ("xtlogit").

Standard errors and Hypotheses Testing

To be completed.

Marginal Effects

As stated previously, the marginal effects depend on the value of x. But we still don't know what values of x to use for computing the marginal effects. One way to go is to use the mean values for every element of x, and evaluate the marginal effects at that point:

$$\left.\frac{\partial p}{\partial x_k}\right|_{x=\overline{x}} = \beta_k f(\overline{x}'\beta)$$

This is the "marginal effects at the mean". You can compute them using the command "dprobit" for the probit model:

```
. dprobit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust
```

At the mean values, the marginal effect of changing the value of obese from 0 to 1 on the probability of smoking is -14 percentage points. You can evaluate the marginal effects at a point different than the mean values:

```
. matrix input x_values = (1,240,1,3,0,1,0,1,10000,10000,1,1,0)
```

```
. dprobit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust at(x_values)
```

You can also use the command "mfx":

```
. probit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust
```

```
. mfx
```

For the logit model you may only use the command "mfx":

```
. logit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust
```

```
. mfx
```

However, analyzing the marginal effects at the mean values of x could be troublesome. Maybe there is no such thing as the "mean" individual (e.g. those values are not representative of any single individual).

In those cases you might be interested in evaluating the "mean marginal effects", that is, the mean of the marginal effects evaluated at every single observation. As you started interested in nonlinearities, this may sound contradictory. However, the "mean marginal effects" give an accurate idea of the actual impact of marginal changes in a variable over the entire sample:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial p}{\partial x_k}\bigg|_{x=x_i} = \frac{1}{n}\sum_{i=1}^{n}\beta_k f(x_i\beta)$$

You can use the command "margeff", which you have to install in advance ("search margeff"):

```
. probit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust
```

```
. margeff
```

It works both with logit and probit models. Moreover, you can calculate these marginal effects "by hand". Estimate the model, predict each $x_i\beta$ and store them in the variable "xb":

```
. probit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust
```

```
. predict xb, xb
```

We need to retrieve the coefficients on the variable under consideration (for instance, on "monage"):

```
. ereturn list
```

```
. matrix coef=e(b)
```

And we can calculate the marginal effect at every observation (to be stored in variable "me"):

```
. gen me=coef[1,2]*normden(xb)
```

Now we can calculate the mean of the marginal effects, or even their median:

```
. tabstat me, stats(mean median)
```

Alternatively, we can estimate nonparametrically the density distribution of the marginal effects over the sample (using the command "kdensity", for obtaining a Kernel estimator):

```
. kdensity me
```

There is no "right" or "wrong" way to evaluate the marginal effects. The important thing is to give the right econometric interpretation for the strategy chosen.

Goodness of Fit

There is a simple way to imitate the R^2 from OLS, denominated pseudo-R^2:

$$LR = 1 - \frac{\ln L}{\ln L_0}$$

Where L is the maximum value of the log-likelihood function under the full specification, and L_0 is the maximum value of the log-likelihood function in a model with only a constant. Then LR measures the increase in the explanatory power from considering a model beyond a simple constant.

As you can see in Menard (2000), there are some "desirable" properties for a goodness-of-fit index that the R^2 satisfyes: To be completed. There is not an index for the logit/probit model that satisfies all of the properties, but you have a broad set of choices. There is no "right" or "wrong" way to measure the goodness of fit, but you should study whether a particular measure is more appropriate according to the underlying model under consideration.

Another strategy is to generate a "percentage of right predictions". As you probably noticed, the predictions are in the (0,1) open interval, and then you cannot compare them with the actual outcomes (either zero or one). But you can use cutoff points (c) to transform the (0,1) predictions in either 0 or 1:

$$\hat{y}_i \equiv 1\left[\hat{p}_i > c\right]$$

$$\hat{y}_i^c \equiv 1\left[\hat{y}_i = y_i\right]$$

$$H = \frac{\sum_{i=1}^{n} \hat{y}_i^c}{n}$$

Where *H* is the "percentage of right predictions". The index is pretty sensitive to the choice of *c*. Furthermore, a trivial model has $H \geq 1/2$. For instance, consider a model explaining the decision to commit suicide. Since less than *0.1%* of the people commit suicide, a trivial model predicting "no person commit suicide" would obtain an $H = 0.999$.

We will create a table of "Type I – Type II" errors. Suppose *c=0.5*. Generate $\hat{y}_i$ and then $\hat{y}_i^c$:

```
. logit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust
```

```
. predict smokes_p
```

```
. gen smokes_hat = (smokes_p>0.3)
```

And then just employ "tabulate":

```
. tabulate smokes smokes_hat
```

The right predictions are on one of the diagonals, and the Type I and Type II errors are on the opposite diagonal. The *H* equals the number of "right" predictions over the sample size.

The choice of *c* is very sensitive to the data under consideration. In order to help you choose an appropriate value, you can plot the predictions. One way to go would be to choose the *c* that maximizes *H*. We frequently use this method to get an idea of the fit of the model, but we think it is too sensitive to calibration as to be considered for rigorous academic research.

Exercise 8.1: Repeat the exercise varying the cutoff point (*c*). Suppose that you are hired by a company that is deeply worried about committing a Type-I-error: would you increase or decrease *c* to reflect this? (Hint: use the Type I-II errors table)

Logit with fixed effects

We can see a logit model with fixed effects for T=2. First remember that in a logit estimation we cannot have perfect predictors. Then, if an individual repeated its outcome every time, then its fixed effect would be a perfect predictor. As a consequence, we must eliminate those observations with constant outcomes.

In the T=2 case, the possible values for (y_1, y_2) will be $(0,1)$ or $(1,0)$. The conditional probability for the first is:

$$P\{y_i = (0,1) \mid \bar{y}_i = 1/2, \alpha_i, \beta\} = \frac{P\{y_i = (0,1) \mid \alpha_i, \beta\}}{P\{y_i = (0,1) \mid \alpha_i, \beta\} + P\{y_i = (1,0) \mid \alpha_i, \beta\}}$$

Use that:

$$P\{y_i = (0,1) \mid \alpha_i, \beta\} = P\{y_{i1} = 0 \mid \alpha_i, \beta\} P\{y_{i2} = 1 \mid \alpha_i, \beta\}$$

And:

$$P\{y_{i2} = 1 \mid \alpha_i, \beta\} = \frac{\exp\{\alpha_i + x_{i2}'\beta\}}{1 + \exp\{\alpha_i + x_{i2}'\beta\}}$$

$$P\{y_{i1} = 0 \mid \alpha_i, \beta\} = 1 - \frac{\exp\{\alpha_i + x_{i1}'\beta\}}{1 + \exp\{\alpha_i + x_{i1}'\beta\}}$$

Then, after some algebra (do it as exercise), it follows that the conditional probability will be given by:

$$P\{y_i = (0,1) \mid \bar{y}_i = 1/2, \alpha_i, \beta\} = \frac{\exp\{(x_{i2} - x_{i1})'\beta\}}{\exp\{(x_{i2} - x_{i1})'\beta\} + 1} = \frac{\exp\{(x_i^*)'\beta\}}{\exp\{(x_i^*)'\beta\} + 1}$$

Which does not depend on α_i. The last looks exactly as a simple logit regression, where x_i^* is in fact the first difference for x_i. In an analogous way you may obtain $P\{y_i = (1,0) \mid \bar{y}_i = 1/2, \alpha_i, \beta\}$.

In summary, the estimator consists in the following: keep only the observations with (y_1, y_2) equal to (0,1) or (1,0). Then generate a dependent variable taking the value 1 for positive changes (0,1), and the value 0 for negative changes (1,0). Then, regress (by an ordinary logit) the transformed dependent variable on the first differences for the regressors ($x_i^* = x_{i2} - x_{i1}$). You can obtain a similar result for T>2.

An example

For this exercise we will use a small panel (T=3) extracted from a Russian database.

```
. use russia1, clear
```

The command “xtlogit” fits random-effects, conditional fixed-effects, and population-averaged logit models:

```
. xtlogit smokes monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2 round2 round3, fe i(eid)
```

You can use the post-estimation command “mfx” (presented in the notes on the binary logit/probit models):

```
. mfx, predict(pu0)
```

It calculates the marginal effects where the explanatory variables are set at their mean values and the fixed effect is set at zero.

We can compare the results to the pooled logit:

```
. xtlogit smokes monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2 round2 round3, fe i(eid)
```

```
. logit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2 round2 round3
```

You will notice that several variables that in the simple model appeared as significant, now including fixed effects are not statistically different from zero any more. Before we studied the association between some variables and the fact that some individuals smoked and others did not. However, in the fixed effects model we are studying the association between the inter-temporal change in some variables and the fact that some individuals have decided to either stop or start smoking. Now you can see that this fixed effects estimator is equivalently to the first-differences in the OLS model.

However, some variables stopped being significant just because they have almost no intertemporal variation (such as "monage" and "highsc").

Probit with random effects

The latent variable specification is:

$$y_{it}^{*} = x_{it}'\beta + u_{it}$$

Where u_{it} have mean zero and unit variance, and it can be further decomposed as $\varepsilon_{it} + \alpha_i$. Additionally:

$$\begin{aligned} y_{it} &= 1 \quad if\ y_{it}^{*} > 0 \\ y_{it} &= 0 \quad if\ y_{it}^{*} \leq 0 \end{aligned}$$

The likelihood contribution of individual i will be the joint probability of observing outcomes $(y_{i1},...,y_{iT})$. Such probability is derived from the joint distribution of $(y_{i1}^{*},...,y_{iT}^{*})$. Recall that in the one-period case:

$$P(y_i = 1) = P(u_i > -x_i'\beta) = \int_{-x_i'\beta}^{\infty} f(u)du$$

In the T-periods case, integrating over the appropriate intervals would involve T integrals (which in practice is done numerically). When $T \geq 4$, the maximum likelihood estimation is made infeasible. This “curse of dimensionality” may be avoided by using simulation-based estimators (Verbeek, 2004, Chapter 10.7.3).

If u_{it} could be assumed as independent:

$$f(y_{i1},...,y_{iT} \mid x_{i1},...,x_{iT},\beta) = \prod_t f(y_{it} \mid x_{it},\beta)$$

The estimation procedure would involve only one-dimensional integrals (as in the cross-section case). Thus, if ε_{it} is assumed to be independent over time:

$$\begin{aligned} f(y_{i1},...,y_{iT} \mid x_{i1},...,x_{iT},\beta) &= \int_{-\infty}^{\infty} f(y_{i1},...,y_{iT} \mid x_{i1},...,x_{iT},\alpha_i,\beta) f(\alpha_i)d\alpha_i \\ &= \int_{-\infty}^{\infty} \left[\prod_t f(y_{it} \mid x_{it},\alpha_i,\beta)\right] f(\alpha_i)d\alpha_i \end{aligned}$$

We can use arbitrary assumptions about the distributions of ε_{it} and α_i. If we assumed that u_{it} is distributed as a multivariate normal, then we would obtain the random effects probit model.

An example

The command "xtprobit" does all the work. Let’s continue with the previous example:

```
. xtprobit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2 round2 round3, re i(eid)
```

We can compare the results to the pooled logit:

```
. probit smokes gender monage highsc belief obese alclmo hattac cmedin totexpr tincm_r work0 marsta1 marsta2 round2 round3
```

Notice that if the explanatory variables were exogenous, both estimators would be consistent, but the former would be more efficient.

You can compute the marginal effects where the explanatory variables are set at their mean values and the fixed effect is set at zero:

```
. mfx, predict(pu0)
```

Problem Set #8

Part 1

With the dataset "russia.dta":

1. Estimate a model (probit or logit) for the probability of drinking alcohol. Use different specifications (remember you can always use "outreg"). Analyze the statistical significance of every coefficient. Interpret them qualitatively.

2. Compute the marginal effects (both the "mean marginal effects" and the "marginal effects at the mean values"). Compare the marginal effects between a logit and a probit specification.

3. Estimate a model for male individuals and a model for female individuals. Choose the "typical individuals" (i.e. the most representative male and the most representative female). For each gender, graph the marginal effects of age for different ages. Graph also the predicted probabilities. Explain extensively.

4. Read Menard (2000) thoroughly. Choose two indexes of goodness-of-fit (different from those introduced in the notes), calculate them and comment.

Part 2

For this exercise we will use the dataset "russia1.dta". Consider as dependent variable the decision whether to smoke or not. Include as explanatory variables all the variables you consider relevant, such as age, health conditions, income, expenditures, and so on.

1. Estimate a binary logit model with fixed effects. Present and explain carefully the results. You should want to report (and interpret!), for example, the marginal effects.

2. Obtain the fixed effects estimator "by hand" (i.e. without using the command "xtlogit"). If you do not want to derive the estimator for T=3, just use information on only two periods. Compare your results with those obtained with "xtlogit".

3. Estimate the traditional logit model, but including dummies for individuals. Show that it yields different estimates than the previous model. Explain why.

4. Choose some of the coefficients and comment the differences between the fixed effects and the “pooled” specification (use your economic intuition!!).

5. Estimate the probit and logit models with random effects. Are the results more similar to the pooled or the fixed effects model? If you care a lot about consistency (as you should!), would you use the random effects models?

Chapter 9
Hazard Models

Introduction to Hazard functions

Sometimos you are interested in explaining the duration of a certain event. For instance, you may be interested in explaining the time it takes for a bank to be either dissolved or acquired by another bank, the duration of unemployment, and so on.

Let T denote the time spent in the initial state. In the above example, T would be the number of months until a bank is dissolved or acquired. Since T is continuous, the distribution of T is expressed by the following cumulative density function:

$$\mathrm{F(t) = P(T \leq t)}$$

The survivor function is the probability surviving past t and is defined as:

$$\mathrm{S(t) = 1 - F(t) = P(T > t)}$$

Having survived up to time t, the likelihood of leaving the initial state within the time interval t until $\mathrm{t+h}$ may be written as follows:

$$\mathrm{P(t \leq T < t+h \,|\, T \geq t)}$$

Dividing by h we can obtain the average likelihood of leaving the initial state per period of time over the interval $\mathrm{[t, t+h)}$. Making h tend to zero we get the hazard function (defined as the instantaneous rate of leaving the initial state):

$$\lambda(\mathrm{t}) = \lim_{h \to 0^+} \frac{\mathrm{P(t \leq T < t+h \,|\, T \geq t)}}{h}$$

The hazard and survival functions provide alternative but equivalent characterizations of the distributions of T. To see this, rewrite the conditional probability as follows:

$$\mathrm{P(t \leq T < t+h \,|\, T \geq t)} = \frac{\mathrm{P(t \leq T < t+h)}}{\mathrm{P(T \geq t)}} = \frac{F(t+h) - F(t)}{1 - F(t)}$$

And then notice that:

$$\lim_{h \to 0^+} \frac{F(t+h) - F(t)}{h} = F'(t) = f(t)$$

Finally:

$$\lambda(\mathrm{t}) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

Additionally, it is easy to show that (Verbeek, 2004):

$$\mathrm{F(s)} = 1 - \exp(-\int_0^{\mathrm{s}} \lambda(\mathrm{t})\mathrm{dt})$$

For instance, consider the simplest case: the hazard rate is constant, $\lambda(\mathrm{t}) = \lambda$. It implies that T follows the exponential distribution: $\mathrm{F(t)} = 1 - \exp(-\lambda \mathrm{t})$.

Let $\mathrm{x_i}$ be a vector of explanatory variables. A widespread group of models are the so-called proportional hazard models. The idea behind those models is that the hazard function can be written as the product of a baseline hazard function that does not depend on $\mathrm{x_i}$, and a individual-specific non-negative function that describes the effect of $\mathrm{x_i}$:

$$\lambda(\mathrm{t}, \mathrm{x_i}) = \lambda_0(\mathrm{t})\exp(\mathrm{x_i'}\beta)$$

Where λ_0 is the baseline hazard function (i.e. that of an hypothetical individual with $\mathrm{x_i} = 0$), and $\exp(\mathrm{x_i'}\beta)$ is the proportional term that stretches and shrinks the baseline function along the y axis (because the adjustment is the same for every t). Take logarithm and then take the derivative with respect to $\mathrm{x_{ik}}$:

$$\frac{\partial \ln \lambda(\mathrm{t}, \mathrm{x_i})}{\mathrm{x_{ik}}} = \beta_\mathrm{k}$$

The coefficient β_k is the (proportional) effect of a change in $\mathrm{x_k}$ on the hazard function. When $\mathrm{x_k}$ is increased, the hazard function is stretched if $\beta_\mathrm{k} > 0$ (i.e. the instantaneous likelihood of leaving the initial state increases for every t), and if $\beta_\mathrm{k} < 0$ it is shrunk.

An example

Open the database “lung.dta” and see its content:

. use lung, clear

. describe

. browse

First of all, use “stset” to declare the data to be survival-time data. You have to indicate the permanence variable ("time") y and the failure variable ("dead"):

```
. stset time, failure(dead)
```

We can already draw a Kaplan-Meier non-parametric survival estimator (see for example Lancaster, 1990):

```
. sts graph, by(sex)
```

We are going to estimate non-proportional models (notice that x_{i} cannot be time-varying). The models with a proportional hazard ratio parameterization included in the command "streg" are those with the following baseline hazard functions: exponential, Weibull, and Gompertz. For instance, if $\lambda_0(t)$ follows a Weibull distribution:

```
. streg age sex wt_loss, nohr distribution(weibull)
```

The "nohr" option means that coefficient estimates must be shown. This option affects only how results are displayed, not how they are estimated. As we already explained, when x_{k} is increased the hazard function is stretched if $\beta_{\mathrm{k}} > 0$ and shrunk if $\beta_{\mathrm{k}} < 0$. If we had not included the "nohr" option, the hazard ratios ($= \exp(\beta_{\mathrm{k}})$) would have been shown instead of the coefficients:

```
. streg, hr
```

The estimates imply that, at each survival time, the hazard rate for females is only 60% of the hazard rate for males.

We can also consider a non-parametric baseline hazard, which would suggest the non-parametric Cox model:

```
. stcox age sex wt_loss, robust basesurv(base_survival)
```

The command "stcurve" plots the survival, hazard, or cumulative hazard function after "stcox" or "streg":

```
. stcox age sex wt_loss, basesurv(hazard_aux1)

. stcurve, survival

. stcox age sex wt_loss, basehc(hazard_aux2)

. stcurve, hazard

. streg age sex wt_loss, distribution(weibull)

. stcurve, survival

. stcurve, hazard
```

Additionally, we can graph the survival function evaluated at a particular point of the explanatory variables:

```
. streg age sex wt_loss, distribution(weibull)

. stcurve, survival at(age=62.44, sex=1, wt_loss=9.83)

. stcurve, survival at1(sex=0) at2(sex=1)
```

Estimating hazard functions using the logit model

Jenkins (1995) points out that the likelihood functions of a logit model and a duration model (under some sampling schemes) are pretty similar. As a consequence, the discrete-time duration model based on data derived from some sampling schemes can be estimated as a regression model for a binary dependent variable.

We only need to rearrange the database. Each individual in the database must have as many observations as time intervals that it spent in the sample. We can use the command "expand":

```
. clear

. set mem 10m

. use lung

. expand time
```

We need a time variable (i.e. the first observation for individual i will correspond to its first day in the simple, the second observations will correspond to its second day in the sample, and so on).

```
. sort id

. quietly by id: gen day = _n
```

Now we must simple generate the new dependent variable ("death"). For every observation different than the last one (for each individual), it must take the value 0. For the last observation of each individual, “death” must take the value 0 if the individual did not die (i.e. it was censored) or 1 otherwise (if died).

```
. gen death=0

. quietly by id: replace death = dead if _n==_N
```

We can now see the results:

```
. browse id day death
```

We need to create a time-control. If we wanted to create the correlate for the non-parametric hazard model, we would need to create a set of dummies for time. Or we can simply use a log specification:

```
. gen ln_t=log(time)
```

Finally, we can estimate the model:

```
. logit death age sex wt_loss ln_t
```

Problem Set #9

Use the fictional database “duration.dta” on cancer patients. Estimate a hazard model, and then estimate the model using the strategy proposed by Jenkins (1995).

Chapter 10

Count-Data Models

Count-data Models

The central characteristics of a count-dependent-variable are: i. It must be a positive integer, including zero; ii. It has not an obvious maximum or upper limit; iii. Most of the values must be low, and particularly there must be lots of zeros.

Consider the following example, illustrated by the same database that we will use thoroughly the notes: the visits to the doctor in the last two weeks ("count.dta").

