NINE SESSIONS TO INTRODUCE YOU TO STATA®

DOING SURVEY RESEARCH LAB EXERCISES

Printed version: January 10, 2019
GitHub repository: `https://github.com/Yuji-Shimohira-Calvo/DSR`

Yuji Shimohira-Calvo
University of Edinburgh
Department of Sociology
`y.shimohira@ed.ac.uk`

Contents

1 Lab Worksheet I 1

Getting data 1

2 Lab Worksheet II 5

Stata’s interface 5

3 Lab Worksheet III 10

Preparing data 10

4 Lab Worksheet IV 13

Univariate analysis and graphing 13

5 Lab Worksheet V 17

Bivariate analysis 17

6 Lab Worksheet VI 23

Statistical inference I 23

7 Lab Worksheet VII 29

Statistical inference II 29

8 Lab Worksheet VIII 32

Simple regression analysis 32

9 Lab Worksheet IX 36

Multiple regression analysis 36

1 Lab Worksheet I

Getting data

Let's start by opening the following URL: `http://ukdataservice.ac.uk`. The **UK Data Service (UKDS)** offers a single point of access to a vast number of secondary data ranging from qualitative data to international macro-data. This comprehensive archive of high-quality data is funded by the **Economic and Social Research Council (ESRC)**. For this first lab session we are going to use data coming from the **European Social Survey (ESS)** 2012 for the United Kingdom.

- Click on the **Get Data** tab situated at the central panel of the website.
- You could now type "european social survey" in the search box. However, you are going to use instead the **Quick Access** side panel. Now click on **Cross-national surveys** and scroll down to "European Social Survey." Click on it.

What you see now is a page that shows basic information about all rounds of the ESS. For instance, under "Title Details" you can see that the ESS project is funded by the Commission of the European Communities, the European Science Foundation, and other national funding bodies of Europe. If you scroll down you will find information about subject categories, main topics of each round, and how to access the data among many other useful things. Let's clink now on "Access online" (marked with a white star on a blue background at the top of the page). You should now at the website of the ESS project, where you can easily navigate through one tab to another. These tabs offer well-structured information on everything you need to know about the ESS. Under **About ESS** you will find institutional information like how the project is funded, what countries participate in the making of the ESS, and so on. The **Methodology** tab presents technical details like sampling strategy, pre-testing, piloting, and so on. **Data and Documentation** allows you to navigate through all rounds of the ESS in three different ways: by year, by country, and by theme. Let's now take a look at the questionnaire used in the United Kingdom in 2012 (sixth round):

- Mouse hover over **Data and Documentation** and click on **Data and Documentation by Country**. This will show you a list of all countries that have taken part in the ESS.
- Scroll down and click on "United Kingdom." Under **Documents** click on "ESS6 Questionnaires GB" (you can download a PDF file of the questionnaire or read it with your browser's built-in PDF reader). Keep the questionnaire visible in order to answer the questions below.

Take a look at the introduction (page 3). Please write down your answers in the boxes situated below each question.

? The introduction seems to be structured around three elements. How would name each of them? Why do think the introduction is structured in this way?

Let's move on to "Section B", which is about politics and governance. Note that all questions are named first with a letter corresponding with the section they belong to (B in this case) followed by a number corresponding to the order of that question within its section.

? Looking at questions B11-B17, what do you think they are trying to measure here? Do you think the list of responses is comprehensive? Why/why not?

? Try to identify in the questionnaire the sociodemographic questions. In which section are they? What kind of characteristics do they ask about? Why do you think these questions are situated towards the end of the questionnaire rather than at the beginning?

? Go now to page 66. The questionnaire says that, in order to improve the questionnaire in the future, the interviewer will ask some final questions about similar topics that the interviewee has already responded to. How do you think this helps the designers of the questionnaire? Also, find the equivalent questions for IF13, IF14, and IF15, and jot down concisely what has changed.

The UKDS also allows you to run a search by topic or theme. Imagine you are interested in exploring political participation. The ESS provides a considerable range of questions on political participation, but you may be interested in seeing how different surveys approach it (this is a good way to critically engage with a topic). For this, let's go back to the UKDS webpage (`http://ukdataservice.ac.uk`) and in the search box situated to the right type "political participation" (make sure "Data" is selected and not "Website").

As you can see 822 results show up. You can sort them by relevance, date, title, and downloads. If you sort the results by "most downloaded data" you will see that the first result to appear (at the time of writing) is the **British Household Panel Survey (BHPS)**. The BHPS is a survey that, like the ESS, covers a broad range of different topics. This time we are interested only in political participation, so let's sort the results again by "relevance" and click on "Audit of Political Engagement 9, 2011" (APE9). Moreover, the UKDS website also has a search-box in which you can type variable names (keep in mind that in a survey questionnaire a question is an *operationalisation* of a variable). You can find this "variable search-box" at `http://discover.ukdataservice.ac.uk/variables`, but you can also get to it by clicking on the **Get Data** tab and then click again on the box situated to the right (which says "variable and question bank"). For the time being we are going to work with the Audit of Political Engagement 9, 2011 survey. Now try to answer the following questions:

? Get a copy of the questionnaire and jot down the main differences you see when you compare it to the ESS questionnaire (keep it open too).

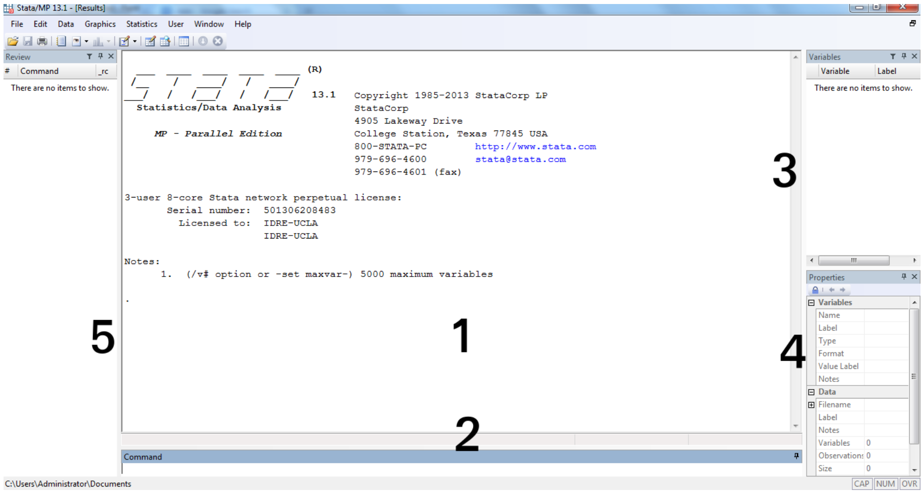
? Perhaps you have noticed that questions Q4A and Q5 (APE9 questionnaire) ask respondents about a series of things that they did (or did not) do in the last two or three years. On the other hand, the ESS asks only about actions performed (or not performed) in the last twelve months. Write in the box below whether you would choose one or the other and the reasons why.

? Take a look at questions Q5 (APE9) and B11-B17(ESS6). What other differences do you see in their phrasing? Which one would you choose and why? Also, would it make sense to operationalise political participation with a ranking scale? Why/why not?

2 Lab Worksheet II

Stata's interface

This is probably the first time you use Stata. As described in their own website, "Stata is a complete, integrated statistics package that provides everything you need for data analysis, data management, and graphics."



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   13.1   Copyright 1985-2013 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

3-user 8-core Stata network perpetual license:
       Serial number:  501306208483
         Licensed to:  IDRE-UCLA
                       IDRE-UCLA