. use count, clear

. tab doctorco

. hist doctorco, discrete

Visits to the doctor in the last two weeks (doctorco) - Actual frequency distribution

Count	0	1	2	3	4	5	6	7	8	9	Total
Frequency	4,141	782	174	30	24	9	12	12	5	1	5,190
Percent	79.79	15.07	3.35	0.58	0.46	0.17	0.23	0.23	0.1	0.02	100

Notice that the "count nature" is clear: i. The variable only takes integer values, including the zero; ii. There is no obvious upper limit; iii. The 95% of the observations take values either 0 or 1, and 80% of the sample is composed by zeros.

The Poisson Regression Model

We need to estimate an appropriate model for $E(y\,|\,x)$, where x is a vector of explanatory variables. The model will comply:

$$E(y\,|\,x) = \exp(x'\beta)$$

Which guarantees $E(y\,|\,x) > 0$ and provides a rather simple interpretation for the marginal effects:

$$\frac{\partial \ln E(y\,|\,x)}{\partial x_k} = \beta_k$$

Then, the coefficients are semi-elasticities (notice that the coefficients on binary variables will be read as proportional changes).

Remember that a random variable has a Poisson distribution if:

$$f(y) = P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} \quad y = 0,1,2,...$$

It is straightforward to show that:

$$E(Y) = V(Y) = \mu$$

The property that the expectation equals the variance is called equidispersion. The Poisson regression model corresponds to:

$$f(y \mid x) = \frac{e^{-\mu(x)}\mu(x)^y}{y!} \quad y = 0,1,2,...$$

Where $\mu(x) = \exp(x'\beta)$ and x is a vector of K explanatory variables (including an intercept). Consequently: $E(Y \mid x) = \exp(x'\beta)$.

Maximum likelihood estimation of the Poisson Model

Suppose $y \mid x \square \, Po(\mu = \exp(x'\beta_0))$, and that we have a random sample i.i.d. (y_i, x_i) with i=1,…,n. The maximum-likelihood function is:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

With $\mu_i = \exp(x_i'\beta)$. Take logarithm:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i x_i'\beta - \exp(x_i'\beta) - \ln y_i! \right]$$

Then $\hat{\beta}$ is such that:

$$\sum_{i=1}^{n} \left[y_i - \exp(x_i'\hat{\beta}) \right] x_i = 0$$

Like in the logit model, there is no explicit solution and you must find the estimates numerically. Under the right specification the estimator is consistent, asymptotically normal and efficient.

Furthermore, the Poisson estimator is a Quasi-MLE. If $E(Y \mid x) = \exp(x'\beta)$ holds, then $\hat{\beta}$ is consistent and asymptotically normal for any underlying distribution. However, we will need to use the robust estimate of the variance-covariance matrix to make valid inference.

Exercise 10.1: Obtain the maximum-likelihood estimates "by hand" (i.e. using the command "ml").

An example: Visits to the doctor

We will use the database on visit to the doctor in the last two weeks, along with demographic, economic and health variables. Enter "describe" to see the variables' label with a description of the data:

```
. describe
```

The "poisson" command performs the estimation:

```
. poisson doctorco sex age agesq income levyplus freepoor freerepa illness actdays hscore chcond1 chcond2, robust
```

As always, we utilize the option "robust" for obtaining robust standard errors. As we anticipated, the marginal effects are easily interpretable. For instance, the coefficient on "income" indicates that an increase in ten thousands dollars in annual income is associated with a 20% decrease in the visits to the doctor. Remember that we are always speaking about associations. We would need an extremely exhaustive structural model or a clever natural experiment to infer causality.

The post-estimation command "estat gof" performs a goodness-of-fit test of the model. If the test is significant, the Poisson regression model is inappropriate. In this case, you could try a negative binomial model.

```
. estat gof
```

Negative Binomial Regression Model

The Poisson distribution is a particular case of the Negative Binomial distribution. Equivalently, the Poisson Regression Model is a particular case of the Negative Binomial Regression Model. In the negative binomial (maximum-likelihood) regression model the count variable is believed to be generated by a Poisson-like process, except that the variation is greater than that of a true Poisson. This extra variation is referred to as overdispersion. You can refer to Rainer Winkelrmann's "Econometric Analysis of count data" to see some theoretical and practical details.

The command for the NB regression is "nbreg":

```
. nbreg doctorco sex age agesq income levyplus freepoor freerepa illness actdays hscore chcond1 chcond2
```

If the parameter "alpha" is null, then the Poisson model is right. In order to test the Poisson suitability, Stata construct an LR-test comparing the log-likelihood values attained in each model. The result seems natural if you see the difference between the mean and the variance of the (raw) dependent variable:

```
. sum doctorco
```

See "xtpoisson" and "xtnbreg" for closely related panel estimators.

Problem Set #10

Open the database "couart.dta", from Scott Long (1997). The dependent variable is the number of articles in last 3 years of PhD for a group of 915 American graduate students. Enter "describe" to see the data definitions.

1. Is "art" a count-variable?

2. Run the Poisson Regression Model. Interpret the coefficients.

3. Run a Negative Binomial Regression Model. Test if the Poisson model is suitable.

Chapter 11
Selection Models

The internal validity of the econometric models is commonly threatened by selection problems. Consider the following example:

$$y^* = x'\beta + u$$

Call s to the selection variable, which takes the value 1 if y^* is observed, and 0 otherwise. Imagine a super-sample (y_i^*, x_i, s_i) of size N, and that we only observe the sub-sample (y_i^*, x_i) for those with $s = 1$. Consider as an example a wage equation for women: we only observe wages for those women who work.

If we had an i.i.d. sample (y_i^*, x_i), the consistency would depend on $E(u \mid x) = 0$. The problem is that we now have a sample conditional on s = 1. Take expectation:

$$E(y \mid x, s = 1) = x'\beta + E(u \mid x, s = 1)$$

Then OLS using the sub-sample will be inconsistent unless:

$$E(u \mid x, s = 1) = 0$$

Notice that not any selection mechanism makes OLS inconsistent. If u is independent from s, then OLS is consistent. Additionally, if the selection depends only on x then OLS is also consistent.

A selectivity model

Consider the following system of equations:

$$\begin{cases} y_{1i} = x'_{1i}\beta_1 + u_{1i} \\ y^*_{2i} = x'_{2i}\beta_2 + u_{2i} \end{cases}$$

Called the regression equation and the selection equation, respectively. Define the binary variable $y_{2i} \equiv 1\left[y^*_{2i} > 0\right]$. The pair (y_{2i}, x_{2i}) is observed for the entire sample. However, the pair (y_{1i}, x_{1i}) is observed iff $y_{2i} = 1$ (called the selected sample).

For instance, in our previous example y_{1i} would be the wages for women, x_{1i} would be the determinants of wages, y^*_{2i} would be the net utility from working, and x_{2i} would be the determinants of the latter. The $y_{2i} = 1$ would indicate whether the woman decided to work (if she had positive net utility from working), in which case we would observe her wage (which would be unobservable otherwise).

Let's assume that (u_{1i}, u_{2i}) are independent from x_{2i} and have mean zero. In addition, suppose that u_{2i} □ $N(0, \sigma_2^2)$. Finally, assume:

$$E\left[u_{1i} \mid u_{2i}\right] = \gamma u_{2i}$$

This allows the non-observable from both equations to be related. Take expectation:

$$\begin{aligned}
E(y_{1i} \mid x_{1i}, y_{2i} = 1) &= x_{1i}'\beta_1 + E\left[u_{1i} \mid x_{1i}, y_{2i} = 1\right] \\
&= x_{1i}'\beta_1 + E\left[E(u_{1i} \mid u_{2i}) \mid x_{1i}, y_{2i} = 1\right] \\
&= x_{1i}'\beta_1 + E\left[\gamma u_{2i} \mid x_{1i}, y_{2i} = 1\right] \\
&= x_{1i}'\beta_1 + \gamma E\left[u_{2i} \mid x_{1i}, y_{2i}^* > 0\right] \\
&= x_{1i}'\beta_1 + \gamma E\left[u_{2i} \mid x_{1i}, u_{2i} < x_{2i}'\beta_2\right] \\
&= x_{1i}'\beta_1 + \gamma\lambda(x_{2i}'\beta_2 / \sigma_2) \\
&= x_{1i}'\beta_1 + \gamma z_i \\
&\neq x_{1i}'\beta_1
\end{aligned}$$

Where $z_i = \lambda(x_{2i}'\beta_2 / \sigma_2)$. If we regressed by OLS using the selected sample we would be including γz_i in the error term. If z_i and x_{1i}' were correlated, and if in addition u_{1i} and u_{2i} were correlated ($\gamma \neq 0$), then OLS would be inconsistent.

Two-stage consistent estimator

Define:

$$u_{1i}^* \equiv y_{1i} - x_{1i}'\beta_1 - \gamma z_i$$

And write:

$$y_{1i} \equiv x_{1i}'\beta_1 + \gamma z_i + u_{1i}^*$$

Where, by definition:

$$E\left[u_{1i}^* \mid x_{1i}, y_{2i} = 1\right] = 0$$

If x_{1i} and z_i were observable when $y_{2i} = 1$, then regressing by OLS y_{1i} on x_{1i} and z_i (using the selected sample) would yield consistent estimates for β_1 and γ. The problem is that z_i is not observable, though it could be calculated from σ_2 and β_2.

Given that u_{2i} □ $N(0, \sigma_2^2)$:

$$P(y_{2i} = 1) = P(y_{2i}^* = 1) = P(u_{2i}/\sigma_2 < x_{2i}'\beta_2/\sigma_2) = \Phi(x_{2i}'\delta)$$

Then $P(y_{2i} = 1)$ corresponds to a probit model with coefficient δ. If x_{2i} and y_{2i} are observed for the complete sample, then δ can be estimated consistently using a probit model (notice that despite we can identify δ, we cannot identify β_2 and σ_2 separately).

Finally, β_1 and γ can be consistently estimated using the following two-step procedure:

First stage: obtain an estimate for $\hat{\delta}$ using the probit model $P(y_{2i}=1)=\Phi(x'_{2i}\delta)$ for the complete sample. Then estimate z_i using $\hat{z}_i=\lambda(x'_{2i}\hat{\delta})$.

Second stage: Regress y_{2i} on x_{1i} and $\hat{z}_i$ utilizing the censored sample, which should yield consistent estimates for β_1 and γ.

It can be shown that the second stage is heterocedastic by construction. You can derive robust standard errors, though the adjustment is not as simple as in the sandwich estimator (it would be if z_i were observable). Nevertheless, Stata computes it automatically.

An example: wage equation for women

The objective of this example is estimating a simple wage equation for women. The database is the same presented for the Tobit example. Enter "describe" to see the data definitions:

```
. use women, clear
```

```
. describe
```

The variable "hwage" represents the hourly salary income (in US dollars), which must be set to missing if the person does not work (since the "heckman" command identifies with missing values those observations censored):

```
. replace hwage=. if hwage==0
```

Then we can use the command "heckman". Inmmediately after it you must specify the regression equation, and in the option "select(·)" you must specify the explanatory variables for the selection equation:

```
. heckman hwage age agesq exp expsq married head spouse headw spousew pric seci secc supi supc school, select(age agesq exp expsq married head spouse headw spousew pric seci secc supi supc school) twostep
```

With the option "twostep" it fits the regression model using the Heckman's two-step consistent estimator. Additionally, you can use different specifications for the regression and selection equations:

```
. heckman hwage age agesq exp expsq pric seci secc supi supc, select(age agesq exp expsq married head spouse headw spousew pric seci secc supi supc school) twostep
```

Selection Test

You can compare the estimate with the (seemingly inconsistent) OLS regression:

```
. reg hwage age agesq exp expsq pric seci secc supi supc, robust
```

The differences are indeed minuscule. This is not surprising, since the inverse Mills ratio term is statistically not different from zero. Testing $H_0 : \gamma = 0$ provides a simple test for selection bias. Under H_0 the regresión model with the selectioned samples is homocedastic, and then you can perform it without correcting for heteroskedasticity.

```
. heckman hwage age agesq exp expsq pric seci secc supi supc, select(age agesq exp
expsq married head spouse headw spousew pric seci secc supi supc school) twostep
```

```
. test lambda
```

Two stages "by hand"

```
. gen s=(hwage!=.)
```

```
. probit s age agesq exp expsq married head spouse headw spousew pric seci secc
supi supc school
```

```
. predict xb, xb
```

```
. gen lambda = normden(xb)/normal(xb)
```

```
. regress hwage age agesq exp expsq pric seci secc supi supc lambda if hwage>0,
robust
```

And compare the estimates to those obtained using "heckman":

```
. heckman hwage age agesq exp expsq pric seci secc supi supc, select(age agesq exp
expsq married head spouse headw spousew pric seci secc supi supc school) twostep
```

Notice that obtaining the two-stage estimates "by hand" you cannot use the standard errors directly.

Maximum-likelihood estimation

Under the (more restrictive) assumption that (u_{1i}, u_{2i}) follows a bivariate normal distribution it is possible to construct a maximum-likelihood estimator. If you do not specify the "twostep" option, then Stata fits such an estimator:

```
. heckman hwage age agesq exp expsq married head spouse headw spousew pric seci
secc supi supc school, select(age agesq exp expsq married head spouse headw
spousew pric seci secc supi supc school)
```

This strategy is usually discarded. See for example Nawata y Nagase (????). In the Problem Set you will be asked to compare both estimations.

Drawbacks

The classical problem with the Heckman model is the high correlation between x_{1i} and $\hat{z}_i$. Since the function $\lambda(\cdot)$ is monotonously increasing, if its argument has little variation then it may resemble a linear function. If x_{1i} is similar to x_{2i}, then the correlation between x_{1i} an $\hat{z}_i$ may be high. Indeed, the identification of γ needs strictly $\lambda(\cdot)$ not to be linear. If x_{1i} and x_{2i} have a lot of variables in common, then the second stage will be subject to a problem of high multicollineality. The final result are insignificant coefficients on γ and β_1.

We can obtain a measure of how bad this problem could be. First estimate the Heckman model using the option "mills(newvar)", which stores $\lambda(\cdot)$ in "newvar":

```
. heckman hwage age agesq exp expsq pric seci secc supi supc, select(age agesq exp expsq married head spouse headw spousew pric seci secc supi supc school) twostep mills(mills)
```

And then regress $\lambda(\cdot)$ on x_{1i}:

```
. reg mills age agesq exp expsq pric seci secc supi supc
```

The higher the resulting R^2, the higher the multicollineality problem. In this particular example high multicollineality does not seem to be an issue.

Marginal effects

The marginal effects for the expected value of the dependent variable conditional on being observed, E(y | y observed), are:

```
. mfx compute, predict(ycond)
```

The marginal effects for the probability of the dependent variable being observed, Pr(y observed), are:

```
. mfx compute, predict(psel)
```

Problem Set #11

Open "mroz.dta". Enter "describe" to see the data definitions. It has been extracted from Wooldridge "Econometric Analysis of Cross Section and Panel Data", examples 16.3 and 17.6 (page 529 and 565, respectively). In the Tobit Model we will estimate a reduced form annual hours equation for married women. Of the 753 women in the sample, 428 worked for a wage outside the home during the year and 325 of the women worked zero hours.

1. The goal is the estimation of a wage offer equation for married women, employing usual control variables. The dependent variable is the logarithm of salary income. Explain (intuitively) why you would try a Heckman model instead of using OLS directly. Obtain the Heckman estimates. Compare your estimates to those obtained by using OLS directly.

2. Experiment using different specifications for the first and second stage.

3. Estimate also the maximum-likelihood version of the estimator. Explain their differences in consistency and efficiency.

4. Reproduce the two-stage estimator by hand.

5. Test the hypothesis of whether there is selection.

Chapter 12

Partially continuous variables

Some random variables are discrete: each value in the support can happen with positive probability (e.g. bernoulli, poisson). Some random variables are continuos: each point in the support happens with probability zero, but intervals happens with positive probability (e.g. normal, chi-squared). But some variables are partially continuous: some points in the support have positive probability.

Consider expenditure in tobacco as an example: an individual may not smoke with positive probability; but given that he/she smokes, then the tobacco expenditure is a continuous variable (i.e. the probability that he/she will expend exactly $121,02 is null).

This discontinuity usually responds to either censorship or corner solutions. Consider for example the following simple model:

$$\begin{cases} y^* = x'\beta + u \\ y = \max(0, y^*) \end{cases}$$

With $E(u \mid x) = 0$. The variable y may be expenditure in food, and y^* (the latent variable) food consumption. If we observed an i.i.d. sample of (y_i^*, x_i); i = 1, …, n, then we would be able to estimate consistently β regressing y_i^* on x_i by OLS.

However, we must find out what happens when we have data only on (y_i, x_i). Following our previous example, the surveys measure expenditures in different goods, but not consumption. We may regress y_i on x_i for the entire sample, or we may regress y_i en x_i only for those observations with $y_i > 0$. We will show that both strategies lead to inconsistent estimators.

Consider first the estimation of β by OLS regressing y_i on x_i only for the observations with $y_i > 0$ (the truncated sample). We need to check out what happens with $E(y \mid x, y > 0)$. If $u_i \sim N(0, \sigma^2)$:

$$y \mid x, y > 0 = x'\beta + \left(u \mid x, y > 0\right)$$

$$\begin{aligned} E(y \mid x, y > 0) &= x'\beta + E(u \mid x, y > 0) \\ &= x'\beta + E(u \mid x, u > -x'\beta) \\ &= x'\beta + \sigma \frac{\phi(-x'\beta/\sigma)}{1 - \Phi(-x'\beta/\sigma)} \\ &= x'\beta + \sigma \frac{\phi(x'\beta/\sigma)}{\Phi(x'\beta/\sigma)} \end{aligned}$$

$$= x'\beta + \sigma\lambda(x'\beta / \sigma)$$

Where $\lambda(z) = \phi(z)/\Phi(z)$ is known as the inverse Mills ratio. If we regressed y_i on x_i then $\sigma\lambda(x'\beta / \sigma)$ would appear in the error term. Since the latter is correlated with x_i, the OLS estimate on β would be inconsistent.

We must prove that OLS also yields inconsistent estimates if using the censored sample:

$$E(y \mid x) = E(y \mid x, y > 0)P(y > 0 \mid x) + E(y \mid x, y \leq 0)P(y \leq 0 \mid x)$$

$$= E(y \mid x, y > 0)P(y > 0 \mid x)$$

$$= \left[x'\beta + \sigma\lambda(x'\beta / \sigma)\right] P(u > -x'\beta)$$

$$= \left[x'\beta + \sigma\lambda(x'\beta / \sigma)\right] \Phi(x'\beta/\sigma)$$

Maximum-likelihood estimation

The model with censored data with normal errors is known as Tobit Model. Under the normality assumption it is relatively easy to obtain a consistent maximum-likelihood estimator. Let's begin by dividing the sample (y_i, x_i) in pairs with $y_i = 0$ and pairs with $y_i > 0$. The first happen with probability $P(y_i = 0 \mid x_i)$, while the second follow the density distribution $f(y_i \mid x_i, y_i > 0)$. Define $w_i \equiv 1\left[y_i > 0\right]$. Then:

$$L(\beta) = \prod_{i \mid y_i = 0} P(y_i = 0) \prod_{i \mid y_i > 0} f(y_i \mid x_i, y_i > 0)$$

$$= \prod_{i=1}^{n} \left[P(y_i = 0)^{(1-w_i)} f(y_i \mid x_i, y_i > 0)^{w_i} \right]$$

$$l(\beta) = \sum\nolimits_{i=1}^{n} \left[(1 - w_i) \ln\left(1 - \Phi(x'\beta_i/\sigma)\right) + w_i \ln\left((1/\sigma)\phi((y_i - x_i'\beta)/\sigma\right) \right]$$

Under standard conditions the maximization problem has a unique solution, which can be easily retrieved using maximization algorithms.

Interpreting the results

The maximum-likelihood coefficients have the following interpretation:

$$\frac{\partial E(y^* \mid x)}{\partial x} = \beta$$

They are the partial derivatives on the latent variable (e.g. consumption). But sometimes the interest must be focused on different marginal effects. Recall that:

$$E(y \mid x) = E(y \mid x, y > 0)P(y > 0 \mid x) = \left[x'\beta + \sigma\lambda(x'\beta)\right] \Phi(x'\beta/\sigma)$$

Using the chain-rule:

$$\frac{\partial E(y|x)}{\partial x_k}=\frac{\partial E(y|x,y>0)}{\partial x_k}P(y>0|x)+E(y|x,y>0)\frac{\partial P(y>0|x)}{\partial x_k}$$

The above is the Donald-Moffit decomposition. It is straightforward to show that:

$$\frac{\partial E(y|x)}{\partial x_k}=\beta_k\Phi(x'\beta/\sigma)$$

This is a rather simple re-scaling of the original estimate. However, as in the logit model, the marginal effects depend on x. In our example, this would be the marginal derivative on expenditures: part of the effect comes from the families that start expending because of the change in x, and the other part comes from the increase in expenditures for those families that were already expending.

We can also obtain the following:

$$\frac{\partial E(y|x,y>0)}{\partial x_k}=\beta_k\left[1-\lambda(x'\beta/\sigma)\left[(x'\beta/\sigma)+\lambda(x'\beta/\sigma)\right]\right]$$

It is also a re-scaling. This would be the effect on expenditures only for those families who were already expending something.

An example: wage determination for women

We want to estimate a rather simple model on the determination of income for women. We have data ("women.dta") on 2556 women between 14 and 64 years old living in Buenos Aires in 1998.[2] Enter "describe" to see the data definitions:

```
. use women, clear
```

```
. describe
```

Then, regressing wages on some individual characteristics we can estimate the effect of those variables on the wages' determination. The variable "hwage" stores the hourly wage in US dollars, and it takes the value zero if the woman does not work. As potential explanatory variables and controls we have variables on education, age, number of kids, marital status, etc.

Regarding our theoretical model, the y would be the "potential" wage (e.g. if compelled to work, a woman may be paid less than her reservation wage), and y* would be the "actual" wage (e.g. if a woman is offered less than her reservation wage, then she will not work and she will earn nothing).

Let's estimate the following wage determination equation (inconsistently) by OLS using the censored sample and the truncated sample:

[2] Extracted from an Argentinean household panel called "Encuesta Permanente de Hogares", and provided by Walter Sosa Escudero.

```
. reg hwage age agesq exp expsq married head spouse children hhmembers headw
spousew pric seci secc supi supc school, robust
```

```
. reg hwage age agesq exp expsq married head spouse children hhmembers headw
spousew pric seci secc supi supc school if hwage>0, robust
```

Now we will retrieve the Tobit coefficients:

```
. tobit hwage age agesq exp expsq married head spouse children hhmembers headw
spousew pric seci secc supi supc school, ll(0)
```

With the option "ll(0)" we are indicated that the sample is left-censored at zero. We know that $\beta = \partial E(y^* \mid x) / \partial x$ is the marginal effect of x on the (conditional expected value of) "potential" wages. For instance, the impact of being married on the "latent" wage is -8.6. Following the previously introduced formulas, we can also estimate the marginal effect of x on the expected "actual" wage (at the mean x), and the marginal effect of x on the expected "actual" wage conditional on being uncensored (at the mean x).

We will take advantage of the command "dtobit", which you must install first. For instance:

```
. version 6.0
```

```
. tobit hwage age agesq exp expsq married head spouse children hhmembers headw
spousew pric seci secc supi supc school, ll(0)
```

```
. dtobit
```

```
. version 9.1
```

Notice that we need to specify "version 6.0" before running the Tobit model. The reason is that in version 6.0 Stata stores the estimate for sigma, which is necessary to carry on "dtobit". If you do not want to calculate the marginal effects at the mean x, you may specify the values for x using the option "at(·)".

For instance, the effect of being married on the "actual" hourly salary income is -3, while the effect on the "actual" hourly salary income for working women is only -2.49. The differences with respect to the latent variable are considerable.

Notice that the latter two estimates are respectively the "consistent" versions of the OLS regressions performed at the beginning of the example. Comparing them you can make up an idea of how much biased were those early estimates (in this example, considerably).

Exercise 12.1: Reproduce the latter two marginal effects "by hand" (i.e. as if the command "dtobit" did not exist).

Exercise 12.2: Obtain the maximum-likelihood estimates "by hand" (i.e. using the command "ml").

Exercise 12.3: Choose a continuous explanatory variable. Then graph a scatterplot of that variable and hourly wages, and add the regression lines for each one of the estimated coefficients showed above (including the inconsistent OLS estimates).

For the marginal effects, you can use the command “mfx” as well. The marginal effects for the probability of being uncensored are obtained in the following way: “mfx compute, predict(p(a,b))”, where “a” is the lower limit for left censoring and “b” is the upper limit for right censoring. In our example:

```
. mfx compute, predict(p(0,.))
```

The marginal effects for the expected value of the dependent variable conditional on being uncensored and the marginal effects for the unconditional expected value of the dependent variable are obtained respectively:

```
. mfx compute, predict(e(0,.))
```

```
. mfx compute, predict(ys(0,.))
```

LR Test

Let $L(\theta)$ be the log-likelihood function, let $\hat{\theta}$ be the unrestricted estimator, and let $\overline{\theta}$ be the estimator with the Q nonredundant constraints imposed (e.g. with less explanatory variables). Then, under the regularity conditions, the likelihood-ratio (LR) statistic $LR = 2[L(\hat{\theta}) - L(\overline{\theta})]$ is distributed asymptotically as χ^2_Q under H_0.

For instance, the following model reaches a log-likelihood of about -4513.77:

```
. tobit hwage age agesq exp expsq married head spouse headw spousew pric seci secc supi supc school, ll(0)
```

If the variables "hhmembers" and "children" are added, then the log-likelihood becomes about -4496.83:

```
. tobit hwage age agesq exp expsq married head spouse children hhmembers headw spousew pric seci secc supi supc school, ll(0)
```

The likelihood ratio statistic is about 2(-4513.77 – (-4496.83)) = -33.88. As it follows a χ^2_2, the implied p-value is practically zero (enter "display 1-chi2(2,33.89)"). Therefore, these two variables are jointly significant.

As in every maximum-likelihood model, "lrtest", which does all the calculation above, is offered as a postestimation command:

```
. tobit hwage age agesq exp expsq married head spouse children hhmembers headw spousew pric seci secc supi supc school, ll(0)
```

```
. est store A
```

```
. tobit hwage age agesq exp expsq married head spouse headw spousew pric seci secc supi supc school, ll(0)
```

```
. lrtest A
```

Tobit with random effects

This model is very similar to the probit with random effects. Rewrite the model:

$$y_{it}^{*} = x_{it}'\beta + \alpha_i + \varepsilon_{it}$$

Where α_i and ε_{it} are i.i.d. normally distributed, independent of $(x_{i1},...,x_{iT})$ with zero means and variances σ_α^2 and σ_ε^2. As in the original model:

$$\begin{aligned} y_{it} &= y_{it}^{*} \quad if \;\; y_{it}^{*} > 0 \\ y_{it} &= 0 \quad if \;\; y_{it}^{*} \leq 0 \end{aligned}$$

As in the random effects probit model, the likelihood contribution of individual i will be the following:

$$f(y_{i1},...,y_{iT} \mid x_{i1},...,x_{iT},\beta) = \int_{-\infty}^{\infty} \left[\prod_t f(y_{it} \mid x_{it},\alpha_i,\beta)\right] f(\alpha_i)d\alpha_i$$

We can obtain the desired model simply by replacing $f(\cdot)$ by the normal density function, and integrating over α_i numerically.

An example

Use the dataset "hinc.dta", which is a quasi-generated sample of individuals based on a British household panel:

```
. clear
```

```
. set mem 50m
```

```
. use hinc
```

See the data definitions:

```
. describe
```

Estimate the model using a pooled Tobit:

```
. tobit hincome educ exper expersq married widowed divorcied, ll(0)
```

And now estimate the model using a Tobit with random effects:

```
. xttobit hincome educ exper expersq married widowed divorcied, i(pid) ll(0)
```

You can use almost all the features available for the pooled Tobit model. You will be required to do so in the Problem Set.

Problem Set #12

Part 1

Open "mroz.dta". It has been extracted from Wooldridge "Econometric Analysis of Cross Section and Panel Data", examples 16.3 and 17.6 (page 529 and 565, respectively). Let's estimate a reduced form annual hours equation for married women. Of the 753 women in the sample, 428 worked for a wage outside the home during the year and 325 of the women worked zero hours.

1. Explain (intuitively) why you would try a Tobit model instead of using OLS directly. Estimate the Tobit model. Can you compare OLS and Tobit coefficients directly?

2. Calculate the partial effects on the conditional expected value of wages for the entire sample and for the sub-sample of working women.

3. Calculate the OLS coefficients for the censored and truncated sample and compare them to the latter.

4. Construct a LR test to test whether the coefficients on "kidslt6" and "kidsge6" are jointly significant.

Part 2

Use the dataset "hinc.dta", which is a quasi-generated sample of individuals based on a British household panel:

1. Estimate the random effects Tobit model. Can you compare the coefficients of least squares and Tobit versions of random effects directly?

2. Calculate all the marginal effects.

Chapter 13

Multinomial and Ordered Logit/Probit

Multinomial Logit (Probit)

The logit model for binary outcomes can be extended to the case where the response has more than two outcomes. For instance: political views (Republican, Independent or Democrat), occupational choice (to work, study or stay at home), the commuting to work (by bus, car, train or bicycle), etc. The individual has to choose only one alternative from the group of choices.

Let y denote a random variable taking on the values {0; 1; . . . ; J} for J a positive integer, and let x denote a set of conditioning explanatory variables (a 1xK vector including a constant). For example, in the commuting to work example y would be {j=0: bus; j=1: car; j=2: train; j=3: bicycle}, and x would probably contain things like income, age, gender, race, marital status and number of children. As usual, the (x_i, y_i) are i.i.d.

We are interested in how ceteris paribus changes in the elements of x affect the response probabilities, $P(y=j\,|\,x)$, j = 0, 1,…, J. Since the probabilities must sum to unity, $P(y=0\,|\,x)$ will be determined once we know the probabilities for j = 1,…, J.

We have to think on the multinomial model as a series of binary models. That is, evaluate the probability of the alternative j against alternative i for every $i \neq j$. For instance, consider the binary model $P(y=j\,|\,y\in\{\text{i,j}\},x)$:

$$\frac{P(y=j\,|\,x)}{P(y=i\,|\,x)+P(y=j\,|\,x)}=\frac{P_j}{P_i+P_j}=F(X\beta_j)$$

Obtenemos:

$$P_j = F(X\beta_j)(P_i+P_j)$$

$$\frac{P_j}{P_i}=F(X\beta_j)\frac{P_i+P_j}{P_i}=\frac{F(X\beta_j)}{\dfrac{P_i}{P_i+P_j}}=\frac{F(X\beta_j)}{1-\dfrac{P_j}{P_i+P_j}}=\frac{F(X\beta_j)}{1-F(X\beta_j)}=G(X\beta_j)$$

Notice that:

$$\sum_{j\neq i}\frac{P_j}{P_i}=\frac{\sum\limits_{j\neq i}P_j}{P_i}=\frac{1-P_i}{P_i}=\frac{1}{P_i}-1$$

$$\frac{1}{P_i}=1+\sum_{j\neq i}\frac{P_j}{P_i}$$

Using the expression for P_j / P_i obtained above:

$$\frac{1}{P_i} = 1 + \sum_{j \neq i} G(X\beta_j)$$

$$P_i = \frac{1}{1 + \sum_{j \neq i} G(X\beta_j)}$$

Finally, solve for P_j:

$$P_j = \frac{G(X\beta_j)}{1 + \sum_{i \neq j} G(X\beta_i)}$$

To find an explicit form for P_j we only have to substitute the $G(\cdot)$ by $\exp(\cdot)$, and then we obtain the multinomial logit model:

$$P(y = j \mid x) = \frac{\exp(X\beta_j)}{1 + \sum_{i \neq j} \exp(X\beta_i)}, \quad j = 1, \ldots, J$$

As the response probabilities must sum to 1, we must set the probability of the reference response (j=0) to:

$$P(y = 0 \mid x) = \frac{1}{1 + \sum_{i \geq 1} \exp(X\beta_i)}$$

The ML are obtained through maximum-likelihood estimation:

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=0}^{J} P(y_i = j \mid x)^{1[y_i = j]}$$

McFadden (1974) has shown that the log-likelihood function is globally concave, what makes the maximization problem straightforward.

The partial effects for this model are complicated. For continuous x_k, we can express:

$$\frac{\partial P(y = j \mid x)}{\partial x_k} = P(y = j \mid x) \left[\beta_{jk} - \frac{\sum_{i \geq 1} \beta_{ik} \exp(X\beta_i)}{1 + \sum_{i \geq 1} \exp(X\beta_i)} \right]$$

Where β_{ik} is the k-th element of β_i. Notice that even the direction of the effect is not entirely revealed by β_{ik}. You may find other ways to interpret the coefficients.

Conditional logit model

McFadden (1974) showed that a model closely related to the multinomial logit model can be obtained from an underlying utility comparison. It is called the Conditional logit model. Those models have similar response probabilities, but they differ in some key regards. In the MNL model, the conditioning variables do not change across alternative: for each i, x_i contains variables specific to the individuals but not to the alternatives. On the other hand, the conditional logit model cannot have variables varying over i but not over j. However, using an appropriate transformation you can obtain the MNL model using the conditional technique, as it turns out to actually contain the MNL model as a special case.

An example: a simple model of occupational choice

Utilizaremos la base de datos "status.dta". Enter "describe" to see the data definitions. It has been extracted from Wooldridge "Econometric Analysis of Cross Section and Panel Data", example 15.4 (page 498). It is a subset from Keane and Wolpin (1997) that contains employment and schooling history for a sample of men for 1987. The three possible outcomes are enrolled in school ("status=0"), not in school and not working ("status=1"), and working ("status=2"). As explanatory variables we have education, past work experience, and a black binary indicator.

Open the database:

```
. use status, clear
```

```
. describe
```

Now we can enter the command "mlogit":

```
. mlogit status educ black exper expersq, b(0)
```

With "b(0)" we indicate that the base category is "enrolled in school" ("status=0").

Marginal Effects

We can calculate the marginal effects "by hand" using the formula derived above, or we can simply take advantage of the command "mfx":

```
. mfx compute, predict(outcome(0))
```

```
. mfx compute, predict(outcome(1))
```

```
. mfx compute, predict(outcome(2))
```

Where "outcome(·)" denotes the response under consideration. For example, an addition year of education (at the mean x) changes the probability of going to school by +0.01, the probability of staying home by -0.05, and the probability of working by +0.03. This is completely logical: in general people invest in education in order to get further education (e.g. going to college in order to get a Ph.D. in the future) or they invest in education in order to enter the job market. Thus, investing in education reduces the probability of staying home.

We can also obtain predicted probabilities to provide some useful comparisons. For instance, consider two non-black men, each with 3 years of experience (and then "expersq=9"). Calculate the three predicted probabilities for a man with 12 years of education:

```
. mfx, predict(p outcome(0)) at(12 0 3 9)
```

```
. mfx, predict(p outcome(1)) at(12 0 3 9)
```

```
. mfx, predict(p outcome(2)) at(12 0 3 9)
```

And calculate the predicted probabilities for a man with 16 years of education:

```
. mfx, predict(p outcome(0)) at(16 0 3 9)
```

```
. mfx, predict(p outcome(1)) at(16 0 3 9)
```

```
. mfx, predict(p outcome(2)) at(16 0 3 9)
```

You can see that the 4 years of additional schooling changed the probability of going to school by +0.06, the probability of staying home by -0.1, and the probability of working by +0.04.

Tests for the multinomial logit model

You may install the command "mlogtest", which computes a variety of tests for multinomial logit models. You may select the test you want by specifying the appropriate option. For each independent variable, "mlogtest" can perform a likelihood-ratio or Wald test of the null hypothesis that the coefficients of the variable equal zero across all equations. It can also perform Wald or LR tests of whether any pair of outcome categories can be combined. In addition, "mlogtest" computes both Hausman and Small-Hsiao tests of independence of irrelevance alternatives (IIA) assumption.

For instance, let's perform the Wald tests for combining outcome categories. The null hypothesis is that the difference between all coefficients except intercepts associated with given pair of outcomes are 0 (i.e. the categories can be collapsed).

```
. mlogtest, combine
```

Exercise 13.1: Predict the probabilities of each response and observation: $\hat{p}_{ij}$. Then generate an "educated guess" based on those predicted probabilities:

$$\tilde{p}_{ij} = \max_{j} \left\{ \hat{p}_{ij} \right\}$$

Then enter "tab2 $\hat{p}_{ij}$ $\tilde{p}_{ij}$" and calculate the proportion of "right" and "wrong" predictions from the model.

Nested logit

The conditional logit model has as assumption the independence from irrelevant alternatives (IIA). This means that relative probabilities for any two alternatives depend only on the attributes of those two alternatives. Thus, it implies that adding another alternative or changing the characteristics of a third alternative does not affect the relative odds between two alternatives. This implication is implausible for applications with similar alternatives. For more details, see Wooldridge "Econometric Analysis of Cross Section and Panel Data", page 501.

A different approach to relaxing the IIA assumption is the specification of a "hierarchical model". The most popular of these is the nested logit model. Suppose that the total number of alternatives can be put into S groups of similar alternatives G_s. Thus the first hierarchy corresponds to which of the S groups y falls into, and the second corresponds to the actual alternative within each group. We can propose separate models for those probabilities:

$$P(y \in G_s \mid x) \text{ and } P(y = j \mid y \in G_s, x)$$

The response probability $P(y = j \mid x)$, which is ultimately of interest, is obtained by multiplying both equations. The problem can be solved either in a two-step fashion or using the full maximum likelihood.

For instance, consider the model of commuting to work. First you might want to decide whether to travel by car, to travel "naturally", or by public transportation. Once you decided to travel by car, you have to decide whether to travel alone or carpooling. Once you decided to travel "naturally", you have to decide whether to travel by foot or in bicycle. Once you decided to travel by public transportation, you have to decide whether to travel by bus or by train.

Ordered Logit (Probit)

Another kind of multinomial response is an ordered response. As the name suggests, if y is an ordered response, then the values we assign to each outcome are no longer arbitrary. For example, y might be a credit rating on a scale from zero to six, with y = 6 representing the highest rating and y = 0 the lowest rating.

The fact that 6 is a better rating than 5 conveys useful information, even though the credit rating itself only has ordinal meaning. For example, we cannot say that the difference between 4 and 2 is twice as important as the difference between 1 and 0.

Consider for instance the ordered probit model. Let y be an ordered response taking on the values {0, 1,…, J} for some known integer J. And define the latent variable as:

$$y^* = x\beta + e, \quad e \mid x \sim N(0,1)$$

Where x does not contain an intercept. Let $\alpha_1 < \alpha_2 < \ldots < \alpha_J$ be unknown cut points (a.k.a. threshold parameters), and define:

$$y = \begin{cases} 0 \quad if \;\; y^* \leq \alpha_1 \\ 1 \quad if \;\; \alpha_1 < y^* \leq \alpha_2 \\ \quad\quad \vdots \\ J \quad if \;\; y^* > \alpha_J \end{cases}$$

Finally, compute each response probability:

$$P(y=0\,|\,x) = P(y^* \leq \alpha_1\,|\,x) = P(x\beta + e \leq \alpha_1\,|\,x) = \Phi(\alpha_1 - x\beta)$$

$$P(y=1\,|\,x) = P(\alpha_1 < y^* \leq \alpha_2\,|\,x) = \Phi(\alpha_2 - x\beta) - \Phi(\alpha_1 - x\beta)$$

$$\vdots$$

$$P(y=J-1\,|\,x) = P(\alpha_{J-1} < y^* \leq \alpha_J\,|\,x) = \Phi(\alpha_J - x\beta) - \Phi(\alpha_{J-1} - x\beta)$$

$$P(y=J\,|\,x) = P(y^* > \alpha_J\,|\,x) = 1 - \Phi(\alpha_J - x\beta)$$

The parameters α and β can be estimated by maximum likelihood. The magnitude of the ordered probit coefficient does not have a simple interpretation, but you can retrieve qualitative information directly from its sign and statistical significance. We can also compute marginal effects with respect to x_k.

As always, replacing the normal cumulative distribution by the logistic yields the ordered logit model.

Example: life satisfaction in Russia

Recall the Russian database used in the first two weeks ("russia.dta"). We will estimate a model explaining life satisfaction using health and economic variables. It will also include dummies for geographical areas (area*). The variable "satlif" measures life satisfaction, and takes values from 1 ("not at all satisfied") to 5 ("fully satisfied").