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

.
```

Figure 1: Stata's interface.

Upon starting Stata you will see the following sections:

1. The results and outputs of your commands will be displayed in the **results window**.
2. The **command window** is where you want to type your commands and send them to Stata by pressing the Enter key.
3. If you have a dataset loaded its variables will appear in the **variables pane**.
4. You can browse the properties of variables in the **properties pane**.
5. The **review pane** lists all the actions you have taken in Stata.

At the moment you have no data loaded, that is the reason why your variables window is empty. Today you are going to input manually some data on UK universities. You can find these data in the table below.

Table 1: Some universities in the UK

Name	Students	Ranking	Location	Nobel laureate	Vice chancellor	Year
Edinburgh University	35258	20	Scotland	20	Male	1582
Napier University	17264	64	Scotland	0	Female	1992
University of Essex	11939	47	England	2	Male	1965
University of Cambridge	18271	1	England	91	Male	1209
Cardiff University	30180	27	Wales	2	Male	1997
University of Liverpool	21276	59	England	9	Female	1903
Queen's University Belfast	24955	45	Northern Ireland	3	Male	1908
Glasgow Caledonian	18410	89	Scotland	1	Female	1993
Ulster University	26200	82	Northern Ireland	0	Male	1982
University of Warwick	23420	6	England	1	Male	1965

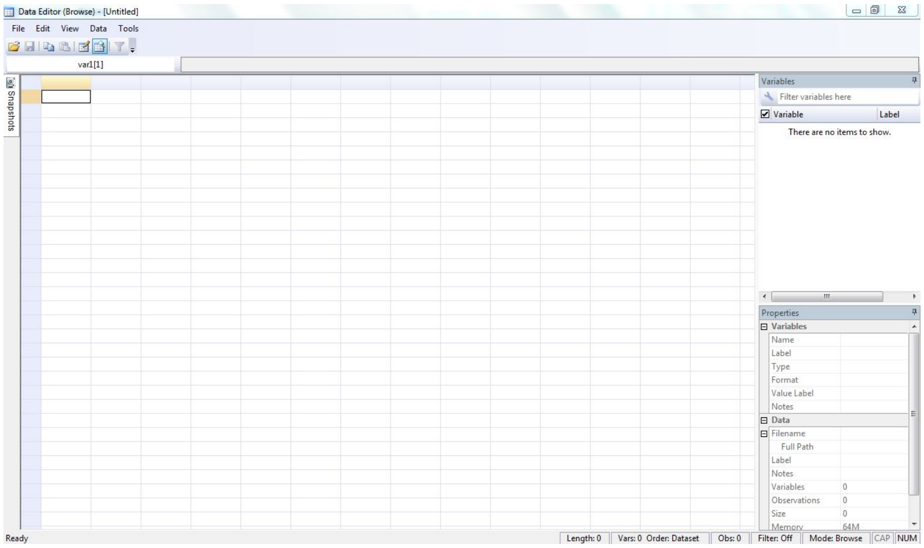
As you can see there are seven variables and ten observations. You need the following variables:

- Name of university.
- Total number of students (numeric).
- Position in the The Guardian Ranking.
- Geographical location (categorical, see below).
- Number of affiliated Nobel laureates (numeric).
- Gender of the Vice-chancellor (binary, see below).
- Year in which it was established (numeric).

The variable “Location” has the following four categories: Scotland = 1, Northern Ireland = 2, Wales = 3, and England = 4. The variable “Vice-chancellor” should be coded as Male = 0 and Female = 1.

You need to think of suitable **variable names**, **variable labels**, and **value labels**. Keep in mind that is customary, but not mandatory, to name variables with small letters.

Stata accepts both upper cases and lower cases, but it is case sensitive! Also, variable names cannot contain spaces and the use of underscores is customary when signalising spaces. After thinking of some variable names, open up the data editor by clicking on the icon with a pencil over a spreadsheet (see Figure 2). Stata's data editor resembles very much Excel. Your "spreadsheet" should be empty, but if it was not you could type `clear all` in the command window to get rid of previous data.



There are no items to show.

Name
Label
Type
Format
Value Label
Notes
Filename
Full Path
Label
Notes
Variables 0
Observations 0
Size 0
Memory 64M

Length: 0 | Vars: 0 Order: Dataset | Obs: 0 | Filter: Off | Mode: Browse

Figure 2: Stata's data editor.

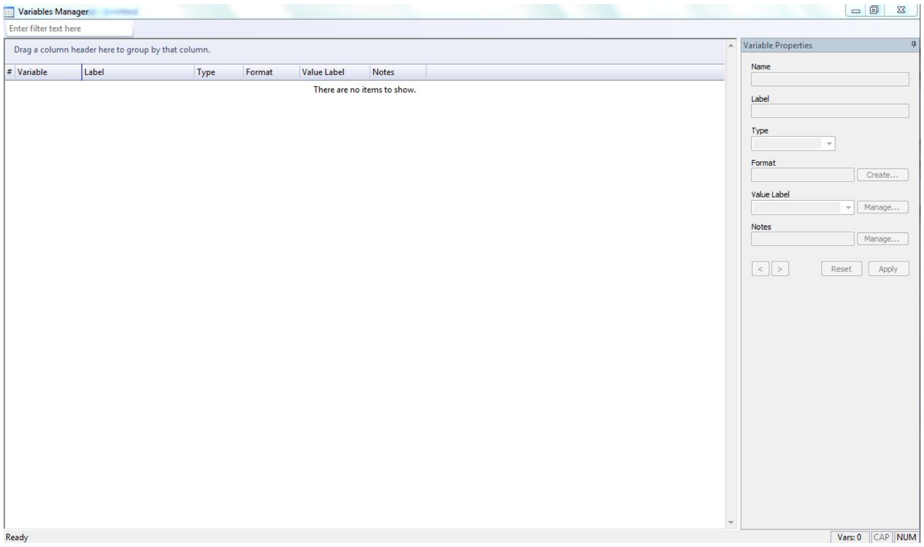
Now it is time to enter the data; start by typing in the first observation (University of Edinburgh). If you press the Tab key you will move onto the next cell. Keep entering the values for all variables for the University of Edinburgh (your first observation). Here it is important to remember that rows represent observations whereas columns represent variables. You might notice that Stata is naming your variables as *var1*, *var2*, and so on. To change **variable names** simply select a variable in the variables pane (top right corner) and change its name in the properties pane (bottom right corner). Repeat the process for the rest of variables. Let me remind you again that we need a number for geographical location and gender (the categories are specified above).

From the properties pane you can also define **variable labels**. The process is identical as before: select a variable in the variables pane, switch to the properties pane and give the variable a label (but this time using the box named "Label"). Variable labels allow spaces. Remember to be precise and accurate with your labelling, but try to keep your labels parsimonious (not too short as to make them ambiguous, but not too long either). Once you have named and labelled your variables you should see the new names and labels in the variables window.

Now that you have assigned names and labels to all your variables, you are going to assign **value labels** to two variables: geographical location and vice-chancellor (remember the above-mentioned specifications). But first try typing in the command window:

```
tabulate gender
```

In the command above, `tabulate` creates a tabulation of a variable named "gender." If you did not name your variable "gender", you will have to substitute in your variable's name instead of "gender." In the results window you can see now a table with two categories: 0 and 1. You know that 0 = Male and 1 = Female, but when you report these results you want your readers to know exactly what is going on. Changing 0 to "Male" and 1 to "Female" is what you do when you assign **value labels** to your vice-chancellor's gender variable.



There are no items to show.

Figure 3: Stata's variables manager.

To assign value labels you need to exit the data editor and open the *variables manager*. This can be accessed by clicking on the icon with a spreadsheet. Once the variables manager shows up (see Figure 3 for the generic, blank window) you need to create first a set of value labels which you will be attaching to a specific variable later on. Click on **Manage...** (situated on the right side of the window). In the new window click on **Create Label**. Yet a new window named **Create Label** will appear on screen. Give to your new set of value labels a name (in the **Label name** box). It is best if you name it with something that resembles the name of the variable you are modifying. For example, if you created a variable named "location" for the geographical location of universities you can name your set of value labels as "locationlab" or "locationl." Keep in mind that sets of value labels cannot use variable names. Now type in the box named **Value** the original value you gave to your variable (i.e., 0 = Male, so you would type "0" here) and in the **Label** box type its new label (i.e., "Male"). Then click on **Add** and repeat the process for 1 = Female. Once this is done click on **OK**.

Now that you have a set of value labels, the last thing you need to do is attach that set to the original variable. For this, go back to the variables manager window (Figure 3), select the variable at stake, and click on the down arrow located to the right of **Value Label**. A dropdown menu in which you can select the appropriate set of value labels will appear. Finally, click on **Apply**. If you use the `tabulate` command again you will see that your tabulation, this time, reports "Female" and "Male" instead of 1 and 0. Well done!

3 Lab Worksheet III

Preparing data

In this exercise you will use some real data. Download from LEARN the dataset named **essuk12.dta**, which corresponds to the sixth round of the European Social Survey. This dataset has been simplified for today's tutorial's sake and it comprises information on the UK in 2012. Open the dataset with Stata by clicking on **File→Open** and select **essuk12.dta**. You should now see some variables listed down in the *variables window* on the right side of your screen.

Today's task is to do some light data management, which entails:

- Exploring your dataset,
- renaming variables, and
- recoding variables.

Data exploration

The command `describe`, as its name indicates, describes data in memory or in file. Many commands in Stata can be abbreviated, thus `describe` can also be used as `de` (notice how I underlined the shortest abbreviation possible above). You can use this command to describe the entire dataset, or just a selection of variables. For example, if you typed

```
describe
```

you would be describing all variables in the dataset. If you wanted to describe only some variables you could type

```
de Tvtot Vote Happy
```

and describe only those three variables of interest. You can get more information on `describe` by using another command:

```
help describe
```

The `help` command opens Stata's manual and shows detailed information on the command you need help with. Remember to consult Stata's manual whenever you are not sure how to use a command. Now try to use `describe` yourself and answer these questions:

1. How many variables are there in the dataset?
2. How many observations does this dataset have?
3. What is the difference between "variable label" and "value label"?

Another useful command is `codebook` (if you type `help codebook` you will see that it cannot be abbreviated). This one also describes the contents of a dataset, however it provides you with a more detailed description of the data. Just like `describe`, you can use `codebook` on the whole dataset or on a selection of variables of interest. Try the following command and answer the questions:

```
codebook Tvtot Gndr Sclmeet
```

1. How many response categories does "Sclmeet" have?
2. How is "Gndr" coded?
3. What does "Tvtot" measure?

Renaming variables

When you download a dataset, variables' names not always make sense. Consider the variable "Sclmeet"; if you use `describe` and `codebook` on this variable you can find what it measures. Let's rename the variable to something else, for example "Socmeet" (mind the capital -S). To achieve this type the following:

```
rename Sclmeet Socmeet
```

You can use `help rename` to find out more about the command `rename` (for instance, it can be abbreviated to `ren`). Now try to rename the variables in your dataset to the following new names (mind all the capital letters):

- Gndr to Gender.
- Agea to Age.
- Eisced to Education.
- Hinctnta to Houseincome.

Although not mandatory, it is common to specify variable names with lower cases. Right now all your variables start with a capital letter, so if you wanted to change their names you could use the command `rename` on each variable and change the first letter only. However, this would take quite a while as you would have to repeat the command twelve times. Stata has a way to deal with problems like this one. Instead of typing the same command twelve times try:

```
re _all , lower
```

Remember that `re` stands for `rename`. Above you used `_all` to tell Stata that you would like to modify all variables in the dataset. Notice that now there is a comma in the command. That comma tells Stata that you are going to use a *command option* (pretty much all commands in Stata allow for options, you can find more about these options in the help page of each command). The option `lower` tells Stata that you want only lower cases.

Recoding variables

Recoding variables is a fundamental part of analysing data. The command `recode`, as its name indicates, recodes variables in your dataset. Use `help` to find out more about its syntax. Let's now try to recode the variable "gender", remember how it is coded (you can find out by using `codebook`). You want to get the following:

- 0 = Male
- 1 = Female

To achieve this you need to type:

```
recode gender (1=0) (2=1), generate(gender2)
```

The command above is telling Stata to recode the variable "gender" so the value 1 is now 0 (1=0), and the value 2 is now 1 (2=1). In short, the "old value" comes first and then you tell Stata the "new value." However, there is one more thing you need to consider. Recoding variables modifies the data, so it is always a good idea to keep intact the original variable (in this case "gender"). The command `recode` allows as a command option another command, in this case the command `generate` (which can be abbreviated to `g`, but people usually use it as `gen`). As you can imagine, `generate` creates a new variable in your dataset. Above, you recoded the variable "gender" and, after recoding it, you created a new variable named "gender2" (mind the round brackets).

Now try yourself to recode the variable "vote" to (do not forget to generate a new variable):

- 1 = Yes
- 0 = No
- -9 = Everything else

Remember that you can use `codebook` to find out how a variable is currently coded. After recoding the variable you can run `codebook` again to see how your new codes look like. You will notice that the new coding does not have appropriate value labels, but luckily you know now how to create and assign value labels!

4 Lab Worksheet IV

Univariate analysis and graphing

Today we are going to use data from the ESS 2012 again. You can find the dataset on LEARN; make sure you download **essuk12v2.dta** (notice the "v2"). In this lab session we are going to do some basic descriptive statistics, including tabulating and graphing. One of the simplest things in data analysis is reporting descriptive statistics and simple graphs that show a single variable. This is what we call **univariate analysis**, and it is very useful when trying to summarise and describe data (which in in turn helps finding patterns in the data). In Stata there are multiple ways to look at descriptive statistics. Last week you were introduced to two useful commands: `describe` and `codebook`. Today you will use one more: `summarize`. Let's take a look at it, try typing

```
summarize
```

and you will see a list of summary statistics for all the variables present in your dataset. This command can be shortened to `su`, and it allows for a selection of variables like this:

```
su gender tvtot
```

In the example above the command `summarize` only reports on two variables ("gender" and "tvtot"). Notice that the command is `summarize` and not `summarise`. This command, as you can see, tells you the number of observations, the mean of the variable, the standard deviation, and the minimum and maximum values a variables takes on. Let's now revise the command `codebook` with an extra option (remember that `codebook` cannot be shortened):

```
codebook gender tvtot , compact
```

As you can see, `codebook, compact` provides you with similar information (but not the same). In this case you get variable labels and unique values. Also notice that the command option follows a comma, which tells Stata that you have decided to use an option. You can try typing the following to see what the option `compact` really does:

```
codebook gender tvtot
```

The option `compact` presents you with a shorter report, which you might want if presenting summary statistics to someone else. Remember that you also have the `describe` command:

```
de gender tvtot
```

Just like the other two commands, you can use `describe` on the entire dataset or on a selection of variables. The important thing to notice is that these three commands report similar yet different information. Now give these commands a try and find the following information:

- The variable labels for all variables in the dataset.
- The mean of the variable “age.”
- The mode of the variable “tvtot.”
- The minimum and the maximum values of the variable “socmeet.”

After completing the exercise above you might have noticed that the maximum value of “socmeet” is oddly high. The answer has to do with response categories. For example, the ESS codes “Don’t know” as “88” and “Not available” as “999.” This coding allows you identify these categories easily across variables and treat them as **missing data**. Different datasets will code these categories in different ways, but it is important to know how these categories are coded before doing data analysis.

Tabulating data

The command `tabulate`, which you can shorten to `tab`, creates one-way or two-way tables of frequencies. The command is followed by one or two variables of your interest, for instance:

```
tabulate socmeet
```

If you were interested in seeing the codes of the response categories you could add a command option to `tabulate` like in the example below (remember that command options follow a comma):

```
tab socmeet, nolabel
```

The option `nolabel` removes value labels from the tabulation, and it can be shortened to `nol`. This option could be useful to see at a glance how a variable is coded (but remember that `codebook` also shows you this information). Now let’s change “Don’t know” to something else that Stata understands as **missing values**. The command `replace` will be handy here:

```
replace socmeet=.a if socmeet==88
```

As you can imagine, `replace` replaces the content of an existing variable, in this case “socmeet.” Notice that in the command above there is an `if` qualifier. This qualifier is telling Stata to perform an action only on those cases that meet the qualification, which in this case is “when the variable “socmeet” equals 88.” In other words, the command above is saying something like “for those observations in which the variable “socmeet” is “88”, replace “socmeet’s” value “88” to “.a”. The baseline dot before the “a” is of paramount importance because that is how Stata recognises missing values. Also notice that Stata distinguishes between “=” and “==” (like many other programming languages). The former (“=”) is used as a set equal operator, whereas the latter ("==") is used to test for equality. You can find more about this topic in the "Stata Cheatsheet" on LEARN. Now that you have coded the missing values in the variable "socmeet" let's tabulate it again:

```
        tab socmeet
```

If you compare this new tabulation to the first one, you will notice that three cases are now missing (i.e., the three cases who responded "Don't know"). You can check how many missing cases a variable has using the `missing` option like this:

```
        tab socmeet, missing
```

? Now try all this by yourself. Let's look for missing values in the variables "age", "gender", and "houseincome." If you find missing values change them to something like ".a" or ".b". Do not forget to use command options like `nolabel` and `missing`.

The `if` qualifier can also be used in the command `tabulate`. For example, say you want to tabulate the gender of those respondents older than 30:

```
        tab gender if age > 30
```

Notice that the variable-target of `tabulate` is "gender" whereas the target of the qualifier `if` is "age." You can be even more specific with your tabulations, for example:

```
        tab gender if age > 30 & age < 50
```

Notice that after `&` ("and") you need to specify again the target-variable (in this case "age"). You can use as many variables as you want. Say you wish to know the gender distribution of those of 20 years of age **or more** and who voted in the last national election:

```
        tab gender if vote==1 & age >= 20
```

? Try now to find out how many observations are below the fifth decile in the variable "hoseincome."

Graphing data

Producing graphs is an extremely powerful (and useful) way of examining your data. Stata allows you to produce a great variety of graphs, as well as it allows you to customise them. Let's start by producing a pie chart of the average number of hours watching TV (variable "tvtot"). Using the menu, go to **Graphics > Pie chart**. Make sure *Graph by categories* is selected, and where it says *Category variable* select "tvtot." Now click on **Submit** and you will get a pie chart of the variable "tvtot." In your results-window you can see the command for this action:

```
 graph pie, over(tvtot)
```

Let's make the plot a bit more attractive. Go to the **Slices** tab and select *Customize all slice properties*. Then click on *Slice properties (all)*, tick *Explode slide* and in the drop-down menu select *Automatic*. In the same window, select *Customize all slice labels* and click on *Label properties (all)*. In *Label type* select *Percent*. Now click on *OK* and submit your graph. Stata reports the command for all this clicking and pointing in the results-window. You can use that command to reproduce the exact same graph granted that the variable "tvtot" remains the same.

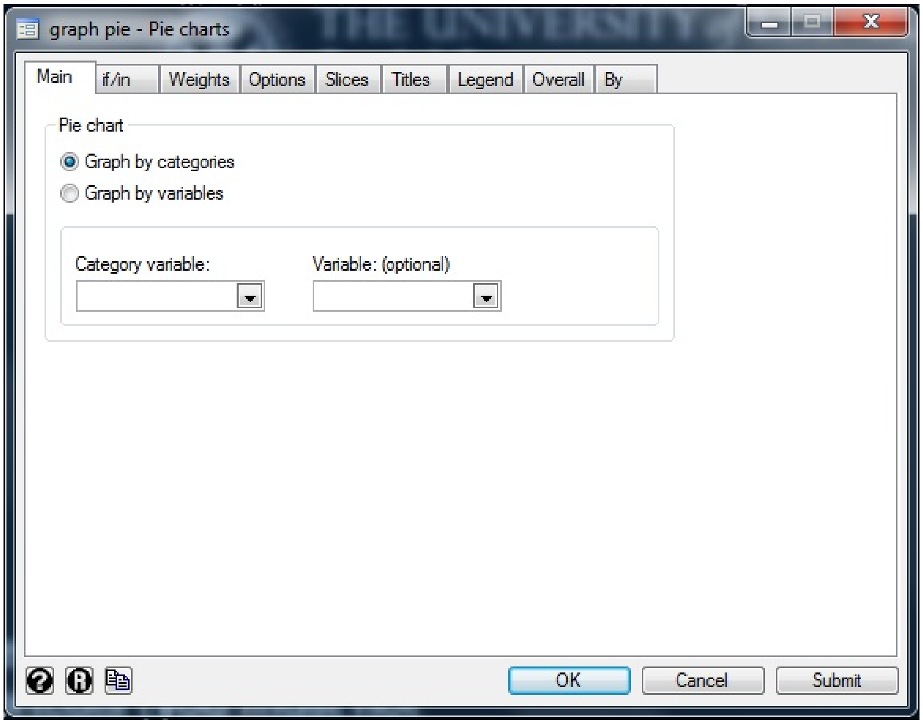


Figure 1: Stata's Graphics > Pie chart.

5 Lab Worksheet V

Bivariate analysis

Today you will learn how to examine the empirical relationship between two variables in Stata. This examination can be both *descriptive* and *inferential*. There is a number of different ways by which you can achieve this, for instance, cross-tabulations, scatterplots, correlation coefficients, and linear regression. However, whether bivariate analysis is descriptive or inferential will ultimately depend on two things: one is the level of measurement of your variables, and the other is whether you can consider one variable "*dependent*" and the other "*independent*." In this exercise we are going to work with categorical variables in the **essuk12v3.dta** dataset which you can download from LEARN (please mind the "v3" in the dataset's name).

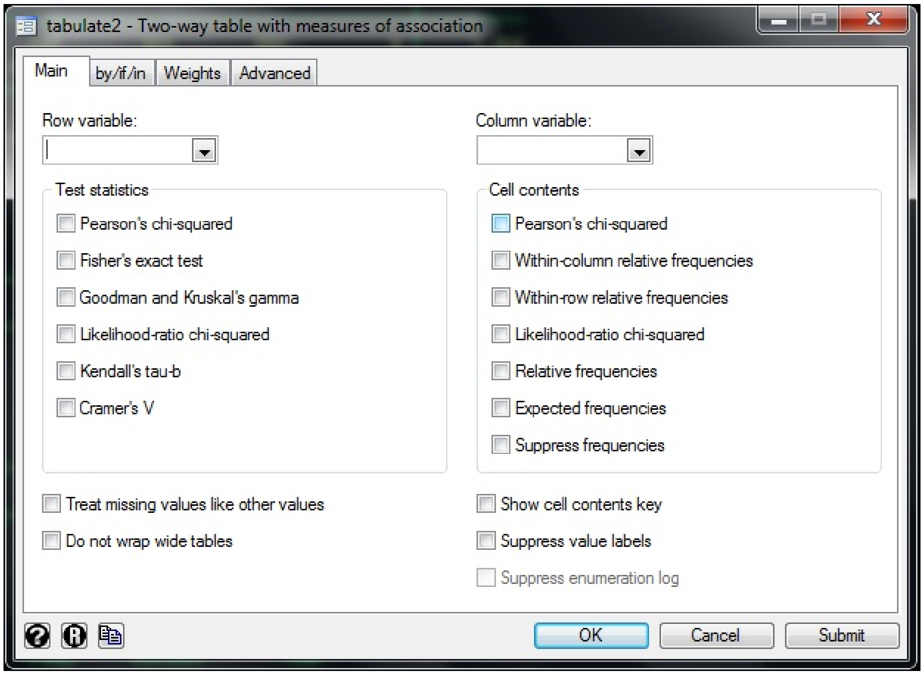
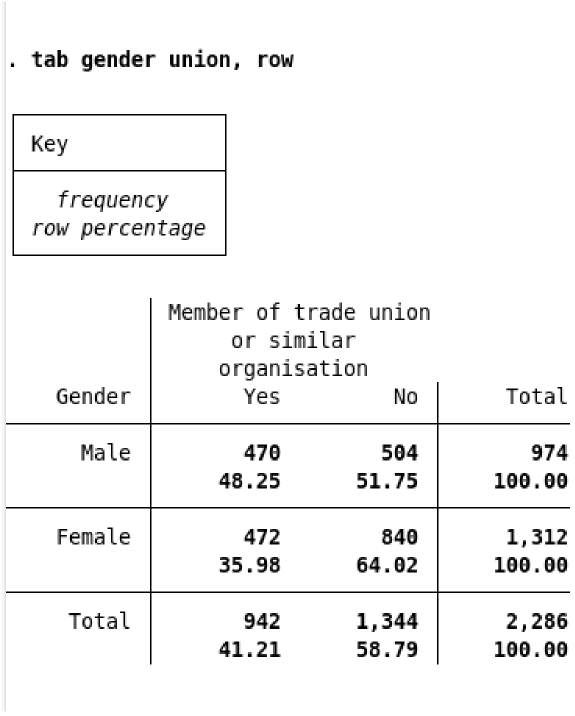


Figure 1: Two-way table with measures of association.

Cross-tabulations

A *cross-tabulation* is a table with rows and columns in which categorical variables (their frequencies) are represented. These tables are also called "*contingency tables*" because the category of a case in one variable is *contingent* on its category in the other variable. For example, the political party a person supports may be contingent on their socio-economic status. If you can identify dependent and independent variables, it is customary to place your dependent variable in columns and your independent variable in rows. Let's start with a cross-tabulation of whether a person has ever been member of a trade union and their gender. In this example, we are going to assume that whether a person has ever been member of a trade union is more likely if the person is a man (remember that assumptions are given by social theory; numbers are useless without theory!). Here "gender" is our independent variable and "trade union membership" is our dependent variable.

To get a cross-tabulation of the variables "gender" and "trade union membership" we can use the command `tabulate`, but this time we are going to feed in two variables. Alternatively, you can use Stata's menu: go to *Statistics* > *Summaries, tables, and tests* > *Frequency tables* > *Two-way table with measures of association*. You should now see something like Figure 1 above. Select "gender" as your *Row variable* and "union" as your *Column variable.*Under *Cell contents* tick *Within-row relative frequencies*, this will give you percentages that add up to 100 percent for each row.