First open the database, generate the geographical dummies, and transform the household expenditures to thousand rubles:

```
. use russia, clear

. gen area1=(geo==1)

. gen area2=(geo==2)

. replace totexpr= totexpr/1000
```

Then estimate the ordered probit model:

```
. oprobit satlif monage obese smokes operat hattac totexpr econrk resprk powrnk work0 area1-area2, robust
```

We can interpret directly the sign and statistical significance of the coefficients. Only the slopes on expenditures, power ranking, respect ranking and economic ranking are statistically different from zero, and they are all positive. As expected, more power, respect, status and consumption raises life satisfaction.

We can use "mfx" to retrieve marginal effects at the mean x:

```
. mfx, predict(outcome(1))

. mfx, predict(outcome(2))

. mfx, predict(outcome(3))

. mfx, predict(outcome(4))

. mfx, predict(outcome(5))
```

An increase in household expenditures of ten thousand rubles would change the probability of being not at all satisfied by -0.04, the probability of being less than satisfied by -0.02, the probability of being both yes and no satisfied by 0.02, the probability of being rather satisfied by 0.03, and the probability of being fully satisfied by 0.01. Once again: more consumption implies a lower probability of being relatively dissatisfied and than a higher probability of being satisfied.

Exercise 7.2: Predict the probabilities of each response and observation: $\hat{p}_{ij}$. Then generate an "educated guess" based on those predicted probabilities:

$$\tilde{p}_{ij} = \max_{j} \left\{ \hat{p}_{ij} \right\}$$

Then enter "tab2 $\hat{p}_{ij}$ $\tilde{p}_{ij}$" and calculate the proportion of "right" and "wrong" predictions from the model.

Problem Set #13

Part 1 (Multinomial logit)

The problem set uses data on choice of heating system in California houses, extracted from Kenneth Train's "Discrete Choice Methods with Simulation". The file is "heat.dta", and it is ready to be used. Enter "describe" to see the data definitions. The observations consist of single-family houses in California that were newly built and had central air-conditioning. The choice is among heating systems. Enter "tab heat" to see the different alternatives.

1. Estimate a multinomial logit model. Can you infer anything directly from the coefficients obtained? Try different specifications. Collapse some options for one of those specifications. Choose one to go on. Do not worry if the statistical significance is bad.

2. Calculate marginal effects or the differences in probabilities (as in the notes). Interpret carefully the results.

3. Calculate an index of goodness-of-fit.

Part 2 (Ordered probit)

Use the Russian database "russia.dta". There is a variable "evalhl" that measures the self-reported health evaluation. We need to explain that variable with demographic and health regressors.

1. Why would you use the ordered probit/logit model instead of OLS? Estimate an ordered probit model.

2. Can you retrieve any information directly from the coefficients? Calculate the marginal effects or the differences in probability (as in the notes). Interpret the results carefully.

3. Calculate an index of goodness-of-fit.

Chapter 14
Quantile Regression

An Introduction

Notice that we can define the sample mean as the solution to the problem of minimizing a sum of squared residuals, we can define the median as the solution to the problem of minimizing a sum of absolute residuals (Koenker et al., 2001). Since the symmetry of the absolute value yields the median, minimizing a sum of asymmetrically weighted absolute residuals yield other quantiles. Then solving:

$$\min_{\mu}\sum \rho_{\tau}(y_i - \mu)$$

Yields the τ-th sample quantile as its solution, where the function ρ_{τ} is: $\rho_{\tau}(z) = 1[z < 0]\cdot z(1-\tau) + 1[z \geq 0]\cdot z\tau$. With that optimización we can estimate the quantiles of the unconditional mean, but it is easy to define conditional quantiles in an analogous fashion. Given a random sample $\{y_1, y_2, ..., y_n\}$, we solve:

$$\min_{\beta}\sum \rho_{\tau}(y_i - \mu(x_i, \beta))$$

If $\mu(x_i, \beta)$ is formulated as a linear function of β, then the problem can be efficiently solved by linear programming methods.

An example

Open the database:

```
. use russia, clear
```

Command “qreg” fits quantile regression models. Option “quantile(x)” estimate quantile “x”. The default is quantile(.5). Por ejemplo, vamos a correr una median regression of real household expenditures contra real household income y una constante:

```
. qreg totexpr tincm_r, quantile(.5)
```

Using other quantiles:

```
. qreg totexpr tincm_r, quantile(.25)
```

```
. qreg totexpr tincm_r, quantile(.75)
```

Command “bsqreg” estimates a quantile regression with bootstrap standard errors. Option “reps()” specifies the number of bootstrap replications to be used. The default is 20, which is arguably too small.

```
. bsqreg totexpr tincm_r, quantile(.5) reps(100)
```

Graphical comparisons

Solo para mejorar el gráfico (aunque genere un bias) vamos a dropear los incomes y expenditure mayores a 50,000 rubles:

```
. drop if totexpr>50000 | tincm_r>50000
```

Vamos a hacer el gráfico de regresión parcial para los quantiles .25, .5 y .75, y para la regresión OLS:

```
. qreg totexpr tincm_r, quantile(.25)

. predict fitted25

. qreg totexpr tincm_r, quantile(.5)

. predict fitted50

. qreg totexpr tincm_r, quantile(.75)

. predict fitted75

. reg totexpr tincm_r

. predict fittedols

. graph twoway scatter totexpr tincm_r || line fitted25 fitted50 fitted75 fittedols tincm_r, legend(label(1 "Expenditures") label(2 "Quantile .25") label(3 "Quantile .5") label(4 "Quantile .75") label(5 "OLS"))
```

Problem Set #14

For this Problem Set we will use the dataset "incomes.dta", which is a quasi-generated sample of individuals, based on a British household dataset. Enter "describe" to see the data definitions. We want to estimate Mincer equations to study the returns to education.

The extract used here consists of individuals between 18 and 50 years old who are employed. The additional explanatory variables are: a constant, years of experience, experience squared, and a dummy variable for nonwhites.

1. Show descriptive statistics for each year (mean, standard deviation, and quantiles 0.1, 0.25, 0.5, 0.75 and 0.9).

2. Run OLS and quantile regressions (for quantiles 0.1, 0.25, 0.5, 0.75 and 0.9) for each of the sample years, 1996 to 2005. Build a table with the coefficients for years of education. Represent graphically the evolution of those coefficients.

3. Present similarly the marginal effects for years of experience.

4. Graph the relation (regressing by least squares and quantile regression) between years of education and hourly income (once again, for every year).

5. Find out whether the following affirmations are true:

a. The mean return to education and the returns at each quantile changed in a similar pattern.

b. The returns to education at these quantiles differ significantly. In general, the returns are higher at the higher quantiles.

c. During the 2000's, there were sharp increases in the returns to education, with greater increases at the higher quantiles.

Chapter 15
Robust inference

Bootstrapped standard errors

While a consistent estimator may be easy to obtain, the formula for the variance-covariance matrix is sometimes more difficult, or even mathematically intractable. As a consequence, some applications need a lot of assumptions to construct an estimate for that matrix.

Bootstrapping is a nonparametric approach for evaluating the distribution of a statistic based on random resampling. It is very useful because it only assumes that the sample is representative of the population. To retrieve such estimates, you have to follow three steps: 1. Draw random samples with replacement repeatedly from the sample dataset; 2. Estimate the desired statistic corresponding to these bootstrap samples; 3. Calculate the sample standard deviation of the sampling distribution.

Stata regression commands usually have bootstrapped standard errors as an option. For instance, in the above model:

```
. xtreg test treatment female educ income, fe i(school) vce(boot)
```

Please notice that the bootstrapped estimates vary among every estimations. For instance, run the above code again:

```
. xtreg test treatment female educ income, fe i(school) vce(boot)
```

The bootstrapped option has some useful options. The "reps()" option specifies the number of bootstrap replications to be performed. The default is 50, and 50-200 replications are generally adequate for estimates of standard errors. The "size()" option specifies the size of the samples to be drawn. The default is the same size as the data. For further information see for instance Andrews et al. (2000), who proposed a three step method of choosing the number of bootstrap repetitions for bootstrap standard errors, confidence intervals, confidence regions, hypothesis tests and p-values.

If the option "strata()" is specified, bootstrap samples are taken independently within each stratum. Finally, the "cluster()" option specifies the variables that identify resampling clusters. This is the cluster-robust version for bootstrap. If this option is specified, the sample drawn during each replication is a bootstrap sample of clusters. This resampling method is called a “pairs cluster bootstrap”. In this case the default size is the number of clusters in the original dataset. For instance:

```
. xtreg test treatment female educ income, fe i(school) vce(boot) cluster(school)
```

In Guan (2003) you can find a Montecarlo simulation that compares the power of conventional and bootstrapped versions of the clustered estimates. Guan (2003) also compares standard ML estimates versus bootstrapped in a Nonlinear least squares regression, and it gives an example on 2SLS.

Jackknife estimator

The jackknife drops in turn each observation (in the clustered version, each cluster), computes the leave-one-out estimate and then calculates the variance in that sample of coefficients. The option is similar to the bootstrap:

```
. xtreg test treatment female educ income, fe i(school) vce(jack)
```

However, (by construction) only the option "cluster" is available:

```
. xtreg test treatment female educ income, fe i(school) vce(jack) cluster(school)
```

The description of other nonparametric techniques, such half-sampling, subsampling, balanced repeated replications and the delta method, can be found in Efron (1981).

Exercise 15.1. Obtain the bootstrapped and jackknife standard errors "by hand" (i.e. performing the corresponding steps detailed above).

T-test bootstrapping

Following Cameron et al. (2006), the bootstrap-t procedure, proposed by Efron (1981) for confidence intervals, computes the following Wald statistic for each bootstrap replication:

$$w_b^* = \left(\hat{\beta}_{1,b}^* - \hat{\beta}_1\right) \Big/ s_{\hat{\beta}_{1,b}^*}, \quad b = 1,\ldots,B$$

Where b denotes each resampling, $s_{\hat{\beta}_{1,b}^*}$ is a cluster-robust standard error for $\hat{\beta}_{1,b}^*$, and $\hat{\beta}_1$ is the original OLS coefficient. As we are interested in cluster-robust estimates, the resampling is done at the cluster-level. Note that w_b^* is centered on $\hat{\beta}_1$. Then obtain the original Wald estimate:

$$w = \left(\hat{\beta}_1 - \beta_0\right) \Big/ s_{\hat{\beta}_1}$$

Obtaining $s_{\hat{\beta}_1}$ from the CRVE standard errors, which controls for both error heteroskedasticity across clusters and quite general correlation and heteroskedasticity within cluster (obtained using jointly the “robust” and “cluster()” options).

Finally, we test HO: $\beta_1 = 0$ against Ha: $\beta_1 \neq 0$. We reject H0 at level α if $w < w_{\alpha/2}^*$ or $w > w_{1-\alpha/2}^*$, where w_q^* denotes the qth quantile of $w_1^*,\ldots,w_B^*$.

Obtain w and $\hat{\beta}_1$:

```
. use education, clear
. xtreg test treatment female educ income, fe i(school) cluster(school)
. quietly: ereturn list
. matrix V=e(V)
. matrix b=e(b)
. scalar se=sqrt(V[1,1])
. scalar b1=b[1,1]
. scalar wfirst=b1/se
```

Then obtain w_b^* (say) 100 times:

```
. forvalues j=1(1)100 {
. quietly: preserve
. quietly: bsample , cluster(school)
. quietly: xtreg test treatment female educ income, fe i(school) cluster(school)
. quietly: ereturn list
. quietly: matrix V=e(V)
. quietly: matrix b=e(b)
. quietly: scalar se=sqrt(V[1,1])
. quietly: scalar beta=b[1,1]
. quietly: scalar w`j'=(beta-b1)/se
. quietly: restore
. }
```

Mostremos el w:

```
. display wfirst
```

Save the 100 variables w_b^* in 100 observations of a new variable, “wstar”:

```
. gen wstar=.
. forvalues t=1(1)100 {
. quietly: replace wstar= w`t' if _n==`t'
.  }
```

We need to see if $w < w^*_{\alpha/2}$ or $w > w^*_{1-\alpha/2}$. Then show the quantiles α /2% and $(1-\alpha)$ /2%, and compare it to w (variable "wfirst"):

```
. tabstat wstar, stat(p1 p5 p95 p99)
```

Two-Stage Models

Thorough this notes we will follow Hardin (2002)[3] and Greene (2000, Chapter 17.7). Just start noticing that some econometric models involve a two-step estimation procedure:

First stage: $E[y_1 \mid x_1, \theta_1]$

Second stage: $E[y_2 \mid x_2, \theta_2, E(y_1 \mid x_1, \theta_1)]$

There are two standard approaches to such estimation. The first is a full information maximum likelihood model (FIML), which consists in specifying the joint distribution $\mathrm{f}(\mathrm{y}_1, \mathrm{y}_2 \mid \mathrm{x}_1, \mathrm{x}_2, \theta_1, \theta_2)$ and maximizing its corresponding log-likelihood function. The second approach is denominated the limited information maximum likelihood (LIML) two-step procedure. Since the first model only involves estimating the first parameter vector (θ_1), we can estimate it separately. Then we can estimate the second parameter vector conditional on the results of the first step estimation. This is performed by maximizing the following conditional log-likelihood:

$$L = \sum_{\mathrm{i}=1}^{\mathrm{n}} \ln \mathrm{f}\left(\mathrm{y}_{2\mathrm{i}} \mid \mathrm{x}_{2\mathrm{i}}, \theta_{2\mathrm{i}}, (\mathrm{x}_{1\mathrm{i}}, \hat{\theta}_{1\mathrm{i}}) \right)$$

The Murphy-Topel estimator

For the second approach, Greene (2000) briefly summarizes the results from Murphy and Topel (1985). Assume that θ_1 is a q×1 vector of unknown parameters associated with an n×q matrix of covariates X. Additionally, θ_2 is a p×1 vector of unknown parameters associated with an n × p matrix of covariates W. The Murphy-Topel formula for the variance estimate of θ_2 is then given by:

$$V_2^{MT} = V_2 + V_2\left(CV_1C' - RV_1C' - CV_1R'\right)V_2$$

Where:

V_1 (a $q \times q$ matrix) is the Asymptotic variance matrix of $\hat{\theta}_1$ based on $L_1(\theta_1)$

V_2 (a $p \times p$ matrix) is the Asymptotic variance matrix of $\hat{\theta}_2$ based on $L_2(\theta_2 \mid \theta_1)$

[3] He also presents a sandwich estimate of variance as an alternative to the Murphy-Topel estimate.

	First Stage				Second Stage		
dep var: z	**coef**	**std error**	**p-value**	dep var: y	**coef**	**std error**	**p-value**
age	-0.073	0.035	0.035	age	0.073	0.048	0.125
income	0.219	0.232	0.345	income	0.045	0.178	0.800
ownrent	0.189	0.628	0.763	expend	-0.007	0.003	0.022
selfemp	-1.944	1.088	0.074	zhat	4.632	3.968	0.243
cons	2.724	1.060	0.010	cons	-6.320	3.718	0.089

C (a $p \times q$ matrix) equals $E\left[\frac{\partial L_2}{\partial \theta_2}\frac{\partial L_2}{\partial \theta_1'}\right]$

R (a $p \times q$ matrix) equals $E\left[\frac{\partial L_2}{\partial \theta_2}\frac{\partial L_1}{\partial \theta_1'}\right]$

Matrices V_1 and V_2 can be estimated in many ways. For instance, in the following example we will use the robust versions.

An example

We will consider an example from Hardin (2002), using data from Greene (1992). The interesting dependent variable is the number of major derogatory reports recorded in the credit history for the sample of applicants to a credit card. There are 1319 observations in the sample for this variable (10% of the original dataset), 1060 of which are zeros. A Poisson regression model will be then a reasonable choice for the second stage.

In a first stage we will estimate of a model of credit card application's outcome, using a logit model. Then we will include the predicted values of the latter model as explanatory variable in the Poisson regression. Notice that using the full information approach would involve writing a very complex joint distribution. Furthermore, maximizing such likelihood might yield inconsistent estimates. Then, we will pursue the two-step approach.

From the original study, the author makes only 100 observations available for our use. Denote "z" the variable indicating if the application for credit card was accepted, and "y" the number of major derogatory reports. Open the database:

```
. use credit, clear
```

We can fit the models:

```
. logit z age income ownrent selfemp, robust
```

```
. predict zhat
```

```
. poisson y age income expend zhat, robust
. drop zhat
```

This output shows the "naive" (robust) standard errors (i.e. which assume that there is no error in the generation of $\hat{z}_i$ in the first-stage).

Within Stata it is not too difficult to obtain the Murphy–Topel variance estimates for a two-stage model. We need to gather all the information for calculating the matrix V_2^{MT}. First we have to obtain V_1 and V_2 (V1r and V2r in the following code). We will also obtain $\hat{z}_i$, $\hat{y}_i$ and the coefficient on $\hat{z}_i$ in the second stage (zhat, yhat and zz in the code), as we will need them later:

```
. logit z age income ownrent selfemp, robust

. matrix V1r = .99 * e(V)

. predict zhat

. poisson y age income expend zhat, robust

. matrix V2r = .99 * e(V)

. predict yhat

. scalar zz = _b[zhat]
```

You may notice that we have multiplied the robust variance-covariance matrices by 0.99. This is because Stata applies a small sample adjustment $n/(n-1)$ to the robust variance estimates, and then we can undo that adjustment by multiplying accordingly.

Then we have to calculate R and C. First we have to do a little bit of algebra. A logistic model where z is the outcome (whether the application is successful) and X is the matrix of covariates has log likelihood given by:

$$L_1 = \sum_{\mathrm{i=1}}^{\mathrm{n}} \left(z_i x_i \theta_1 - \ln\{1 + \exp(x_i \theta_1)\} \right)$$

A Poisson model where y is the outcome (number of derogatory reports) and W is the matrix of covariates has log-likelihood given by:

$$L_2 = \sum_{\mathrm{i=1}}^{\mathrm{n}} \left(y_i w_i \theta_2 - \exp(w_i \theta_2) - \ln \Gamma(y_i + 1) \right)$$

Differentiating we obtain:

$$\frac{\partial L_1}{\partial \theta_1} = \sum_{i=1}^{n} x_i \left(z_i - \hat{z}_i \right) x_i' = X' diag\left(z_i - \hat{z}_i \right) X$$

$$\frac{\partial L_2}{\partial \theta_1} = \sum_{i=1}^{n} x_i \left(y_i - \hat{y}_i \right) \hat{z}_i \left(1 - \hat{z}_i \right) \hat{\theta}_2 x_i' = X' diag\left(\left(y_i - \hat{y}_i \right) \hat{z}_i \left(1 - \hat{z}_i \right) \right) X$$

$$\frac{\partial L_2}{\partial \theta_2} = \sum_{i=1}^{n} w_i \left(y_i - \hat{y}_i\right)\hat{w}_i = W' diag\left(y_i - \hat{y}_i\right)W$$

Where $\hat{\theta}_2$ is the estimate obtained from the two-step procedure. The "matrix accum" command calculates X'X. Let's obtain C and R:

```
. gen byte cons = 1
```

```
. matrix accum C = age income ownrent selfemp cons age income expend zhat cons [iw=(y-yhat)*(y-yhat)*zhat*(1-zhat)*zz], nocons
```

```
. matrix accum R = age income ownrent selfemp cons age income expend zhat cons [iw=(y-yhat)*(z-zhat)], nocons
```

And keep only the desired partition:

```
. matrix list C
```

```
. matrix C = C[6..10,1..5]
```

```
. matrix list R
```

```
. matrix R = R[6..10,1..5]
```

Finally, we can calculate the Murphy–Topel estimate:

```
. matrix M = V2r + (V2r * (C*V1r*C' - R*V1r*C' - C*V1r*R') * V2r)
```

```
. matrix list M
```

To show the Murphy–Topel estimates, Hardin (2002) suggested a rather useful way. In case it already exists, drop the program "doit":

```
. capture program drop doit
```

Then define it:

```
. program define doit, eclass
```

```
. matrix b = e(b) // stores the coefficients in b
```

```
. ereturn post b M // saves the old coefficients along with the new standard errors
```

	Second Stage		
dep var: y	**coef**	**std error**	**p-value**
age	0.073	0.314	0.816
income	0.045	1.483	0.976
expend	-0.007	0.019	0.723
zhat	4.632	30.591	0.880
cons	-6.320	24.739	0.798

```
. ereturn local vcetype "Mtopel" // names the new standard errors

. ereturn display // shows the regression table

. end
```

Thus, after running a regression and calculating "M", just run "doit":

```
. doit
```

Hole (2006) shows how the calculation of the Murphy-Topel variance estimator for two-step models can be simplified in Stata by using the scores option of predict.

Problem Set #15

For this Problem Set we will use the dataset "inactivity.dta". We want to estimate a model of days of labor inactivity, where the interesting explanatory variable will be the ex-ante probability of being ill (notice that there would not be problems if the interesting variable were if the individual is sick ex-post).

1. Is "actdays" a count-variable?

2. Run the Poisson Regression Model and a Negative Binomial Regression Model using "sex", "age", "agesq", "income", "levyplus", "freepoor" and "freerepa" as explanatory variables. Test if the Poisson model is suitable.

3. Estimate as a first stage a Probit model of the probability of being ill in the last two weeks. Use as explanatory variables "sex", "age", "agesq", "income", "hscore", "chcond1" and "chcond2". Then estimate the second stage: a Negative Binomial Regression of the number of inactivity days using the explanatory variables utilized in the former exercise plus the predicted value of the Probit model (i.e. the ex-ante probability of being ill). Present the coefficients and the "naive" standard errors.

4. Compute and show the Murphy-Topel estimates. Compare them with the "naive" versions.

Chapter 16

Program Evaluation and Matching

Introduction to Program Evaluation

The underlying model is one with two outcome variables: y_{i1} is the outcome for individual i with treatment, and y_{i0} is the outcome for individual *i* without treatment. Denote w_i the binary variable that indicates whether individual i is benefited from the treatment. The original sin in statistical methods applied to social sciences is that we can only observe either y_{i1} or y_{i0}: at a precise moment of time a person is either employed or unemployed, sick or healthy, benefiting from a program or not, etc.

Esto puede verse muy claro tratando de estimar el el average treatment effect on the treated (ATT): $\mathrm{E}[y_{i1} \mid w_i = 1] - \mathrm{E}[y_{i0} \mid w_i = 1]$. Vemos que si bien podemos estimar el primer termino, no tenemos observaciones para el segundo: i.e. no podemos observar cual seria el outcome de un individuo que tratamos si no hubiera recibido el tratamiento.

On the contrary, phisisics can reproduce “all” the other characteristics in an experiment and change only the property that they are interested in. If we tried to do the same with human subjects it would render useless: gravity is the same now or in twenty days, but the world and the human subject in consideration can change dramatically. There is an overwhelmingly interesting epistemologic discussion behind this simple introduction, and the reader will probably find useful to read Holland (1986) and Heckman (2000).

The econometric trick we have analyzed so far is to approximate the result by looking at cross section, time series and panel data. You cannot see a person sick and healthy at the same time, but you can observe people through time and you will find that sometimes they were sick and some times they were not. Or you can see a cross section of sick and healthy people, or the combination of the two dimensions.

Por ejemplo, si usamos como estimador naive la diferencia de medias entre individuos tratados y no tratados:

$$\underbrace{E\left[\mathrm{Y}_i|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i|\mathrm{D}_i = 0\right]}_{\text{Observed difference in average health}} = \underbrace{E\left[\mathrm{Y}_{1i}|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i}|\mathrm{D}_i = 1\right]}_{\text{average treatment effect on the treated}}$$
$$+\underbrace{E\left[\mathrm{Y}_{0i}|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i}|\mathrm{D}_i = 0\right]}_{\text{selection bias}}$$

El segundo termino, el *selection bias*, es nulo solo si los individuos no tratados pueden ser usados de contrafactuales de los tratados.

Como ejemplo, supongamos que la variable dependiente es el salario, y el tratamiento es ir al college. Entonces el selection bias es cero solo si el salario que hubieran percibido las personas que fueron al collegue si no hubieran ido fuera el mismo que las personas que no fueron al collegue. Esto es probablemente falso: según el modelo de Spence (¿??) las personas mas talentosas (y por lo tanto productivas) son las que van al collegue. Como el salario incrementa con la productividad: A<B, por lo que el selection bias es negativo.

Al menos en el ejemplo tenemos una idea de cual es el signo del selection bias. En general no tenemos idea ni del signo. Notese que el problema es el mismo que en una Mincer equation (regression of wages on education plus controls), donde la variable de interes no necesariamente tiene que ser binaria.

Hay una forma trivial de eliminar el selection bias: randomizing the treatment. You would not even need to use control variables in a regression, as you should simply compare means between the treated and untreated population (notice that this would not work if people were able to either voluntarily opt out from the treatment if chosen, or either be included in not chosen).

En lo que sigue vamos a seguir trabajando con the average treatment effect: $\mathrm{E}[\mathrm{y_i} \mid w_i = 1] - \mathrm{E}[\mathrm{y_i} \mid w_i = 0]$. El outcome para el individuo i se puede escribir como:

$$\mathrm{y_i} = (1 - \mathrm{w_i}) \cdot \mathrm{y_{i0}} + \mathrm{w_i} \cdot \mathrm{y_{i1}} = \mathrm{y_{i0}} + \mathrm{w_i}\ (\mathrm{y_{i1}} - \mathrm{y_{i0}})$$

If treatment is randomized across individuals, the treatment indicator $\mathrm{w_i}$ is independent of the outcomes:

$$\mathrm{ATE} = \mathrm{E}[\mathrm{y_i} \mid w_i = 1] - \mathrm{E}[\mathrm{y_i} \mid w_i = 0] = \mathrm{E}[\mathrm{y_{i1}} \mid w_i = 1] - \mathrm{E}[\mathrm{y_{i0}} \mid w_i = 0] = \mathrm{E}[\mathrm{y_{i1}}] - \mathrm{E}[\mathrm{y_{i0}}]$$

Appliying the analogous principle, a consistent estimate of ATE would be:

$$\mathrm{A\hat{T}E} = \sum_{i=1}^{N} \mathrm{y_i w_i} - \mathrm{y_i}\ (1 - \mathrm{w_i})$$

For inference you can use your favourite mean-comparison test. However, randomized experiments are very costly. Burtless (1995) summarizes 190 social experiments dating from 1991 with a total cost in current dollars of one billion. In 1995 the budget of the National Science Foundation for economics was only 25 million dollars, so this is indeed a big number.[4]

Notice that we do not need independence, but a weaker assumption of mean independence:

$$\mathrm{E}[\mathrm{y_{i1}} \mid w_i] = \mathrm{E}[\mathrm{y_{i1}}]$$

[4] However, we should consider only the intrinsic costs of the experiment. If (say) the program involves the assignment of school vouchers, the expenses on vouchers should be considered as “schooling” or “poverty alleviation programs”, and not “scientific expenses”.

$$\mathrm{E}[\mathrm{y}_{\mathrm{i0}} \mid w_i] = \mathrm{E}[\mathrm{y}_{\mathrm{i0}}]$$

Si bien en la realidad no tenemos random assignment, the assumption of Ignorability of Treatment (Rosenbaum and Rubin, 1983) propuso que quizas podemos cumplir ese assumption si utilizamos los control variables adecuados: conditional on the variables x, the outcome variables $(\mathrm{y}_{\mathrm{i1}}, \mathrm{y}_{\mathrm{i0}})$ and the treatment variable w_i are independent.

In what follows we require the weaker assumption of mean independence conditional on x:

$$\mathrm{E}[\mathrm{y}_{\mathrm{i}} \mid w_i = 1, x] - \mathrm{E}[\mathrm{y}_{\mathrm{i}} \mid w_i = 0, x] = \mathrm{E}[\mathrm{y}_{\mathrm{i1}} \mid w_i = 1, x] - \mathrm{E}[\mathrm{y}_{\mathrm{i0}} \mid w_i = 0, x] =$$

$$= \mathrm{E}[\mathrm{y}_{\mathrm{i1}} \mid x] - \mathrm{E}[\mathrm{y}_{\mathrm{i0}} \mid x] = ATE(x)$$

We could simply use the sample analogs of the conditional expectations above, and then obtain ATE using a weighted average over x using its sample frequency distribution. But there is an even simpler way that involves a regression. The advantage of this approach is that we already have a lot of theory and estimators about regression methods, and writing the treatment effects problem as a regression allows us to extend all those existing tools (instead of having to develop new ones for problems already studied).

Write y_{i1} and y_{i0} as deviations around a mean:

$$\mathrm{y}_{\mathrm{i1}} = \mu_1 + \mathrm{v}_{\mathrm{i1}}$$

$$\mathrm{y}_{\mathrm{i0}} = \mu_0 + \mathrm{v}_{\mathrm{i0}}$$

Where both v_{i1} and v_{i0} have zero expectation. We can express the outcome as:

$$\mathrm{y}_{\mathrm{i}} = \mu_0 + w_i\,(\mu_1 - \mu_0) + \mathrm{v}_{\mathrm{i0}} + w_i\,(\mathrm{v}_{\mathrm{i1}} - \mathrm{v}_{\mathrm{i0}})$$

Take conditional expectation:

$$\mathrm{E}[\mathrm{y}_{\mathrm{i}} \mid w_i, x] = \mu_0 + w_i\,(\mu_1 - \mu_0) + \mathrm{E}[\mathrm{v}_{\mathrm{i0}} \mid w_i, x] + w_i\,(\mathrm{E}[\mathrm{v}_{\mathrm{i1}} \mid w_i, x] - \mathrm{E}[\mathrm{v}_{\mathrm{i0}} \mid w_i, x])$$

Where we used the Ignorability of Treatment assumption. We will assume that the conditional expectation of v_{i0} and v_{i1} are linear in x. Then:

$$\mathrm{E}[\mathrm{y}_{\mathrm{i}} \mid w_i, x] = \delta + w_i\,(\mu_1 - \mu_0) + x\beta + w_i\,(x - E[x])\theta$$

Where we re-arranged the above expression to demean the last term. It does not alter the results, it simply allows us to calculate ATE directly since the unconditional expectation of $x - E[x]$ is zero. The expression multiplying w_i in the second term is exactly the ATE. Therefore, we can obtain an estimator of ATE by running the above regression:

$$\mathrm{y}_{\mathrm{i}} = \beta_0 + w_i\,\beta_1 + x\beta_2 + w_i\,(x - \bar{x})\beta_3 + \upsilon_i$$

And $\hat{\beta}_1$ is our estimator for the average treatment effect. As we mentioned before, you can extend this idea to nonlinear models, you can use robust estimators for the standard errors, etc. And with the same regression we can also report: $\mathrm{A\hat{T}E(x)} = \hat{\beta}_1 + (x - \bar{x})\hat{\beta}_3$.

En lo que sigue vamos a introducir el metodo de matching.

Matching

Dejemos la weaker assumption de mean independence y volvamos a la condicion de Ignorability of Treatment original de Rubin: una vez que condicionamos por x, $(\mathrm{y_{i1}, y_{i0}})$ es independiente de w_i.

La idea detrás de matching es muy rudimentaria, y deberia resultar sumamente familiar a todos. Por ejemplo, este razonamiento en la vida cotidiana: "A Juan y Pedro los percibo igualmente talentosos, fueron al mismo colegio, la misma universidad y hasta se visten igual. Juan decidio irse a hacer un MBA a USA y Pedro no. Juan ahora gana mucho mas que Pedro. El MBA debe explicar gran parte de eso".

Una de las formas mas comunes de interpretar informacion es "matchear" un individuo con su mejor aproximacion a un "contrafactual". Consideren como ejemplo de tratamiento ir al college. Sin random assigment seria ingenuo comparar el promedio de salarios de las personas que fueron y las que no fueron a la universidad. Pero si yo encuentro a muchos pares de Juanes y Pedros, el promedio de los salarios de los Juanes menos el de los Pedros suena un estimador muy razonable. De hecho, en las ciencias sociales hay muchisimos experimentos involucrando gemelos homocigotas y heterocigotas separados al nacer, que consisten en unos pares mas que interesantes.

Pero elegir los pares no puede hacerse arbitrariamente, y los gemelos homocigotas no agotan el numero de problemas a estudiar en la economia. La estrategia entonces seria: considera al grupo de alumnos del ultimo anio del high school en la ciudad de Newark que cumplan una larga serie de condiciones como ser americanos, blancos, no tener antecedentes penales, estar en el mismo tier de tests estandarizados, vivir con sus padres, que los padres esten todavia en pareja, tener entre 1 y 2 hermanos, etc. Este ejercicio se parece mucho a matchear a Juanes y a Pedros: solo comparo entre individuos con muchisimas caracteristicas en comun, tal que la variable college es "casi" ortogonal.

Graficamente esto implicaria dividir cada variable en x en regiones discretas (0 hermanos, 1 hermano, etc.) y luego generar una gran tabla K-dimensional (el numero de variables en xi, menos la constante) con todas las posibles combinaciones de los valores discretos en x. El $\mathrm{A\hat{T}E(x)}$ entonces seria la diferencia de medias entre tratados y no tratados en la celda correspondiente a x. Para obtener el ATE simplemente tenemos que promediar los $\mathrm{A\hat{T}E(x)}$ para todas las celdas, ponderandolos por la frecuency distribution correspondiente a cada celda.

Matching es simplemente un modelo de selection on observables. La ventaja de usar matching en lugar de el regression method expuesto mas arriba es poder capturar non-linearities a la hora de condicionar en x. No obstante, podriamos utilizar por ejemplo interacciones entre variables y funciones de variables como controles, ya que el modelo de regresion esta limitado a ser lineal en los parametros y no en las variables.

No obstante, el matching puede ser dificil de implementar en la practica, porque la dimensionalidad del problem crece con el numero de variables control. Y aparecen otros problemas: por ejemplo, gran parte de las celdas pueden estar “vacias”. Para lidiar con este tipo de problemas se generaron otros metodos (parametricos y no-parametricos) de “matchear” individuos.

The propensity score theorem (Rosenbaum and Rubin, 1983) states that if potential outcomes are independent of treatment status conditional on a multivariate covariate vector, xi, then potential outcomes are independent of treatment status conditional on a scalar function of covariates, the propensity score: P(xi)=E[wi | xi] (la probabilidad de recibir treatment dado xi).

The proof is straightforward. Assume that the Ignorability of Treatment holds: yji is orthogonal to wi, conditional on xi. We must show that this implies that P[wi = 1 | yji, p(xi)] does not depend on yji: i.e. yji is orthogonal to wi, conditional on p(xi).

<u>To be completed.</u>

Es decir, no se necesita necesariamente matchear uno a uno las celdas definidas por x, sino que basta con condicionar en el propensity score. Given that the propensity score is a scalar, the above condition facilitates matching because the dimension is reduced from *K* (dimension of x) to one.

Vamos a obtener un estimador utilizando esta idea. Begin working the following expression:

$$\begin{aligned}[\mathrm{w_i - p(x_i)}]y_\mathrm{i} &= [\mathrm{w_i - p(x_i)}][(1-\mathrm{w_i})y_\mathrm{i0} + \mathrm{w_i}y_\mathrm{i1}] \\ &= \mathrm{w_i}y_\mathrm{i1} - \mathrm{p(x_i)}(1-\mathrm{w_i})y_\mathrm{i0} - \mathrm{p(x_i)w_i}y_\mathrm{i1}\end{aligned}$$

Using the fact that $\mathrm{w_i^2} = w_i$ and $\mathrm{w_i}(1-\mathrm{w_i}) = 0$. Take conditional expectation:

$$E[\mathrm{w_i}y_\mathrm{i1} - \mathrm{p(x_i)}(1-\mathrm{w_i})y_\mathrm{i0} - \mathrm{p(x_i)w_i}y_\mathrm{i1} \mid w_i, x_i] =$$

$$= \mathrm{w_i} E[y_{\mathrm{i1}} \mid w_i, x_i] - (1 - \mathrm{w_i})\mathrm{p(x_i)} E[y_{\mathrm{i0}} \mid w_i, x_i] - \mathrm{p(x_i)w_i} E[y_{\mathrm{i1}} \mid w_i, x_i]$$

In the last step we used the Ignorability of Treatment condition. Now take expectation conditional on x_i:

$$E[(\mathrm{w_i} - \mathrm{p(x_i)})y_\mathrm{i} \mid x_i] = \mathrm{p(x_i)} E[y_{\mathrm{i1}} \mid x_i] - (1 - \mathrm{p(x_i)})\mathrm{p(x_i)} E[y_{\mathrm{i0}} \mid x_i] - \mathrm{p(x_i)}^2 E[y_{\mathrm{i1}} \mid x_i]$$

$$= \mathrm{p(x_i)}[1 - \mathrm{p(x_i)}][E[y_{\mathrm{i1}} \mid x_i] - E[y_{\mathrm{i0}} \mid x_i]] = \mathrm{p(x_i)}[1 - \mathrm{p(x_i)}]ATE(x_i)$$

Solve for $ATE(x_i)$:

$$ATE(x_i) = E\left[\frac{[\mathrm{w_i} - \mathrm{p(x_i)}]y_\mathrm{i}}{\mathrm{p(x_i)}[1 - \mathrm{p(x_i)}]} \mid x_i \right]$$

We can use the analogous principle to obtain an estimator:

$$A\hat{T}E = \frac{1}{N} \sum_{i=1}^{N} \frac{[\mathrm{w_i} - \hat{\mathrm{p}}(\mathrm{x_i})]y_\mathrm{i}}{\hat{\mathrm{p}}(\mathrm{x_i})[1 - \hat{\mathrm{p}}(\mathrm{x_i})]}$$

Where $\hat{\mathrm{p}}(\mathrm{x_i})$ is obtained in a first stage, fitting por ejemplo un probit de $\mathrm{w_i}$ on $\mathrm{x_i}$ and luego obtaining the predicted value $\hat{\mathrm{p}}(\mathrm{x_i}) = \Phi(\mathrm{x_i}\hat{\beta})$.

Exercise 16.1: Calculate this formula "by hand" using the data in the following example. Compare the results to the output from standard Stata commands.

Pero veremos que una vez que tenemos el propensity score esta no es la unica forma de "matchear" individuos con similar propensity score, y utilizaremos metodos no-parametricos mas ricos.

Stata Commands

Stata does not have a built-in command for the propensity score matching. However, there are several user-written modules for this method. The following modules are among the most popular: "psmatch2.ado" (developed by Leuven and Sianesi, 2003), "pscore.ado" (Becker and Ichino, 2002) and "nnmatch.ado" (Abadie, Drukker, Herr, and Imbens, 2004).

You can find these modules using the .net command. For instance, for pscore:

```
. net search pscore
```

In this notes we will use the "pscore" command (for any particular reason).

An example

Vamos a escribir un modelo de en la cual los individuos son estudiantes del colegio secundario que reciben un plan de ayuda de útiles escolares, y queremos estudiar el impacto de esos planes sobre la performance de los alumnos en un examen estandarizado de matemática (variable "score"). El plan intenta ser randomizado, pero por distintos problemas de implementación la asignación no es random. Los datos son generados, de forma tal que sabemos que el efecto tratamiento sobre los tratados es exactamente 5. Como variables de control se tienen el sexo del estudiante, el income de sus padres, si en el hogar hubo episodios violentos el último año, si los padres están divorciados, y el tamaño del hogar. Estas variables son las únicas que determinan la selección para el treatment y además tienen efectos sobre los scores (i.e. no hay selection sobre los unobservables). Abramos la base de datos:

```
. use psm, clear
```

Podemos ver qué ocurriría si corremos una regresión lineal:

```
. reg evalhl plan_sel monage alclmo smokes diseases totexpr vacatn subord height marsta1, robust
```

Vemos que OLS hace una labor aceptable para medir el ATT.

Y ahora hacemos el regression method.

Ahora vamos a seguir con matching. El comando pscore hace los siguientes pasos:

1. Estimate the probit (or logit) model.

2. Split the sample in k equally spaced intervals of the propensity score (default k=5).

3. Within each interval test that the average propensity score of treated and control units do not differ.

4. If the test fails in one interval, split the interval in halves and test again.

5. Continue until, in all intervals, the average propensity score of treated and control units do not differ.

6. Within each interval, test that the means of each characteristic do not differ between treated and control units. This is a necessary condition for the Balancing Hypothesis.

7. If the means of one or more characteristics differ, inform the user that the balancing properties is not satisfied.

El comando se toma todo el trabajo de general “clusters” del propensity score basicamente porque no esta limitado a utilizar la formula de ATE que mostramos mas arriba, sino que permitira promediar los ATE(x) en formas no-parametricas incluso mas interesantes.