```
. tab gender union, row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
+-------------------+

           | Member of trade union
           |     or similar
           |    organisation
    Gender |       Yes         No |     Total
-----------+----------------------+----------
      Male |       470        504 |       974 
           |     48.25      51.75 |    100.00 
-----------+----------------------+----------
    Female |       472        840 |     1,312 
           |     35.98      64.02 |    100.00 
-----------+----------------------+----------
     Total |       942      1,344 |     2,286 
           |     41.21      58.79 |    100.00 
```

Figure 2: Cross-tabulation of gender and union membership.

Remember that `tabulate` can be shortened to `tab`. The column on the right provides you with the total for each row. The first numbers in each cell of the table are frequencies, whereas the numbers below are percentages. As you can see in Figure 2, there are 974 males in total, 470 of them have ever been member of a trade union. On the other hand, there are 1,312 females in total and 472 of them have ever been member of a trade union. Because each row (i.e., each category of "gender") has a different number of observations, it is not easy to compare *absolute frequencies*. It appears that more females have ever been in a trade union, however this is not true when we look at it in *relative terms*. Percentages are greatly useful in cases like this one: in absolute terms, this dataset shows more females that have ever been member of a trade union, but when taking into consideration the number of males and females, it appears that more males, in relative terms, have ever been member of a trade union (48.25 percent versus 35.98 percent). Another important thing to keep in mind is the "direction" of the percentages; in this example the proportions have been computed in rows (horizontally), thus you should carry your comparison in columns (vertically).

Chi-squared test (χ^2)

The difference between men and women who have ever been member of a trade union is not very big, but it might be considerable (remember, 48.25 percent versus 35.98 percent). Did we get this difference by *chance*? The sample you are working with is pretty big (2,286 cases), so you may think that there is no reason to think that the difference between men and women is given by chance. Nevertheless, it is good practice to check whether it is *statistically significant*. We can use the chi-squared statistic (χ^2) for examining the likelihood that your results occurred by chance. In simple words, the logic of this test is that if it is extremely unlikely to get this difference between men and women, with such a big sample size, then you can be confident that there was a real difference. However, do not confuse this with the researcher's call to assess whether the difference between men and women is substantially (i.e., theoretically) important.

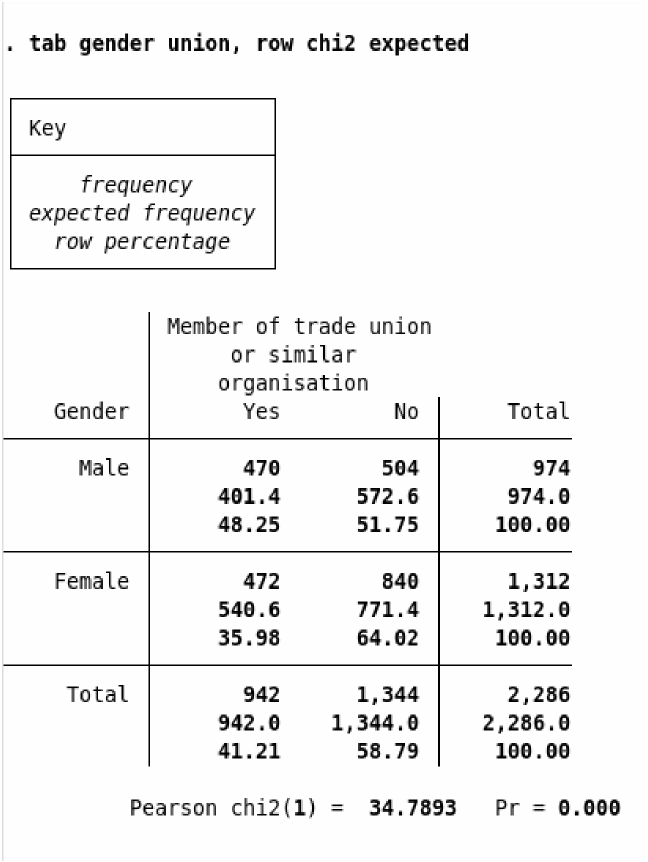
To perform a chi-squared test in Stata you can use the menu shown in Figure 1, but this time tick *Person's chi-squared* under *Test statistics* and, additionally, you can also tick *Expected frequencies* under *Cell contents*. Alternatively, you can use the following command:

```
tabulate gender union, row chi2 expected
```

The option `row` gives you percentages in rows, whereas `chi2` performs the test and `expected` reports the expected frequencies. These expected frequencies are the number of observations expected if there were no relationship between the two variables in the table. Figure 3 below shows the chi-squared test: you can see that there are 470 men who have ever been member of a trade union (or 48.25 percent of men), but now you also know that you would expect to have 401.4 men here by chance. On the other hand, 472 women have ever been member of a trade union (35.98 percent of women), but you would expect to have 540.6 if the variables were truly independent. In other words, this means that there are 470–401.4 = 68.6 more men that have been member of a trade union than you would expect. At the bottom in Figure 3 you can see the value of Pearson's chi-squared, which is 34.7893 (the number in parenthesis is the degrees of freedom, which is 1 in a 2 by 2 table). The formal way to communicate the results of this test is:

$\chi^2(1, N = 2,286) = 34.7893; p < 0.001$

Because $p < 0.001$ (i.e., statistically significant) you can reject the null hypothesis that the two variables are independent. In other words, the data suggest that "gender" and "trade union membership" are not totally independent. Note that in Figure 3 Stata reports $p = 0.000$. However, this does not mean that the p-value is 0. Stata, in fact, reports an estimate of the probability to three decimal places. Say we got $p = 0.00023$; Stata would round this figure to $p = 0.000$. It is worth-noting that the value $p = 0.00023$ means that you would find this distribution of data only 0.023 percent of the time if the distribution was due to chance.



```
. tab gender union, row chi2 expected

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
+--------------------+

           | Member of trade union
           |     or similar
           |    organisation
    Gender |       Yes         No |     Total
-----------+----------------------+----------
      Male |       470        504 |       974
           |     401.4      572.6 |     974.0
           |     48.25      51.75 |    100.00
-----------+----------------------+----------
    Female |       472        840 |     1,312
           |     540.6      771.4 |   1,312.0
           |     35.98      64.02 |    100.00
-----------+----------------------+----------
     Total |       942      1,344 |     2,286
           |     942.0    1,344.0 |   2,286.0
           |     41.21      58.79 |    100.00

          Pearson chi2(1) =  34.7893   Pr = 0.000
```

Figure 3: Cross-tabulation of gender and union membership.

How does the chi-squared test work? (In simple words)

Pearson's chi-squared test applies to categorical data, and it evaluates how likely it is that an observed distribution of data is due to chance. Sometimes it is also called "goodness of fit" statistic because it evaluates how well the observed distribution of data fits the expected distribution if there were no relationship between the variables. Chi-squared is used to test the likelihood that your results *occurred by chance*. This is the reason why we can think of Pearson's chi-squared test as a means to test the null hypothesis that the variables are independent. The test compares the frequency of each cell with the frequency one would expect to obtain were the variables independent. The expected distribution is given by the number of observations in the table's rows and columns.

Measures of association

Measures of association summarise, with a number, the strength of a relationship. Pearson's chi-squared test is useful in many contexts, but it is not always the best option. A chi-squared test depends greatly on the sample size; this means that the bigger the sample, the bigger chi-squared will be (and you may potentially misinterpret a chi-squared value in big samples, thus concluding that there is a strong relationship between two variables when, in fact, it may be rather weak). To solve this issue, we have two main alternatives: the coefficient phi (ϕ), and Cramer's V. The way these two work is by simply dividing chi-squared by the maximum value chi-squared itself can be for a given table and a given number of observations (so for a 2 by 2 table this is N, and to compute phi we simply calculate the square root of that division. Many people interpret phi as follows:

- $0 \geq \phi \leq 0.19$ = weak.
- $0.2 \geq \phi \leq 0.49$ = moderate.
- $\phi \geq 0.5$ = strong.

One thing to keep in mind is that a chi-squared test tells us about statistical significance, whereas measures of association like phi or Cramer's V tell us about the strength of the relationship (that is why the value of phi or V does not change with the sample size). Because phi is used in 2 by 2 tables (and Cramer's V is used for bigger tables), Stata always calculates Cramer's V (you could say one is a special case of the other). To calculate Cramer's V for "gender" and "union membership" we use the the command option `V` as shown below:

```
tabulate gender union, chi2 V
```

The result of the command above shows a phi of 0.1234 (remember that phi is for 2 by 2 tables like ours). You can see now that even though chi-squared was statistically significant, phi tells us that the relationship between "gender" and "trade union membership" is rather weak! Another thing to note is that phi might take on negative values (always from -1 to 1) and Cramer's V goes from 0 to 1.

6 Lab Worksheet VI

Statistical inference I

This lab worksheet walks you through the very basic concepts of statistical inference (no need for Stata today, pencil and paper will do). For the purposes of this exercise, it will be very helpful if you already know a bit about the following concepts:

- Population.
- Sample.
- Sample statistic.
- Sampling distribution.
- Central Limit Theorem.
- Normal distribution.

The main learning goals is to understand what inferential statistic is, but we will also look at how we can collect random samples from larger populations as well as how we go about constructing sampling distributions using repeated random samples. Figure 1 below shows the basic "statistical procedure." As you can see, it has three main components: a **population** or target group of interest, a **sample** or sub-group of the target group, and **probability**, which allows us to say things about the population based on a sample. After collecting a sample, we produce some summaries (i.e., we describe and explore the data). However, this does not really speak of the population but of the observed data in a sample. These observed data might differ from the population, so we rely on probability and inference to draw conclusions. Below Figure 1 there are some questions for you to answer.

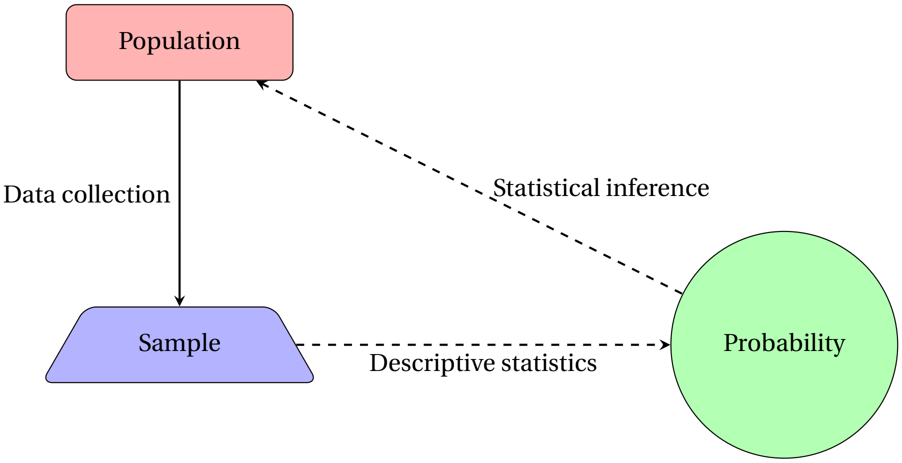


Figure 1: Doing statistics.

? Provide some examples of populations and samples in the box below.

? Might the following samples (simple random sampling) be representative?

Target population:
Lawyers in Greater London area.
Sampling and sample:
All lawyers working in the Greater London area are assigned an X-digit identification number. 500 lawyers are selected at random by using a random-number generator.
How this sample could have been non-representative:
This samples would not be representative if only lawyers with a British passport had been selected (presumably there are lawyers with non-British passports working in the Greater London area).

Target population:
University students studying in Scotland.
Sampling and sample:
Using Scottish universities' student records, a random sample of 1,000 students was pooled. This sample was designed to comprise students older than eighteen and younger or equal than twenty-four. All nationalities, ethnic groups, and sexes were included.
Is this sample representative? Why/why not? How could it be representative?

Target population:
University students studying in Scotland.
Sampling and sample:
50 interviewers were hired in order to interview university students from all Scottish universities. Interviewers were sent to the field only on odd working calendar-days. Data collection was ended after obtaining 1,000 completed questionnaires.
Is this sample representative? Why/why not? How could it be representative?

Target population:
University students studying in Scotland.
Sampling and sample:
Researchers gather a comprehensive list of university students in Scotland. Students are grouped into "male" and"female" categories. Conveniently, 60 percent of university students in Scotland are female. Therefore, the researchers select at random 600 female students and 400 male students.
Is this sample representative? Why/why not? How could it be representative?

Next we are going to construct a sample using simple random sampling. For this, we will use real data from the European Social Survey 2012, but this time we will only focus on respondents from Scotland. On the last page of this worksheet you will find a table with 100 self-reported happiness levels. Happiness was measured from 0 to 10 (being 10 the happiest one can be). Our task is to construct a sample of 15 individuals, so:

? What is the value of *N*? And the value of *n*?

? How would you chose a random sample of 15 individuals from *Table 1*?

One "manual" way to construct a random sample involves using a table of random numbers. The numbers in the table below were placed in a random order by a computer, so technically is not purely random (but it is good enough for our exercise). We need to use this table of random numbers to chose 15 individuals from *Table 1* that reports happiness levels in Scotland. First we must devise a "path" or "method", for example: "two to the right, one down, one left, one up, two left, two up, and repeat" (it can be easier, of course). Now we place a finger randomly on the table and simply follow our "path." The figure we end up on will be our first sampled individual. Finally we repeat this procedure as many times as necessary.

? Annotate below the 15 respondents' ID and their happiness levels.

056	034	074	055	037	064	084	051	050	026	026	078	095	019	073	011	071	017	052	075	052
047	035	071	065	009	077	098	009	002	024	057	059	100	003	021	067	030	087	017	079	091
028	044	024	085	075	077	073	083	054	041	054	013	064	036	067	074	080	064	012	068	027
096	088	066	005	087	069	096	016	083	082	057	031	083	026	050	078	042	076	049	057	006
084	079	067	002	096	040	082	030	041	033	056	062	090	032	034	053	072	062	018	048	084
023	060	049	029	090	007	008	005	015	086	072	086	044	069	068	099	006	011	095	043	073
058	028	093	097	037	092	001	027	047	088	014	089	063	015	058	056	010	074	007	080	088
038	089	010	099	034	028	072	014	061	046	038	061	094	022	063	040	058	003	093	050	053
016	055	065	081	060	021	039	040	037	046	091	004	018	076	001	100	031	038	043	027	075
004	090	059	025	029	069	024	032	059	052	019	045	094	094	047	063	087	041	079	039	085
020	070	020	042	030	065	033	077	046	066	078	070	093	025	054	068	071	089	035	025	081
085	048	060	097	012	092	053	044	045	042	051	022	036	049	081	033	031	062	043	048	032
080	036	095	091	056	034	074	055	037	064	084	051	050	026	026	078	095	019	073	011	071
017	052	075	052	047	035	071	065	009	077	098	009	002	024	057	059	100	003	021	067	030
087	017	079	091	028	044	024	085	075	077	073	083	054	041	054	013	064	036	067	074	080
064	012	068	027	096	088	066	005	087	069	096	016	083	082	057	031	083	026	050	078	042
076	049	057	006	084	079	067	002	096	040	082	030	041	033	056	062	090	032	034	053	072
062	018	048	084	023	060	049	029	090	007	008	005	015	086	072	086	044	069	068	099	006
011	095	043	073	058	028	093	097	037	092	001	027	047	088	014	089	063	015	058	056	010
074	007	080	088	038	089	010	099	034	028	072	014	061	046	038	061	094	022	063	040	058
003	093	050	053	016	055	065	081	060	021	039	040	037	046	091	004	018	076	001	100	031
038	043	027	075	004	090	059	025	029	069	024	032	059	052	019	045	094	094	047	063	087
041	079	039	085	020	070	020	042	030	065	033	077	046	066	078	070	093	025	054	068	071
089	035	025	081	085	048	060	097	012	092	053	044	045	042	051	022	036	049	081	033	031
062	043	048	032	080	036	095	091	056	034	074	055	064	091	011	077	077	052	052	005	021
045	099	037	097	044	078	001	079	074	062	097	091	035	004	024	036	028	051	083	085	027

We could have calculated the average happiness level using all 100 respondents. However, this would take us more time (although it is easily doable) than sampling 15 individuals and calculating the average happiness. The formula below is how you calculate a mean (use the box below for your computation):

$$\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

With the arithmetic mean you can calculate now the standard deviation as follow:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

In turn, with the standard deviation we can calculate the standard error:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

Finally we can use the arithmetic mean to create the sampling distribution (of the mean). If we were to repeat the process above 10 times (i.e., if we constructed 10 samples of 15 individuals each) and graph their means (10 means) we would get a graph that might not look "normally distributed." However, if we repeated this 1,000 times the graph will look "more normal" (and the more samples you draw, the "more normal" your distribution would look like). However, remember that we created a sample of 15 individuals; we probably need a bigger sample size (about 25) to obtain a "normal" distribution. There are online applications that recreate this process, but if you have time you can try it yourself "manually."

Table 2: Happiness in Scotland, 2012

ID	**Happiness**	**ID**	**Happiness**	**ID**	**Happiness**
01	8	**41**	7	**81**	8
02	9	**42**	8	**82**	7
03	6	**43**	9	**83**	8
04	7	**44**	10	**84**	9
05	5	**45**	8	**85**	6
06	10	**46**	5	**86**	7
07	9	**47**	9	**87**	6
08	8	**48**	6	**88**	7
09	8	**49**	8	**89**	10
10	1	**50**	9	**90**	10
11	7	**51**	7	**91**	9
12	4	**52**	8	**92**	7
13	8	**53**	10	**93**	7
14	7	**54**	9	**94**	10
15	8	**55**	8	**95**	9
16	10	**56**	8	**96**	8
17	8	**57**	9	**97**	8
18	8	**58**	9	**98**	5
19	9	**59**	8	**99**	8
20	8	**60**	7	**100**	8
21	9	**61**	9		
22	9	**62**	10		
23	8	**63**	7		
24	8	**64**	8		
25	8	**65**	9		
26	8	**66**	10		
27	9	**67**	9		
28	10	**68**	10		
29	8	**69**	9		
30	9	**70**	8		
31	7	**71**	6		
32	9	**72**	10		
33	8	**73**	10		
34	9	**74**	8		
35	4	**75**	10		
36	8	**76**	10		
37	7	**77**	8		
38	9	**78**	8		
39	6	**79**	9		
40	7	**80**	5		

7 Lab Worksheet VII

Statistical inference II

This worksheet revises the concepts that you already know about statistical inference (if you need a reminder check *Lab Worksheet VI: Statistical inference I* and the pertinent lecture-slides). You will not need to use Stata for the exercises below. All exercises start with a question mark.

? When is a mean (average) a statistic, and when is it a parameter?

? If you were constructing a simple random sample from a comprehensive population of 10,000 individuals, how would be your sample's mean distribution **in terms of normality** if you were to select 50, 1,000, or 2,000 individuals from that population?

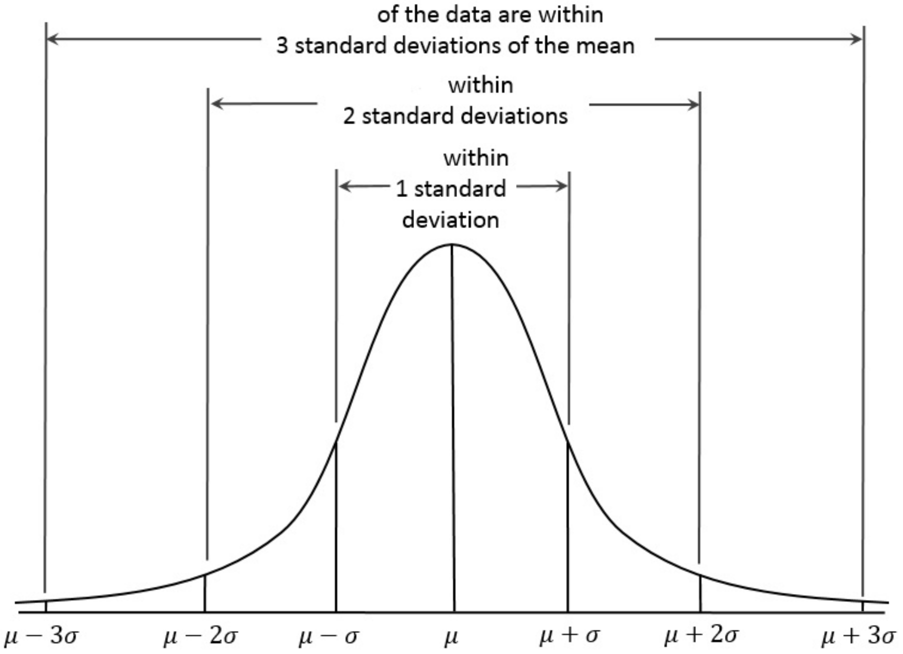
? Can you name the following formulas?

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

$$\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

? Can you complete the blank spaces in the figure below?



? How is the distribution depicted above called? What theorem makes it useful for data analysis?

? A simple random sample of students has a mean age of 25 and *n*=500. Can you calculate the standard error with the information provided? If not, what are you missing?

? A random sample of 100 students did an IQ test. The resulting average IQ was of 106 with a standard deviation of 16. Can you find a 95 percent confidence interval? (Use the formula below).

$$\overline{x} \pm Z_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

You can use the above formula to compute any confidence interval. The suffix α (alpha) denotes the confidence level. $Z_{\alpha/2}$ is the critical level for a given alpha. Remember that a critical level of a 95 percent confidence interval is about 1.96. For a 95 percent confidence interval the value of alpha is of 0.05 (because $1-0.95 = 0.05$). Using this information try to calculate the confidence interval for the exercise above.

? Can you complete the table below?

$Z_{\alpha/2}$	α	$1-\alpha$
1.645	0.1	
1.96	0.05	
2.576		0.99

? The coffee machine of your university's cafeteria is set up in such way that the average content of a cup of coffee equals a quantity μ. Who knows why but you decided to construct a random sample of 200 cups of coffee. This resulted in an average content of 34cl. Assume your standard deviation was of 6cl. Can you find two confidence intervals one for 90 percent and another for 99 percent?

8 Lab Worksheet VIII

Simple regression analysis

This lab worksheet guides you through the basics of doing simple regression analysis in Stata. The exercise does not really teach you the theory behind regression analysis, instead it focuses on the technical aspect of doing with Stata what you learned in your lectures. If you feel you need a reminder you can look up in the course's materials what regression analysis, regression coefficients, and R^2 are. It is also fundamental to know the difference between dependent and independent variables (the latter are also called *predictors* or *regressors*).

The data we are going to use come from the European Social Survey Round 8 (2016) in Great Britain. The dataset is called **essuk16** and you can find it on LEARN. As we already know, the first step when dealing with new data is to explore them. Use the following commands to find out more about this dataset:

```
describe
summarize
codebook, compact
```

These commands are helpful because they tell us things like what the dataset's variable labels mean. Tabulations are useful when exploring data too. Let's see how the data are distributed (remember you can shorten most Stata commands):

```
tabulate happy
tabulate trstprt
tabulate trstplt
tabulate atchctr
```

From the above we want to see things like "a happiness of 8 out of 10 seems to be the most frequent (about 30 percent)" or that "about 60 percent of the British population trust their political parties 4 out of 10 or less." We can obtain information from individual variables by using tabulations, but we could also graph the same data:

```
graph bar (percent), over(trstprt, label(angle(45)))
```

The way we explore new datasets partially depends on personal preference, however it ultimately comes down to "what we need to do" with the data (perhaps a barplot is more useful if we need to communicate these data to an audience that is not very familiar with tabulations).

Now that you know a bit more about this new dataset, we are going to perform a simple regression analysis. For the sake of this exercise we are going to assume that the following hypothesis is true: *"How happy a person is impacts on their trust in political parties."* We should also consider a few more things:

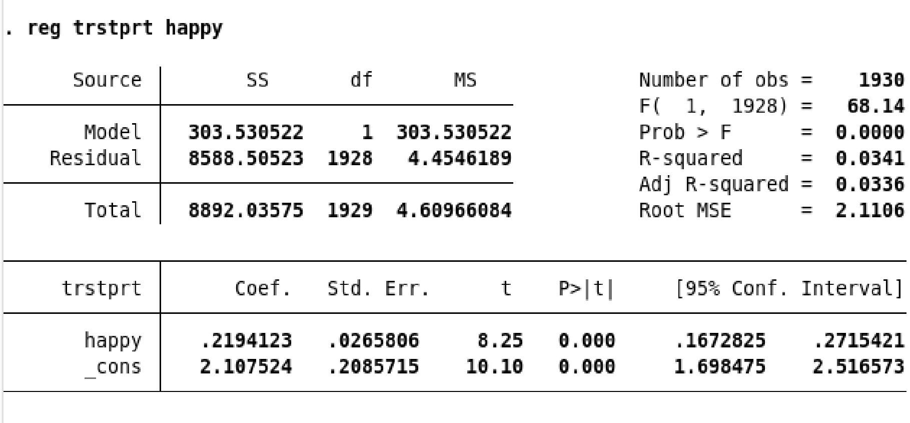
- If our theory poses that the happier a person is the more trust they will have in political parties, we can expect a positive regression coefficient.
- If our theory poses that the happier a person is the less trust they will have in political parties, we can expect a negative regression coefficient.
- To verify the points above we need to know how our variables are coded (i.e., “do they go from least to most?”)

We should acknowledge a final note; the two variables we are going to use in our regression analysis are scales, which technically are not continuous but ordered-categorical. Some authors argue that these data can be used in linear regression insofar some assumptions are met. These assumptions are well beyond the scope of this exercise, so we are going to ignore them for simplicity’s sake.

To run a simple regression model we type the following in Stata:

```
regress trstprt happy
```

The command `regress` can be shortened to `reg`. Remember that you can find out more about any Stata command with the command `help`. The first variable in our regression command represents our *dependent variable*, whilst the rest of variables (in this case there is only one) are our *independent variables*. In other words, we are assessing the impact of happiness on trust in political parties. Figure 1 below shows the regression output.