Antes de calcular el propensity store debemos garantizarnos que no incluyamos ninguna observación con missing values. Recordemos que si "varlist" es la lista de las variables que queremos utilizar, entonces deberíamos ingresar:

```
. foreach var in varlist {

. drop if `var'==.

. }
```

Corremos el comando "pscore":

```
. pscore treated female income hhsize divorced violence, pscore(myscore) blockid(myblock)
```

Vemos que chequea la balancing property. If p(X) is the propensity score then $X \perp D \mid p(X)$. The balancing hypothesis is that individuals with the same propensity score must have the same distribution of observable characteristics independently of treatment status. In other words, for a given propensity score, exposure to treatment is random and therefore Di=1 and Di=0 units should be on average observationally identical. Uno puede testear si esto efectivamente se cumple una vez calculados los propensities scores y generados los estratos.

Stata no nos permitiría continuar si la balancing property no se satisface. El comando "pscore" tiene algunas opciones: "detail" displays more detailed output; "logit" uses a logit instead of a probit model; "level(.)" allows to set the significance level of the tests of the balancing property (the default is 0.01); "numblo(.)" sets the number of blocks.

El primer comando es el "atts", que calcula el ATT utilizando el stratification matching:

```
. atts score treated, pscore(myscore) blockid(myblock)
```

Vemos que el valor estimado es sumamente similar al obtenido por OLS. Notar que utilizamos como argumentos en el comando "atts" a "myscore" y "myblock", que debimos guardar previamente con el comando "pscore". Esas variables guardan la probabilidad de ser seleccionado para el tratamiento y el bloque de propensities scores, respectivamente.

El resto de los comandos no requiere indicar el "pscore" y el "blockid". Simplemente pueden ser ejecutados después del comando "pscore" (en caso contrario, calcularán primero el propensity score con las variables indicadas). Por ejemplo, el comando "attr" calcula el ATT utilizando el radius matching:

```
. attr score treated female income hhsize divorced violence, pscore(myscore)
```

No obstante, seguiremos con la especificación anterior:

```
. attr score treated, pscore(myscore) blockid(myblock)
```

The option "bootstrap" bootstraps the standard error of the treatment effect:

```
. attr score treated, pscore(myscore) boot
```

El comando "attk" calcula el ATT utilizando el kernel-based matching:

```
. attk score treated, pscore(myscore) blockid(myblock)
```

El comando "attnd" calcula el ATT utilizando el nearest neighbor matching:

```
. attnd score treated, pscore(myscore) blockid(myblock)
```

Para extender todos estos conceptos a ordered and continuous treatments, puede ver Angrist and Krueger (1999).

Common support

La comparación de los propensity score entre los tratamientos y los controles puede ser una herramiento de diagnóstico útil cuando uno quiere evaluar que tanto similares son los tratamientos y los controles, y que tan confiable es la estrategia de estimación. Por ejemplo, uno esperaría que el rango de variación del propensity score debería ser lo mismo para los tratamientos y los controles.

En nuestro ejemplo, podemos graficar la distribución de los propensity scores para el grupo control y el grupo tratamiento, en determinados stratos:

```
. gen one=1
```

```
. graph twoway scatter one myscore if treated==1 || scatter one myscore if treated==0, by(myblock)
```

When using pscore, la opción "comsup" restricts the analysis of the balancing property to all treated plus those controls in the region of common support, and creates a dummy variable named "it:comsup" to identify the observations in the common support.

Para el resto de los comandos (att*), la opción "comsup" restricts the computation of the ATT to the region of common support. Lo que hace es ???. En nuestro ejemplo:

```
. atts score treated, pscore(myscore) blockid(myblock) comsup
```

However, imposing the common support restriction is not necessarily better (see Lechner, 2001).

Differences in Differences Matching (2 periods)

Tomemos el ATT de las primeras diferencias:

E(Yi1t’ - Yi0t | Di = 1) – E(Yi0t’ - Yi0t | Di = 1)

Where t’ is the time period after treatment and t is the time period before treatment. Entonces la condición de uncounfoundness en este caso sería:

E(Y0it’ - Y0it| Zi, Di=0) = E(Y0it’ - Y0it| Zi, Di=1)

Por lo que el ATT de las primeras diferencias sería:

E(Yi1t’ - Yi0t | Zi, Di = 1) – E(Yi0t’ - Yi0t | Zi, Di = 0)

Donde el segundo término si lo observamos. The required assumption is that the mean change in the no-treatment outcome is the same for treated and non-treated individuals. Y la implementación es sumamente fácil: se debe hacer matching utilizando datos de los before-afters.

Abramos la base "psm1.dat", en las cuales hay datos para los scores un período antes del tratamiento:

```
. use psm1, clear
```

Vamos a generar las primeras diferencias. Como los individuos están ordenados, podemos hacer:

```
. gen score_dif = score- score[_n+300] if _n<301
```

Dropeamos las observaciones que no nos sirven:

```
. drop if _n>300
```

Estimemos differences-in-differences por OLS:

```
. reg score_dif treated female income hhsize divorced violence, robust
```

Y ahora differences-in-differences matching:

```
. pscore treated female income hhsize divorced violence, pscore(myscore) blockid(myblock)
```

```
. atts score_dif treated, pscore(myscore) blockid(myblock)
```

Los resultados son mejores que con la estrategia original.

Comparing OLS and Matching

To be completed.

Aproximating randomization with nonrandom data

To be completed.

Problem Set #16

To be completed.

Chapter 17

Time Series

To be completed.

Problem Set #17

To be completed.

Chapter 18

Nonparametric Methods

Nonparametric kernel density estimation

Intuitively, you can think of density estimation as a histogram with infinitely small bins. This has obviously direct applications for describing data. For instance, you can present the evolution of the income distribution in a particular population. But we are more interested in appliying this concept to regression methods, which we will pursue after this subsection.

Let $f(\cdot)$ be the density distribution of the random variable X. Let $x_1, x_2, \dots, x_n$ be a random sample of observations. How could we estimate the density in a particular point x_0 (i.e. $\hat{f}(x_0)$) from the sample?

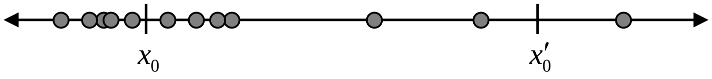
The parametric methods assume a particular functional form for $f(\cdot)$ and then estimates the parameters of such distribution. For instance, we can conjecture that $X \to N(\mu, \sigma^2)$ and then get consistent estimates:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \hat{\sigma}^2 == \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 .$$

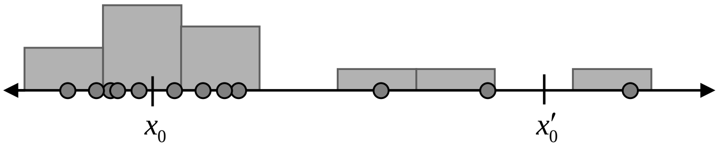
Finally, we can plug them back in the formula for the density distribution:

$$\hat{f}(x_0) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{1}{2\hat{\sigma}^2}(x_0 - \hat{\mu})^2}$$

Nonparametric estimators try to get consistent estimates without assuming a particular functional form for $f(\cdot)$. First you should ask yourself: what does it mean that $f(x_0) > f(x_0')$? Intuitively, it means that if we observe samples from the random variable, we should observe in general more points around x_0 than around x_0':



The histogram is indeed a rudimentary density estimator. In the case of the histogram we define “around” as fixed intervales in the domain of X. Such intervals start in the initial point *a* and have fixed lenght given by *h*. The relative area of each bar is given by the relative frequency of observations inside each corresponding interval. That is, the height of the bars equals the number of observations in the interval divided by $n \cdot h$. You can check that the areas of the bars sum to one.



There are three basic problems with histograms. In the first place, the results are sometimes very sensitive to the choice of the initial point. Secondly, they are sometimes sensitive to the choice of the bandwidth. Finally, it is discontinuous at the end of the intervals.

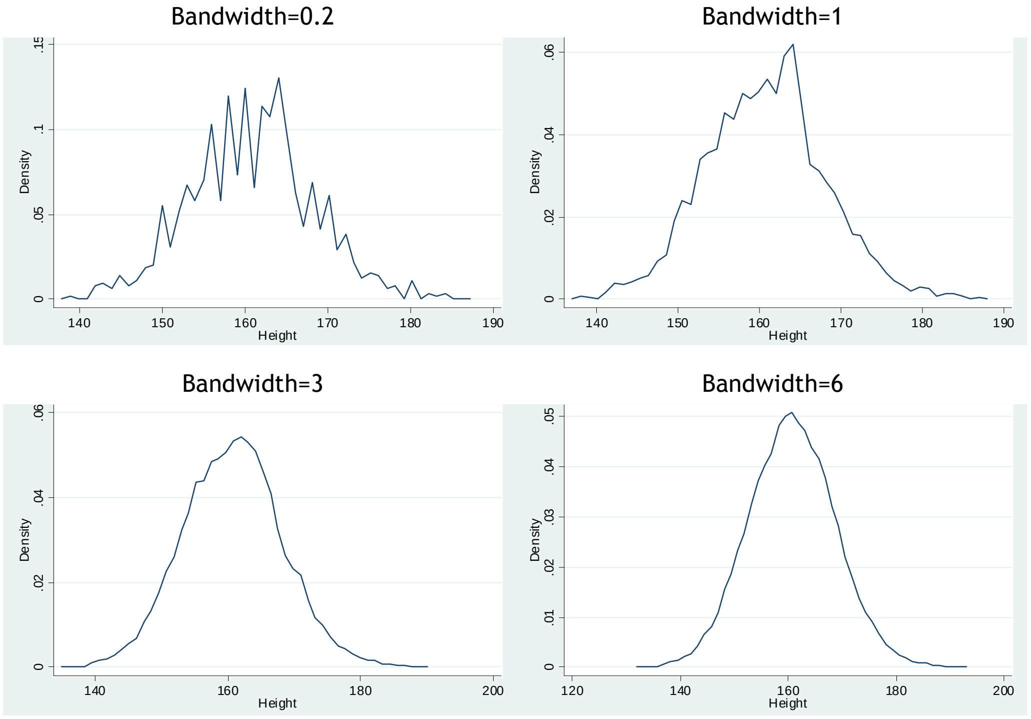
We can illustrate this problems generating some random data with known distribution and then manipulating the commands to yield very different pictures:

```
. clear

. set seed 13941

. set obs 30

. gen x=invnorm(runiform())*100

. hist x, width(10)

. hist x, width(20)

. hist x, width(50)

. hist x, width(100)
```

And we can play similarly changing the starting value. Notice that all of those problems could be easily solved if the underlying random variable was discrete (you would have to use the “discrete” option). We go then to a second nonparametric estimator (known as “naive”) that will solve the first issue, and after some modifications will partially solve the rest.

If we want the density at point x_0, we must define a neighborhood around x_0 as the segment with center x_0 and length h. Now let the density at x_0 be the proportion of observations in the sample that lies in its corresponding interval:

$$\hat{f}(x_0) = \frac{1}{n}\sum_{i=1}^{n} I\left(x_0 - \frac{h}{2} \le x_i \le x_0 + \frac{h}{2} \right) = \frac{1}{n}\sum_{i=1}^{n} I\left(-\frac{1}{2} \le \frac{x_i - x_0}{h} \le \frac{1}{2} \right)$$



Where $I(\cdot)$ is an indicator function that takes the value 1 if the condition in parenthesis is satisfied and zero otherwise. Intuitively, we are aproximating the probability in the interval (the area below the density function) with a rectangle with base *h* and height $\hat{f}(x_0)$.

Each x_0 has its own interval, and then there is no dependence on the initial point. See for example the points x_0 and x_0' in the previous figure, and notice also that the intervals can overlap.

Notwithstanding, the naive estimator still depends on the choice of *h*: if *h* is greater then observations further from x_0 will be used in the estimation of $\hat{f}(x_0)$. Intuitively, a greater *h* implies an smoother estimate for $\hat{f}(\cdot)$.

The corresponding Stata command is “kdensity” along with the option “rectangle”:

```
. use russia, clear
```

```
. kdensity height if gender==0, rectangle
```

Notice that Stata automatically chose an “optimal” bandwitdth (according to a “desirable” combination of bias and variance). Let’s show the effect of changing the bandwidth:

```
. kdensity height if gender==0, rectangle bw(0.2)
```

…

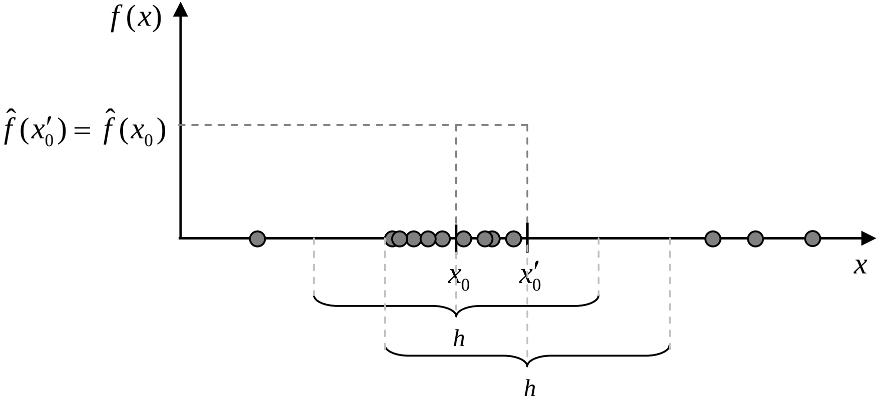
```
. kdensity height if gender==0, rectangle bw(6)
```

You can see the results in the Figure above.

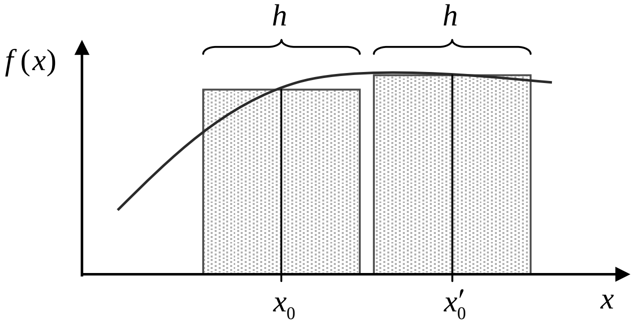
One of the problems with the naive estimator is that the function I(·) use to “weight” the observations is discontinuous in the boundaries of each interval. Moreover, it also treats equally observations very near x_0 than observations far from (as long as they are inside the interval of width *h* centered in x_0). You can see this clearly in the following figure:

We have $\hat{f}(x_0) = \hat{f}(x_0')$ because there are the same number of observations in the intervals of x_0 and x_0', even though you can clearly see that the almost all of the observations are closer to x_0.

The Kernel method solves some of the problems with the naïve estimator. The indicator function $I(\cdot)$ is replaced by a new ponderator $K(\cdot)$, which must satisfy the following two properties: i. $K(c) \geq 0$; ii. $\int_{-\infty}^{\infty} K(c)dc = 1$. The function $I(\cdot)$ is also known as the rectangular or uniform Kernel. It is straightforward to check that it satisfyies the two properties above.



Another Kernel is the Gaussian, which corresponds to the density distribution function of the standard normal (which, as any other density distribution, satisfy the



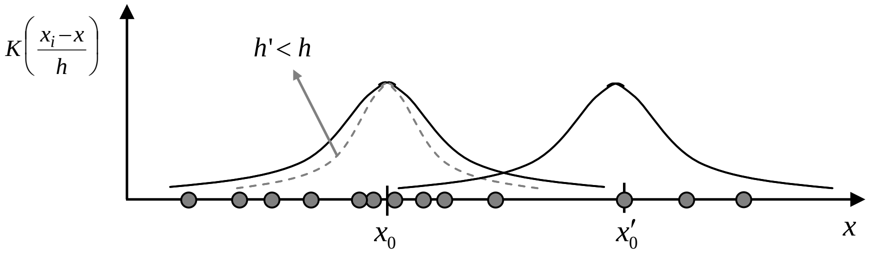
two properties above):

$$K(c) = \frac{1}{\sqrt{2\pi}} e^{-\frac{c^2}{2}}$$

And the the density estimator is:

$$\hat{f}(x_0) = \frac{1}{n \cdot h} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\frac{x_i - x_0}{h}\right)^2}{2}}$$

The intuition is simple. As in the naive estimaror, to estimate the density at x_0 this formula is also a "weighted" average of the all x_i's. The Gaussian kernel is a function that has a global maxima at zero and decreases smoothly and symetrically as the argument goes to minus or plus infinity. Since $(x_i - x_0)$ equals zero for $x_i = x_0$, the Kernel gives maximum weight to observations near x_0, and the weight decreases as the observations are further from x_0:



The advantage with respect to the naive estimator is that $K(\cdot)$ is continuous. The bandwidth h stills controls what means to be "near" x_0.

To be completed.

Regresion no-parametrica y semi-parametrica

To be completed.

Problem Set #18

To be completed.

Solutions to the Problem Sets

Problem Set #1

1. use russia, clear

```
. centile totexp, centile(11 22 33 44 55 66 77 88 100)
```

```
. return list
```

```
. gen objective=(totexp< r(c_1))+2*( r(c_1) <= totexp & totexp < r(c_2) )+ 3*( r(c_2)<=
totexp & totexp < r(c_3) )+ 4*( r(c_3) <= totexp & totexp < r(c_4) )+ 5*( r(c_4) <=
totexp & totexp < r(c_5) )+ 6*( r(c_5) <= totexp & totexp < r(c_6) )+ 7*( r(c_6) <=
totexp & totexp < r(c_7) )+ 8*( r(c_7) <= totexp & totexp < r(c_8) )+ 9*( totexp >=
r(c_8) )
```

```
. tab obj econ, chi2
```

The option “chi2” calculates and displays Pearson's chi-squared for the hypothesis that the rows and columns in a two-way table are independent. The low p-value indicates that the variables are not independent, as expected. We can use “cell” to cell display the relative frequency of each cell:

```
. tab obj econ, cell chi2
```

If people were perfectly informed about their relative wealth, we should see that most observations lay in the diagonal. On the contrary, we observe that rich people tend to subestimate their wealth and poor people tend to overestimate their position. See Cruces et al. (2009) for an explanation.

However, expenditure is a very incomplete description of the objective welfare of the people. People may be thinking about their raw incomes instead of their standard of living when answering the question about subjective relative welfare. However, income is much more volatile than consumption, so we should expect less precision from this measure. We should simply change “totexp” in the code:

```
. centile tincm, centile(11 22 33 44 55 66 77 88 100)
```

```
. return list
```

```
. gen objective_inc=(tincm< r(c_1))+2*( r(c_1) <= tincm & tincm < r(c_2) )+ 3*(
r(c_2)<= tincm & tincm < r(c_3) )+ 4*( r(c_3) <= tincm & tincm < r(c_4) )+ 5*( r(c_4)
<= tincm & tincm < r(c_5) )+ 6*( r(c_5) <= tincm & tincm < r(c_6) )+ 7*( r(c_6) <=
tincm & tincm < r(c_7) )+ 8*( r(c_7) <= tincm & tincm < r(c_8) )+ 9*( tincm >= r(c_8) )
```

```
. tab objective_inc econ, cell chi2
```

We see that both income and expenditures describe a similar picture. Another possible improvement is dividing the analysis by region. When a person in Moscu is guessing his position in the wealth distribution it is more probably that he is thinking about other people in Moscu that (say) people in Siberia. Therefore, we should expect a best “fit” between subjective and objective measures if we run the analysis in a region-by-region basis:

```
. centile totexp if geo==1, centile(11 22 33 44 55 66 77 88 100)

. return list

. gen obj1=(totexp< r(c_1))+2*( r(c_1) <= totexp & totexp < r(c_2) )+ 3*( r(c_2)<=
totexp & totexp < r(c_3) )+ 4*( r(c_3) <= totexp & totexp < r(c_4) )+ 5*( r(c_4) <=
totexp & totexp < r(c_5) )+ 6*( r(c_5) <= totexp & totexp < r(c_6) )+ 7*( r(c_6) <=
totexp & totexp < r(c_7) )+ 8*( r(c_7) <= totexp & totexp < r(c_8) )+ 9*( totexp >=
r(c_8) )

. replace obj1=. if geo!=1

. tab obj1 econ if geo==1, cell
```

And you can do the same with the other two regions.

2. tab powr econ, cell chi2

```
. tab powr resp, cell chi2

. tab econ resp, cell chi2

. bysort gender: tab power econ, cell chi2
```

And so on.

3. Define “orh” as the over-reporting of height in centimeters:

```
. gen orh=htself-height
```

And the same for economic rank:

```
. gen ore=econ-obj
. graph twoway scatter orh ore, by(gender) title("Height and wealth over-reporting") subtitle("All regions") note("1") caption("Source:  RLMS")
```

```
. pwcorr orh ore, sig
```

The variables are not significantly correlated.

Problem Set #2

There is no “right” answer to the first part of the Problem Set. The second part is simply “playing” with the parameters in the simulation.

Problem Set #3

To be completed.

Problem Set #4

```
. use schooling, clear
```

1.

```
tabstat age66 smsa66r reg66* nearc4 momdad14 sinmom14 daded momed black kww smsa76r ed76 reg76r, stats(mean)
. tabstat age66 smsa66r reg66* nearc4 momdad14 sinmom14 daded momed black kww smsa76r ed76 reg76r if smsa76r & nodaded==1 & nomomed==1, stats(mean)
. tabstat age66 smsa66r reg66* nearc4 momdad14 sinmom14 daded momed black kww smsa76r ed76 reg76r if smsa76r & nodaded==1 & nomomed==1 & lwage76!=., stats(mean)
```

2.

```
reg lwage76 ed76 exp76 exp762 black reg76r smsa76r, robust
. outreg using table2, se 3aster replace
. reg lwage76 ed76 exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r, robust
. outreg using table2, se 3aster append
. reg lwage76 ed76 exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded, robust
. outreg using table2, se 3aster append
. reg lwage76 ed76 exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8, robust
. outreg using table2, se 3aster append
. reg lwage76 ed76 exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r
momed daded nomomed nodaded f1-f8 momdad14 sinmom14, robust
. outreg using table2, se 3aster append
```

3.

```
reg lwage76 ed76 exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r
momed daded nomomed nodaded f1-f8 momdad14 sinmom14, robust
. outreg using table4ols, se 3aster replace
. ivreg lwage76 (ed76 exp76 exp762 = age76 age76sq nearc4) black reg76r smsa76r
reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14 sinmom14,
robust
. outreg using table4iv, se 3aster replace
. reg lwage76 ed76 kww exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r
momed daded nomomed nodaded f1-f8 momdad14 sinmom14, robust
. outreg using table4ols, se 3aster replace
. ivreg lwage76 (ed76 exp76 exp762 = age76 age76sq nearc4) kww black reg76r
smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14
sinmom14, robust
. outreg using table4iv, se 3aster append
. reg lwage76 ed76 kww exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r
momed daded nomomed nodaded f1-f8 momdad14 sinmom14 if iq!=., robust
. outreg using table4ols, se 3aster replace
. ivreg lwage76 (ed76 exp76 exp762 kww = age76 age76sq nearc4 iq) black reg76r
smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14
sinmom14, robust
. outreg using table4iv, se 3aster append
. ivreg lwage76 (ed76 exp76 exp762 = age76 age76sq nearc4a) black reg76r smsa76r
reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14 sinmom14,
robust
. outreg using table4iv, se 3aster append
. ivreg lwage76 (ed76 exp76 exp762 = age76 age76sq nearc2 nearc4) black reg76r
smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14
sinmom14, robust
. outreg using table4iv, se 3aster append
```

4. I cannot test the exogeneity conditions. However, if I knew that the instruments are exogenous, then I can test the education's exogeneity using the Hausman's specification test. Using the baseline specification:

```
. ivreg lwage76 (ed76 exp76 exp762 = age76 age76sq nearc4) black reg76r smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14 sinmom14, robust
. est store iv
. reg lwage76 ed76 exp76 exp762 black reg76r smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14 sinmom14, robust
. est store ols
. hausman iv ols
```

I cannot reject the null hypothesis that the difference in the coefficients is not systematic. Then, I can use OLS directly.

5. Sargan test:

```
. ivreg lwage76 (ed76 exp76 exp762 = age76 age76sq nearc2 nearc4 nearc4a) black reg76r smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14 sinmom14, robust
. predict resid, residual
. reg resid age76 age76sq nearc2 nearc4 nearc4a black reg76r smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14 sinmom14
. ereturn list
. display chi2(2,e(N)*e(r2))
```

The null hypothesis that all instruments are exogenous is not rejected at any conventional level. We cannot conclude that at least one of those instruments is invalid.

6.

```
reg ed76 age76 age76sq nearc4 black reg76r smsa76r reg662-reg669 smsa66r momed daded nomomed nodaded f1-f8 momdad14 sinmom14, robust
.test nearc4==0
```

The F statistic is greater than 10. Then, (following SW) it is not a weak instrument.

7. Be creative!

Problem Set #5

```
. set mem 5m
. use crime, clear
```

First we must create two dummy variables: inst3 (indicating whether the nearest institution was more than two blocks away) and inst2 (indicating whether the nearest institution was exactly two blocks away):

```
. gen inst2=(dist==2)
. gen inst3=(dist>2)
```

Generate the interactions with "post" and replace the database:

```
. gen inst2p= inst2*post
. gen inst3p= inst3*post
. save crime, replace
```

1. For, Table 2, column (A), (B), (C) and (D):

```
. replace month=7.5 if month==72 | month==73
. bysort month: tabstat cartheft if inst3==1, stats(mean sd count)
. bysort month: tabstat cartheft if inst==1, stats(mean sd count)
. bysort month: tabstat cartheft if inst1==1, stats(mean sd count)
. bysort month: tabstat cartheft if inst2==1, stats(mean sd count)
```

Copy and paste to MS Excel to build the table. Columns (E) and (G) can be obtained by subtracting appropriately.

For Table A1, first row:

```
. use crime, clear
. bysort barrio: tabstat blockid if month==5, stats(count)
. tabstat blockid if month==5, stats(count)
```

Second row:

```
. bysort barrio: tabstat blockid if month==5 & inst==1, stats(count)
. tabstat blockid if month==5 & inst==1, stats(count)
```

Third row:

```
. bysort barrio: tabstat cartheft if month>3 & month<8, stats(sum)
. tabstat cartheft if month>3 & month<8, stats(sum)
. bysort barrio: tabstat cartheft if month==72 | month==73, stats(sum)

. tabstat cartheft if month==72 | month==73, stats(sum)

. bysort barrio: tabstat cartheft if month>7 & month<13, stats(sum)

. tabstat cartheft if month>7 & month<13, stats(sum)
```

For (the monthly version of) Figure 2, you must create a "small" database containing the monthly average car theft for the different groups (in the same block than Jewish institution, one block away, two blocks away and more than two blocks away):

```
. use crime, clear

. replace month=7.5 if month==72

. replace month=7.8 if month==73

. preserve

. keep if inst==1

. collapse (mean) cartheft, by(month)

. gen inst=0

. save inst,replace

. restore

. preserve

. keep if inst1==1

. collapse (mean) cartheft, by(month)

. gen inst=1

. save inst1,replace

. restore

. preserve

. keep if inst2==1

. collapse (mean) cartheft, by(month)

. gen inst=2

. save inst2,replace

. restore

. keep if inst3==1
. collapse (mean) cartheft, by(month)
. gen inst=3
. append using inst
. append using inst1
. append using inst2
```

And then just make the desired graph:

```
. graph twoway line cartheft month if inst==0 || line cartheft month if inst==1 || line cartheft month if inst==2 || line cartheft month if inst==3, title("Monthly average car thefts") legend(off)
```

```
2. use crime, clear
. drop if month>70
. xtreg cartheft instp month5-month12, fe i(blockid) robust
. xtreg cartheft instp inst1p month5-month12, fe i(blockid) robust
. xtreg cartheft instp inst1p inst2p month5-month12, fe i(blockid) robust
. reg cartheft instp inst1p inst2p month9-month12 if month>7, robust
. xtreg cartheft instp inst1p inst2p if dist<3, fe i(blockid) robust
```

```
3. use crime, clear
. drop if month>70
```

Generate the interactions and run the regressions. Run the F-test when needed:

```
. gen instpb=instp*bank
. gen instpmb=instp*(1-bank)
. gen inst1pb=inst1p*bank
. gen inst1pmb=inst1p*(1-bank)
. gen inst2pb=inst2p*bank
. gen inst2pmb=inst2p*(1-bank)
. xtreg cartheft instpb instpmb inst1pb inst1pmb inst2pb inst2pmb month5-month12, fe i(blockid) robust
. ttest instpb==instpmb
. gen instpp=instp*public
. gen instpmp=instp*(1-public)
. gen inst1pp=inst1p*public
. gen inst1pmp=inst1p*(1-public)
. gen inst2pp=inst2p*public
. gen inst2pmp=inst2p*(1-public)
. xtreg cartheft instpp instpmp inst1pp inst1pmp inst2pp inst2pmp month5-month12, fe i(blockid) robust
. ttest instpp==instpmp
. gen instps=instp*station
. gen instpms=instp*(1-station)
. gen inst1ps=inst1p*station
. gen inst1pms=inst1p*(1-station)
. gen inst2ps=inst2p*station
. gen inst2pms=inst2p*(1-station)
. xtreg cartheft instps instpms inst1ps inst1pms inst2ps inst2pms month5-month12, fe i(blockid) robust
. ttest instps==instpms
. gen any=(bank==1 | station==1 | public==1)
. gen instpa=instp*any
. gen instpma=instp*(1-any)
. gen inst1pa=inst1p*any
. gen inst1pma=inst1p*(1-any)
. gen inst2pa=inst2p*any
. gen inst2pma=inst2p*(1-any)
. xtreg cartheft instpa instpma inst1pa inst1pma inst2pa inst2pma month5-month12, fe i(blockid) robust
. ttest instpa==instpma
```

4. Be creative!

Problem Set #6

To be completed.

Problem Set #7

To be completed.

Problem Set #8

Part 1

```
. use russia, clear
```

1.

```
logit alclmo gender monage highsc belief obese smokes hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust
```

For instance, being male is positively associated with drinking alcohol. The coefficient is also statistically different from zero at the 1% level.

2.

```
logit alclmo gender monage highsc belief obese smokes hattac cmedin totexpr tincm_r work0 marsta1 marsta2, robust

. mfx

. margeff
```

For instance, for a "mean individual" the impact of being male on the probability of drinking alcohol is about 12 percentage points. On the other hand, the mean marginal effect is about ??.

3.

```
gen age=monage/12

. probit alclmo age highsc belief obese smokes hattac cmedin totexpr tincm_r work0 marsta1 marsta2 if gender==0, robust

. quietly: ereturn list

. matrix coef=e(b)

. matrix x_mean = (.67, 4.12, .25, .17, .021, .911, 7525.84, 5705.44, .457, .472, .063, 1)

. gen age_aux=18+(60*_n/100)

. matrix xb=x_mean*coef[1,"highsc".."_cons"]'

. gen alclmo_hat=normal(xb[1,1]+coef[1,1]*age_aux) if _n<101

. gen age_me = coef[1,1]*normalden(xb[1,1]+coef[1,1]*age_aux) if _n<101
. graph twoway line age_me age_aux  if _n<101

. graph twoway line alclmo_hat age_aux  if _n<101
```

The marginal effects are greater at the extremes of "age". However, the differences are so slight that the impact is practically linear. Do the same with "gender==1".

Part 2

```
. clear

. set mem 10m

. use russia1
```

1. xtlogit smokes alclmo monage belief highsc obese totexpr tincm_r work0 work1 ortho marsta1 marsta2 marsta3 round2 round3, fe i(eid)

Compute the marginal effects for the predicted probability of smoking:

```
. mfx compute, predict(pu0)
```

2. Vamos a quedarnos con datos solo para los primeros dos years:

```
. drop if round3==1
```

Generamos las primeras diferencias para todas las variables involucradas:

```
. foreach var in smokes alclmo monage belief highsc obese totexpr tincm_r work0 work1 ortho marsta1 marsta2 marsta3 round2 {

. gen `var'_dif=.

. quietly: bysort eid: replace `var'_dif=  `var'[2]-`var'[1] if _n==1

. }
```

Terminamos de generar la nueva variable dependiente:

```
. drop if smokes_dif==0

. replace smokes_dif=0 if smokes_dif==-1
```

Corremos el modelo “by hand”:

```
. logit smokes_dif alclmo_dif monage_dif belief_dif highsc_dif obese_dif totexpr_dif tincm_r_dif work0_dif work1_dif ortho_dif marsta1_dif marsta2_dif marsta3_dif round2_dif, nocons
```

Y vemos que los estimates son idénticos a los obtenidos con el comando “xtlogit”:

```
. xtlogit smokes alclmo monage belief highsc obese totexpr tincm_r work0 work1 ortho marsta1 marsta2 marsta3 round2, fe i(eid)
```

3. Nos tenemos que quedar con los individuos que al menos una vez cambian from smoker to non-smoker and/or from non-smoker to smoker:

```
. clear

. set mem 500m

. set matsize 800

. use russia1

. drop if smokes!=1 & smokes!=0

. drop if alclmo==. | monage==. | belief==. | highsc==. | obese==. | totexpr==. |
tincm_r==. | work0==. | work1==. | ortho==. | marsta1==. | marsta2==. |
marsta3==. | round2==. | round3==.

. bysort eid: drop if _N==1

. bysort eid: drop if smokes[1]==smokes[2] & _N==2

. bysort eid: drop if smokes[1]==smokes[2] & smokes[2]==smokes[3] & _N==3
```

Como Stata no me deja trabajar con mas de 800 variables, tengo que asegurarme de tener menos de 800 individuos. De lo contrario, no podría incluir una dummy por individuo. Por eso saco al 25% de los individuos:

```
. drop if eid>1.20e+08
```

Ahora corro la regression con el set de dummies:

```
. xi: logit smokes alclmo monage belief highsc obese totexpr tincm_r work0 work1
ortho marsta1 marsta2 marsta3 round2 round3 i.eid
```

Y el modelo consistente de logit con fixed effects:

```
. xtlogit smokes alclmo monage belief highsc obese totexpr tincm_r work0 work1
ortho marsta1 marsta2 marsta3 round2 round3, fe i(eid)
```

The results are pretty different. In summary, the problem is a computacional one. With the individual dummies, there are almost 800 variables. Then, the solution of the maximization problem moving 800 parameters is unfeasible for current processors. See Greene (2001a, 2001b) for further details.

4. To be completed.

5. Run both models:

```
. xtlogit smokes gender alclmo monage belief highsc obese totexpr tincm_r work0
work1 ortho marsta1 marsta2 marsta3 round2 round3, re i(eid)
```

```
. xtprobit smokes gender alclmo monage belief highsc obese totexpr tincm_r work0 work1 ortho marsta1 marsta2 marsta3 round2 round3, re i(eid)
```

Compute the marginal effects for the predicted probability of smoking:

```
. xtlogit smokes gender alclmo monage belief highsc obese totexpr tincm_r work0 work1 ortho marsta1 marsta2 marsta3 round2 round3, re i(eid)

. mfx compute, predict(pu0)

. xtprobit smokes gender alclmo monage belief highsc obese totexpr tincm_r work0 work1 ortho marsta1 marsta2 marsta3 round2 round3, re i(eid)

. mfx compute, predict(pu0)
```

Para consistencia solo me interesa el de fixed effects. Como vimos antes, lo que me interesa es ver como una variable esta asociada a que una persona deje o empiece a fumar.

Problem Set #9

```
. use duration, clear

. stset time, failure(died)
```

Kaplan-Meier non-parametric survival estimator:

```
. sts graph, by(female)
```

The proportional hazard model with a Weibull distribution:

```
. streg female married treatment, nohr distribution(weibull)
```

The non-parametric Cox model:

```
. stcox female married treatment, robust basesurv(base_survival)

. clear

. set mem 10m

. use duration

. expand time

. sort id

. quietly by id: gen month = _n

. gen death=0

. quietly by id: replace death = died if _n==_N
```

. gen ln_t=log(time)

. logit death female married treatment

Problem Set #10

use couart, clear

1. tab art

. hist art, discrete

The answer is yes. It takes integer values (including zero), the 76% of the observations are below 3, and particularly 30% of the sample consists of zeros.

2. poisson art fem ment phd mar kid5, robust

For instance, the coefficient on kid5 indicates that an additional child (less than 5 years old) is associated with a 20% decrease in the number of publications.

3. nbreg art fem ment phd mar kid5

The LR-test rejects the Poisson Regression Model. So, you must use the NB results:

. nbreg art fem ment phd mar kid5, robust

Problem Set #11

1. use mroz, clear

. reg lwage educ exper expersq, robust

. heckman lwage educ exper expersq, select(educ exper expersq nwifeinc age kidslt6 kidsge6) twostep

The estimates are practically the same.

2. heckman lwage educ exper expersq, select(educ exper expersq) twostep

. heckman lwage educ exper expersq, select(educ exper expersq nwifeinc age kidslt6 kidsge6) twostep

. heckman lwage educ exper expersq nwifeinc age kidslt6 kidsge6, select(educ exper expersq nwifeinc age kidslt6 kidsge6) twostep

3. heckman lwage educ exper expersq, select(educ exper expersq nwifeinc age kidslt6 kidsge6)

Under joint normality of the errors from the selection and regression models, the maximum-likelihood estimator will be more efficient than the two-step procedure. However, the joint normality restriction is more restrictive. Thus, the ML estimator is less robust than the two-step procedure. In addition, it is sometimes difficult to get the problem to converge.

4.

```
gen s=(lwage!=.)

. probit s educ exper expersq nwifeinc age kidslt6 kidsge6

. predict xb, xb

. gen lambda = normden(xb)/normal(xb)

. regress lwage educ exper expersq lambda if s==1, robust
```

And compare it:

```
. heckman lwage educ exper expersq, select(educ exper expersq nwifeinc age kidslt6 kidsge6) twostep
```

Notice that obtaining the two-stage estimates "by hand" you cannot use the standard errors directly.

5. The differences between the OLS and Heckman estimates are practically inexistent. Not surprisingly, the inverse Mills ratio term is statistically not different from zero:

```
. test lambda
```

Problem Set #12

Part 1

1.

```
use mroz, clear

. tobit hours nwifeinc educ exper expersq age kidslt6 kidsge6, ll(0)
```

No, we cannot compare the OLS and Tobit coefficients directly. The Tobit coefficients equal the marginal effects on the latent variable.

2.

```
version 6.0

. tobit hours nwifeinc educ exper expersq age kidslt6 kidsge6, ll(0)

. dtobit

. version 9.1
```

For example, conditional on hours being positive, a year of education (starting from the mean values of all variables) is estimated to increase expected hours by about 34.27 hours. On the other hand, without conditioning on hours being positive, a year of education increase expected hours by 48.73 hours. The magnitudes of the partial effects on expected hours are larger than when we condition on working women.

3.

```
reg hours nwifeinc educ exper expersq age kidslt6 kidsge6, robust

. reg hours nwifeinc educ exper expersq age kidslt6 kidsge6 if hours>0, robust
```

The Tobit coefficient estimates are in general the same sign as the corresponding OLS estimates. However, they differ substantially.

4.

```
tobit hours nwifeinc educ exper expersq age kidslt6 kidsge6, ll(0)

. est store A

. tobit hours nwifeinc educ exper expersq age, ll(0)

. lrtest A
```

These two variables are jointly significant.

Part 2

1.

```
clear

. set mem 50m

. use hinc

. xttobit hincome educ exper expersq married widowed divorcied, i(pid) ll(0)
```

No, we could not compare the least squares and Tobit coefficients directly.

2. Compute the marginal effects for the probability of being uncensored:

```
. mfx compute, predict(pr0(0,.))
```

And the marginal effects for the expected value of y conditional on being uncensored:

```
. mfx compute, predict(e0(0,.))
```

Finally, compute the marginal effects for the unconditional expected value of y:

```
. mfx compute, predict(ys(0,.))
```

Problem Set #13

Part 1 (Multinomial Logit)

1. use heat, clear

```
. describe
. tab heat
. mlogit heat ic1 ic2 ic3 ic4 ic5 oc1 oc2 oc3 oc4 oc5 income agehed rooms ncostl scostl mountn valley, b(5)
```

No, you cannot infer anything but the statistical significance.

2. For the marginal effects at the mean x (warning: it may take several minutes. Maybe you may try a simpler model):

```
. mfx compute, predict(outcome(1))
. mfx compute, predict(outcome(2))
. mfx compute, predict(outcome(3))
. mfx compute, predict(outcome(4))
. mfx compute, predict(outcome(5))
```

For instance, the prices and maintenance costs for almost every option appeared as insignificant. Indeed, almost every coefficient appeared insignificant. The reason is that 900 observations are too few to estimate a complete model like this.

However, those (prices and costs) that appeared significant had the right sign: the probability of choosing an option decreases with the prices (and maintenance costs) of that option and increases with the prices (and maintenance costs) of the other alternatives.

There are almost no improvements if we try with a simpler model:

```
. mlogit heat agehed income rooms, b(5)
```

We then have to collapse some options:

```
. mlogtest, combine
. gen heat1=(heat==1|heat==2)+(heat==3|heat==4)*2
. label values heat1 heat1label
. label define heat1label 1 "gas " 2 "electric" 3 "heat pump"
. tab heat1
. mlogit heat1 income rooms ncostl scostl mountn valley, b(0)
```

And the significance problem persists. You should try a conditional logit model (since the prices varies between individual and options). You can also try a hierarchical model, because (for instance) people might choose first between electric, gas and heat pump, and then between room and central.

3. Consider the most complex model:

```
. mlogit heat ic1 ic2 ic3 ic4 ic5 oc1 oc2 oc3 oc4 oc5 income agehed rooms ncostl
scostl mountn valley, b(5)

. forvalues x=1/5 {

. predict p`x', outcome(`x')

. }

. gen heat_hat = (p1==max(p1,p2,p3,p4,p5)) + (p2==max(p1,p2,p3,p4,p5))*2+
(p3==max(p1,p2,p3,p4,p5))*3+(p4==max(p1,p2,p3,p4,p5))*4+(p5==max(p1,p2,p3,p4,p
5))*5

. tab2 heat heat_hat
```

You can see that almost all the predictions were "heat_hat==1". The proportion of right predictions is 63%, buy just because 63% of the sample have chosen the same election. A dummy model "simply predict the most popular option" would have performed even better. We can see that issue disappears when we drop the observations with "heat==1":

```
. use heat, clear

. drop if heat==1

. mlogit heat ic2 ic3 ic4 ic5 oc2 oc3 oc4 oc5 income agehed rooms ncostl scostl
mountn valley, b(5)

. forvalues x=2/5 {

. predict p`x', outcome(`x')

. }

. gen heat_hat = (p2==max(p2,p3,p4,p5))*2 + (p3==max(p2,p3,p4,p5))*3 +
(p4==max(p2,p3,p4,p5))*4 + (p5==max(p2,p3,p4,p5))*5

. tab2 heat heat_hat
```

Now only 45% of the predictions are right. But the model "simply predict the most popular option" would have 39% of right predictions.

Part 2 (Ordered Probit)

1. use russia, clear

```
. oprobit evalhl gender highsc belief monage obese hattac smokes alclmo tincm_r work0 marsta1 marsta2 marsta3, robust
```

2. You should see the sign and the statistical significance of the coefficients. To estimate the marginal effects:

```
. mfx, predict(outcome(1))

. mfx, predict(outcome(2))

. mfx, predict(outcome(3))

. mfx, predict(outcome(4))

. mfx, predict(outcome(5))
```

Being a man, obese and having had a heart attack are some things that raise your probability of reporting yourself as not healthy.

```
. forvalues x=1/5 {

. predict p`x', outcome(`x')

. }

. gen evalhl_hat = (p1==max(p1,p2,p3,p4,p5)) + (p2==max(p1,p2,p3,p4,p5))*2+(p3==max(p1,p2,p3,p4,p5))*3+(p4==max(p1,p2,p3,p4,p5))*4+(p5==max(p1,p2,p3,p4,p5))*5

. tab2 evalhl evalhl_hat
```

The 52% of the predictions were right.

Problem Set #14

```
. clear
. set mem 15m
. set more off
. use incomes
```

1. bysort year: tabstat hincome educ age exper nonwhite, stats(mean p10 p25 p50 p75 p90)

Then put all the information together in a single Table (e.g. using MS Excel).

2. Run the following code:

```
forvalues t=1996(1)2005 {

reg hincome educ exper expersq nonwhite if year==`t', robust
outreg using year`t', se 3aster replace
qreg hincome educ exper expersq nonwhite if year==`t', quantile(.10)
outreg using year`t', se 3aster append
qreg hincome educ exper expersq nonwhite if year==`t', quantile(.25)
outreg using year`t', se 3aster append
qreg hincome educ exper expersq nonwhite if year==`t', quantile(.50)
outreg using year`t', se 3aster append
qreg hincome educ exper expersq nonwhite if year==`t', quantile(.75)
outreg using year`t', se 3aster append
qreg hincome educ exper expersq nonwhite if year==`t', quantile(.90)
outreg using year`t', se 3aster append
}
```

Then put all the information together in a single Table (e.g. using MS Excel). Each row will denote a year, and in the columns you should put the coefficient on education for all the different estimations (first OLS, then quantile(.1), quantile(.25), and so on).

3. Denote β_1 the coefficient on experience and β_2 the coefficient on experience squared. Notice that the marginal effect of experience is: $\beta_1 + 2 \cdot \beta_2 \cdot experience$. Thus, that marginal effect will vary for different values of years of experience. You can arbitrarily choose a level of experience, say 10 years. Then retrieve the coefficients on experience and experience squared from the output of the previous excercise, and calculate $\beta_1 + 20 \cdot \beta_2$. Finally, build a similar Table (rows for years, first column for the OLS estimate, second column for the quantile(.1), and so on).

4. Run the following code:

```
forvalues t=1996(1)2005 {
reg hincome educ if year==`t', robust
predict fittedols
qreg hincome educ if year==`t', quantile(.10)
predict fitted10
qreg hincome educ if year==`t', quantile(.25)
predict fitted25
qreg hincome educ if year==`t', quantile(.50)

predict fitted50

qreg hincome educ if year==`t', quantile(.75)

predict fitted75

qreg hincome educ if year==`t', quantile(.90)

predict fitted90

graph twoway scatter hincome educ if year==`t' & hincome<50 || line fitted10 fitted25 fitted50 fitted75 fitted90 fittedols educ if year==`t' & hincome<50, legend(label(1 "Hourly Wage") label(2 "Quantile .10") label(3 "Quantile .25") label(4 "Quantile .50") label(5 "Quantile .75") label(6 "Quantile .90") label(7 "OLS"))

graph save year`t', replace

drop fitted10 fitted25 fitted50 fitted75 fitted90 fittedols

}
```

Notice that “if hincome<50” is a way to keep the same y-axis for the 9 different graphs. Put the graphs together in a single page to facilitate the interpretation.

5. Using the results from the third exercise you should be able to answer to every point very concisely.

Problem Set #15

```
. use inactivity, clear
```

1. tab actdays

The answer is yes. It takes integer values (including zero), 85% of the observations are zero, and it has no obvious upper limit.

2. poisson actdays sex age agesq income levyplus freepoor freerepa

```
. nbreg actdays sex age agesq income levyplus freepoor freerepa
```

The LR-test rejects the Poisson Regression Model. So, you must use the NB results:

```
. nbreg actdays sex age agesq income levyplus freepoor freerepa, robust
```

3. probit ill sex age agesq income hscore chcond1 chcond2, robust

```
. predict illhat
. nbreg actdays sex age agesq income levyplus freepoor freerepa illhat, robust
```

First Stage				Second Stage			
dep var: ill	**coef**	**std error**	**p-value**	dep var: actdays	**coef**	**std error**	**p-value**
sex	0.148	0.042	0.000	sex	-0.061	0.116	0.597
age	-2.721	0.812	0.001	age	1.815	2.068	0.380
agesq	2.833	0.908	0.002	agesq	-1.366	2.238	0.542
income	-0.088	0.062	0.155	income	0.201	0.185	0.278
hscore	0.221	0.019	0.000	levyplus	-0.337	0.146	0.021
chcond1	0.798	0.046	0.000	freepoor	-0.379	0.250	0.129
chcond2	0.955	0.078	0.000	freerepa	-0.281	0.191	0.142
cons	0.447	0.140	0.001	illhat	4.318	0.312	0.000
				cons	-3.806	0.425	0.000

4. Follow the example given in the notes. To change the first step model to a probit instead of a logit, you have to take into account that the derivative of $\hat{z}$ with respect to model 1's index function now equals $\phi(\hat{x}_i\theta_1)$, since $\hat{z}_i = \Phi(\hat{x}_i\theta_1)$. As a consequence, the line calculating $\hat{C}$ must be changed accordingly (i.e. replacing "zhat*(1-zhat)" by "normalden(xb)").

If you do not want to obtain the derivatives for the NB likelihood function, you can use the "scores" approach from Hole (2006). To use a Negative Binomial Model instead of a Poisson Model in the second stage you have to take into account the additional parameter in the NBM when deriving the Murphy-Topel variance estimate. Now θ_2 has two elements: the regression coefficients β_2 and the auxiliary parameter α. It can be shown that the only correction necessary to allow for the presence of such auxiliary parameter is to replace X_2 with $\tilde{X}_2$ in the corresponding equations, where $\tilde{X}_2$ is defined as follows (Hole, 2006):

$$\tilde{X}_2 = \left(X_2, \frac{\partial L_2 / \partial \alpha}{\partial L_2 / \partial (x_2 \hat{\beta}_2)} \right)$$

```
. use inactivity, clear

. probit ill sex age agesq income hscore chcond1 chcond2, robust

. matrix V1r = .99 * e(V)

. predict illhat

. predict xb, xb

. nbreg actdays sex age agesq income levyplus freepoor freerepa illhat, robust
score(s2 a)

. matrix V2r = .99 * e(V)
. predict actdayshat
. scalar illb = _b[illhat]
. gen a_s = a/s2
. gen byte cons = 1
. matrix accum C = sex age agesq income hscore chcond1 chcond2 cons sex age agesq income levyplus freepoor freerepa illhat cons a_s [iw=s2*s2*normalden(xb)*illb], nocons
. matrix accum R = sex age agesq income hscore chcond1 chcond2 cons sex age agesq income levyplus freepoor freerepa illhat cons a_s [iw=s2*(ill-illhat)], nocons
. matrix C = C[9..18,1..8]
. matrix R = R[9..18,1..8]
. matrix M = V2r + (V2r * (C*V1r*C' - R*V1r*C' - C*V1r*R') * V2r)
. capture program drop doit
. program define doit, eclass
. matrix b = e(b)
. ereturn post b M
. ereturn local vcetype "Mtopel"
. ereturn display
. end
. doit
```

The resulting output is:

	Second Stage		
dep: actdays	**coef**	**std error**	**p-value**
sex	-0.061	0.125	0.624
age	1.815	2.141	0.397
agesq	-1.366	2.320	0.556
income	0.201	0.207	0.332
levyplus	-0.337	0.146	0.021
freepoor	-0.379	0.251	0.130
freerepa	-0.281	0.192	0.144
illhat	4.318	0.354	0.000
cons	-3.806	0.455	0.000

Notice that there are no large differences between the "naive" and the Murphy-Topel estimates (the latter are just slightly larger).

Problem Set #16

To be completed.

Problem Set #17

To be completed.

Problem Set #18

To be completed.

References

[1] Angrist, Joshua (2002), "Vouchers for Private Schooling in Columbia: Evidence from a Randomized Natural Experiment," *American Economic Review*, Vol. 92, pp. 1535-1558.

[2] Angrist, Joshua and Krueger, Alan (1991), "Does Compulsory Schooling Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, Vol. 106, pp. 976-1014.

[3] Angrist, Joshua and Krueger, Alan (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, Vol. 15, pp. 69-85.

[4] Angrist, Joshua and Krueger, Alan (1999), "Empirical Strategies in Labor Economics,"" in *Handbook of Labor Economics*, Vol. 3, edited by Orley Ashenfelter and David Card.

[5] Angrist, Joshua D. and Pischke, Jörn-Steffen (2008), "Mostly Harmless Econometrics: An Empiricist's Companion." NJ: *Princeton University Press*.

[6] Angrist, Joshua and Lavy, Victor (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, Vol. 114 (2), pp. 533-575.

[7] Anderson, T. W. and Hsiao, C. (1981), "Estimation of Dynamic Models With Error Components," *Journal of the American Statistical Association*, Vol. 76 (375), pp. 598-606.

[8] Arellano, M. and Bond, S. (1991), "Some Test of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*, Vol. 58 (2), pp. 277-297.

[9] Arellano, M. and Bover, O. (1995), "Another look at the instrumental variable estimation of error-component models," *Journal of Econometrics*, Vol. 68, pp. 29-51.

[10] Baum, Christopher F. (2001), "Stata: The language of choice for time series analysis?" *Stata Journal*, Vol. 1 (1), pp. 1-16.

[11] Bertrand, M, E. Duflo and S. Mullainathan (2004). "How Much Should We Trust Differences-in-Differences Estimates," *Quarterly Journal of Economics*, Vol. 119, pp. 249-275.

[12] Blundell, Richard, and Bond, Stephen (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, Vol. 87, pp. 115-143.

[13] Bottan, Nicolas and Perez Truglia, Ricardo (2009), “Deconstructing the Hedonic Treadmill,” Mimeo.

[14] Bound, John; Jaeger, David and Baker, Regina (1995), “Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Variables is Weak,” Journal of American Statistical Association, Vol. 90, pp. 443-450.

[15] Buchinsky, Moshe (1994), “Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression,” *Econometrica*, Vol. 62 (2), pp. 405-458.

[16] Buchinsky, Moshe (1991), “Methodological Issues in Quantile Regression” - Chapter 1 of “The Theory and Practice of Quantile Regression," Ph.D. dissertation, Harvard University.

[17] Buse, A. (1992), “The Bias of Instrumental Variable Estimators,” *Econometrica*, Vol. 60, pp. 173-180.

[18] Cruces, Guillermo; Perez Truglia, Ricardo and Tetaz, Martin (2009), “Biased self-perceptions of income rankings and attitudes towards redistribution,” Working Paper, CEDLAS.

[19] Cameron, Colin A.; Gelbach, Jonah and Miller, Douglas L. (2006), “Bootstrap-Based Improvements for Inference with Clustered Errors,” *NBER* Technical Working Paper No. 344.

[20] Card, David (1993), “Using Geographic Variation in College Proximity to Estimate the Return to Schooling,” NBER Working Paper No. 4483.

[21] Casella, George and Berger, Roger (2001), “Statistical Inference,” Duxbury.

[22] Chernozhukov, Victor, and Hansen, Christian (2007), “A Simple Approach to Heteroskedasticity and Autocorrelation Robust Inference with Weak Instruments,” Mimeo, MIT.

[23] Di Tella, Rafael and Schargrodsky, Ernesto (2004), “Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack,” *American Economic Review*, Vol. 94, pp. 115-133.

[24] Draper, Norman and Smith, Harry (2007), “Applied Regression Analysis.” *Wiley*.

[25] Efron, Bradley (1981), “Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods,” *Biometrika*, Vol. 68 (3), pp. 589-599.

[26] Eicker, F. (1967), “Limit theorems for regressions with unequal and dependent errors,” Proceedings of the fifth Berkeley symposium on

mathematical statistics and probability, Berkeley: University of California Press, Vol. 1, pp. 59-82.

[27] Flores-Lagunes, Alfonso (2007), “Finite Sample Evidence of IV Estimators under Weak Instruments,” *Journal of Applied Econometrics*, Vol. 22, pp. 677-694.

[28] Greene, William (2003), “Econometric Analysis.” NJ: *Prentice Hall*.

[29] Greene, W. (1992), “A statistical Model for Credit Scoring,” Working Paper No. EC-92-29, NYU.

[30] Guan, Weihua (2003), “From the help desk: Bootstrapped standard Errors,” *The Stata Journal*, Vol. 3, pp. 71–80.

[31] Hamilton, Lawrence (2006), "Statistics with Stata." *Brooks/Cole*.

[32] Hardin, James W. (2002), “The robust variance estimator for two-stage models,” *Stata Journal*, Vol. 2 (3), pp. 253-266.

[33] Hausman, Jerry (1978), “Specification Tests in Econometrics,” Econometrica, Vol. 46, pp. 1251-1271.

[34] Hayashi, Fumio (2000), “Econometrics.” NJ: Princeton University Press.

[35] Heckman, James (1978), “Dummy Endogenous Variables in a Simultaneous Equations System,” Econometrica, Vol. 46, pp. 695-712.

[36] Heckman, James (2000), “Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective,” *Quarterly Journal of Economics*, Vol. 115, pp. 45-97.

[37] Hole, Arne Risa (2006), “Calculating Murphy-Topel variance estimates in Stata: A simplified procedure,” *Stata Journal*, Vol. 6 (4), pp. 521-529.

[38] Holland, Paul W. (1986), “Statistics and Causal Inference,” *Journal of the American Statistical Association*, Vol. 81, pp. 945-970.

[39] Huber, P. J. (1967), “The behavior of maximum likelihood estimates under non-standard conditions,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press, Vol. 1, pp. 221-233.

[40] Imbens, Guido, and Angrist, Joshua (1994), “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, Vol. 62, pp. 467-476.

[41] Jenkins, S. (1995), "Easy Estimation Methods for Discrete-time Duration Models," *Oxford Bulletin of Economics and Statistics*, Vol. 57 (1), pp. 129-138.

[42] Kezdi, Gabor (2004), “Robust Standard Error Estimation in Fixed-Effects Panel Models,” *Hungarian Statistical Review Special*, Vol. 9, pp. 96-116.

[43] Koenker, Roger (2005), “Quantile Regression.” Cambridge: Cambridge University Press.

[44] Koenker, Roger, and Bassett, Gilbert (1978), “Regression Quantiles,” *Econometrica*, Vol. 46, pp. 33-50.

[45] Koenker, Roger and Hallock, Kevin F. (2001), “Quantile Regression,” *Journal of Economic Perspectives*, Volume 15 (4), pp. 143–156.

[46] LaLonde, Robert J. (1986), “Evaluating the Econometric Evaluations of Training Programs Using Experimental Data,” *American Economic Review*, Vol. 76, pp. 602-620.

[47] Lancaster, Tony (1990), “The econometric analysis of transition data,” *Econometric Society Monographs*.

[48] Murphy, K. M. and Topel, R. H. (1985), “Estimation and inference in two-step econometric models,” *Journal of Business and Economic Statistics*, Vol. 3 (4), pp. 370–379.

[49] Nichols, Austin and Schaffer, Mark (2007), “Clustered Errors in Stata,” Mimeo.

[50] Nickell, Stephen (1981), “Biases in Dynamic Models with Fixed Effects,” *Econometrica*, Vol. 49, 1417-1426.

[51] Pagan, Adrian and Aman, Ullah (1999), “Nonparametric Econometrics.” Cambridge: *Cambridge University Press*.

[52] Rabe-Hesketh, Sophia and Everitt, Brian (2007) “A Handbook of Statistical Analices Using Stata.” *Chapman & Hall/CRC*.

[53] Ravallion, Martin and Lokshin, Michael (1999), “Subjective Economic Welfare,” Policy Research Working Paper 2106, World Bank.

[54] Rosenbaum P., and Rubin, D. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, Vol. 70, pp. 41-55.

[55] Rubin, Donald (1973), “Matching to Remove Bias in Observational Studies,” *Biometrics*, Vol. 29, pp. 159-83.

[56] Rubin, Donald (1974), “Estimating the Causal Effects of Treatments in Randomized and Non-Randomized Studies,” *Journal of Educational Psychology*, Vol. 66, pp. 688-701.

[57] Savin, Eugene (1984), “Multiple Hypothesis Testing,” in Zvi Griliches and Michael D. Intriligator (eds.), “Handbook of Econometrics,” Amsterdam, North-Holland, Vol. 2, pp. 828-860.

[58] Stock, James H.; Wright, Jonathan H. and Yogo, Motohiro (2002), “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments,” *Journal of Business & Economic Statistics*, Vol. 20, pp. 518-529.

[59] Verbeek, Marno (2000), “A Guide to Modern Econometrics.” *John Wiley & Sons*.

[60] White, Halbert (1982), “Maximum likelihood estimation of misspecified models,” *Econometrica*, Vol. 50, pp. 1-25.

[61] Wooldridge, Jeffrey M. (2001), “Econometric Analysis of Cross Section and Panel Data.” Cambridge: *MIT Press*.

[62] Wooldridge, Jeffrey M. (2003), “Cluster-Sample Methods in Applied Econometrics,” *The American Economic Review*, Vol. 93 (2), pp. 133-138.