```
. reg trstprt happy

      Source |       SS       df       MS              Number of obs =    1930
-------------+------------------------------           F(  1,  1928) =   68.14
       Model |  303.530522     1  303.530522           Prob > F      =  0.0000
    Residual |  8588.50523  1928   4.4546189           R-squared     =  0.0341
-------------+------------------------------           Adj R-squared =  0.0336
       Total |  8892.03575  1929  4.60966084           Root MSE      =  2.1106

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       happy |   .2194123   .0265806     8.25   0.000     .1672825    .2715421
       _cons |   2.107524   .2085715    10.10   0.000     1.698475    2.516573
------------------------------------------------------------------------------
```

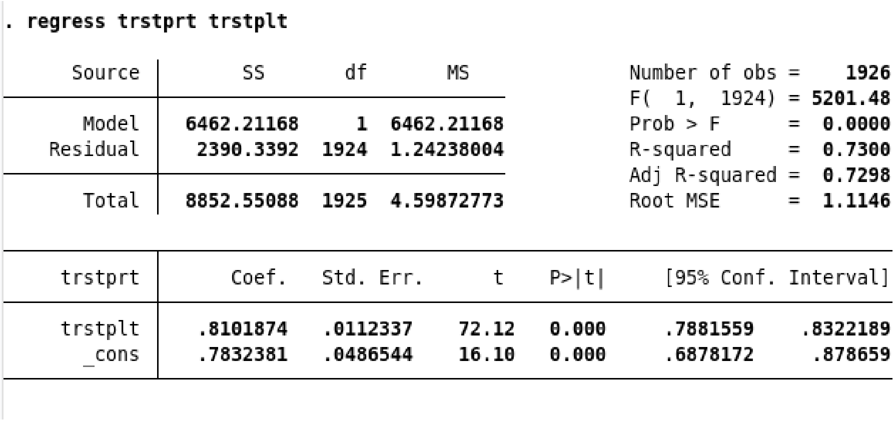
Figure 1: Simple linear regression with happiness as predictor.

We can see that this model was performed on 1,930 respondents (why do you think the model missed 29 people?) The value of R^2 is not impressive (a mere 0.034). This figure indicates that 3.4 percent of the variance in our dependent variable (trust in political parties) is explained by the model (which has only one predictor, happiness). The regression coefficient of happiness is statistically significant at the 0.001

level, but the impact seems rather small: for each one-unit increase in happiness we can expect about 0.22 increase in trust. The constant is statistically significant too, and it indicates the value of the dependent variable when all independent variables are zero (in other words, if we were to graph the regression line the constant is the point at which the line would cut the vertical Y-axis). According to this model, our "starting point" is of about 2 points of trust.

Perhaps happiness is not the best "predictor" of trust in political parties. Let's use this time trust in politicians as a predictor of trust in political parties. We could assume the following (rather redundant) hypothesis: "people with high levels of trust in politicians will also show high levels of trust in political parties."

```
regress trstprt trstplt
```



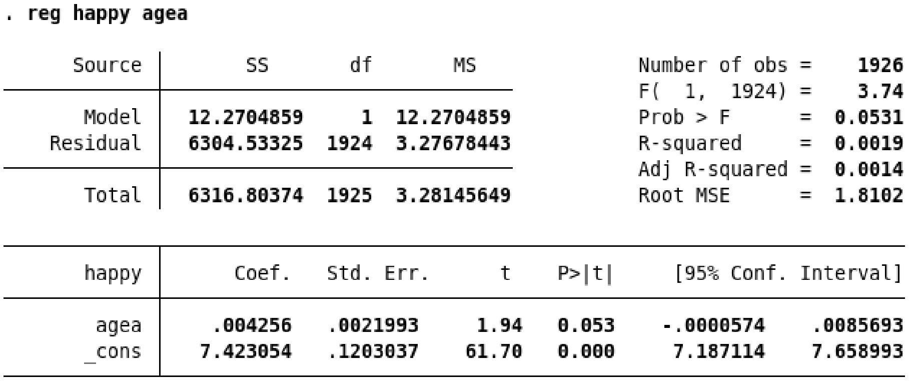
```
. regress trstprt trstplt

      Source |       SS           df       MS      Number of obs   =      1926
-------------+----------------------------------   F(1, 1924)      =   5201.48
       Model |  6462.21168         1  6462.21168   Prob > F        =    0.0000
    Residual |   2390.3392     1,924  1.24238004   R-squared       =    0.7300
-------------+----------------------------------   Adj R-squared   =    0.7298
       Total |  8852.55088     1,925  4.59872773   Root MSE        =    1.1146

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     trstplt |   .8101874   .0112337    72.12   0.000     .7881559    .8322189
       _cons |   .7832381   .0486544    16.10   0.000     .6878172     .878659
------------------------------------------------------------------------------
```

Figure 2: Simple linear regression with trust in politicians as predictor.

We can see several changes. First, the number of observations went down to 1,926. On the other hand, R^2 went up drastically to 0.73 (73 percent of the variance in our dependent variable is explained by our model). The value of R^2 is very high, and in this case it might be a sign that our variables are really measuring the same thing. The regression coefficient is again statistically significant at the 0.001 level (so we can say with confidence that is different from zero). The coefficient itself predicts an increase of 0.81 in trust in political parties for each unit-increase in trust in politicians, i.e., when trust in politicians increase by 1 we can expect an increase in trust in parties of 0.81. Because both variables are measured using the same scale we can easily see that the change is pretty similar (but not identical).

Let's run one last regression. This time we are going to predict happiness using the age of our respondents. Our research question could well be either "older people are happier than younger people" or the opposite.

```
regress happy agea
```



```
. reg happy agea

      Source |       SS       df       MS              Number of obs =    1926
-------------+------------------------------           F(  1,  1924) =    3.74
       Model |  12.2704859     1  12.2704859           Prob > F      =  0.0531
    Residual |  6304.53325  1924  3.27678443           R-squared     =  0.0019
-------------+------------------------------           Adj R-squared =  0.0014
       Total |  6316.80374  1925  3.28145649           Root MSE      =  1.8102

------------------------------------------------------------------------------
       happy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agea |    .004256   .0021993     1.94   0.053    -.0000574    .0085693
       _cons |   7.423054   .1203037    61.70   0.000     7.187114    7.658993
------------------------------------------------------------------------------
```

Figure 3: Simple linear regression with age as a predictor of happiness.

Notice the value of R^2 (0.0019). According to this model, age accounts for a mere 0.19 percent of the variance (information) in happiness. We can also see that the coefficient is not statistically significant at the 0.05 level (yet just so). The coefficient itself would suggest that a one-unit increase in age results in about 0.0043 increase in happiness. This model is not very useful as it does not explain the variance of our dependent variable very well. Exploring the data before and after performing an analysis is critical to finding issues with the distribution of the data that might account for poor-performing models. For example, we could have plotted the average of happiness for each age to see if there is a pattern:

```
graph bar (mean) happy, over(agea, label(labsize(tiny) angle(45)))
```

The resulting graph does not show a clear pattern, and more importantly it does not suggest any linearity at all (remember the “linear” in “linear regression”). Keep in mind that when doing “real” linear regression analysis (i.e., not a lab exercise) the technique requires some important assumptions to be true.

9 Lab Worksheet IX

Multiple regression analysis

In today's worksheet we are going to do some multiple regression analysis on respondents' self-reported happiness using the *European Social Survey*. You can find the dataset **essuk16.dta** on LEARN. Since it is the same dataset we used in the previous worksheet, the same caveats established then also apply here (e.g., regression assumptions are not considered for simplicity's sake) . Remember three key Stata commands widely used to explore datasets:

```
        describe
        summarize
        codebook, compact
```

We should be paying attention to things like the number of variables in the dataset, what they measure, how they are coded, whether we have missing cases and how these are coded. Last time we focused on the variables "happy", "trstprt", "trstplt", and "atchctr." This time let's try:

```
tabulate agea
tabulate region
tabulate gender
```

From the first tabulation we can see that respondents' ages range from 15 to 94. The command `summarize` would have given us this information more easily:

```
summarize agea
```

However, remember that tabulations are not always the best way to represent and explore data. Since "agea" is a continuous variable we could plot it in a histogram:

```
histogram agea, frequency
```

Also, you can explore two variables at the same time using the command `histogram`. For example, we could plot the distribution of respondents' ages in each region:

```
histogram agea, frequency by(region)
```

On the other hand, "region" is a categorical variable, so we could plot it with bars like this:

```
graph bar (percent), over(region, label(angle(45)))
```

The variable "gndr" is also categorical; however if you try plotting it in a bar chart or pie chart you might find it redundant because the variable only has two categories. Perhaps it is more convenient representing "gndr" with a simple tabulation (but again, this is a matter of personal preference). Stata allows you to explore data in a myriad of ways. For instance, consider the following:

```
table gndr, contents(freq mean agea) row
```

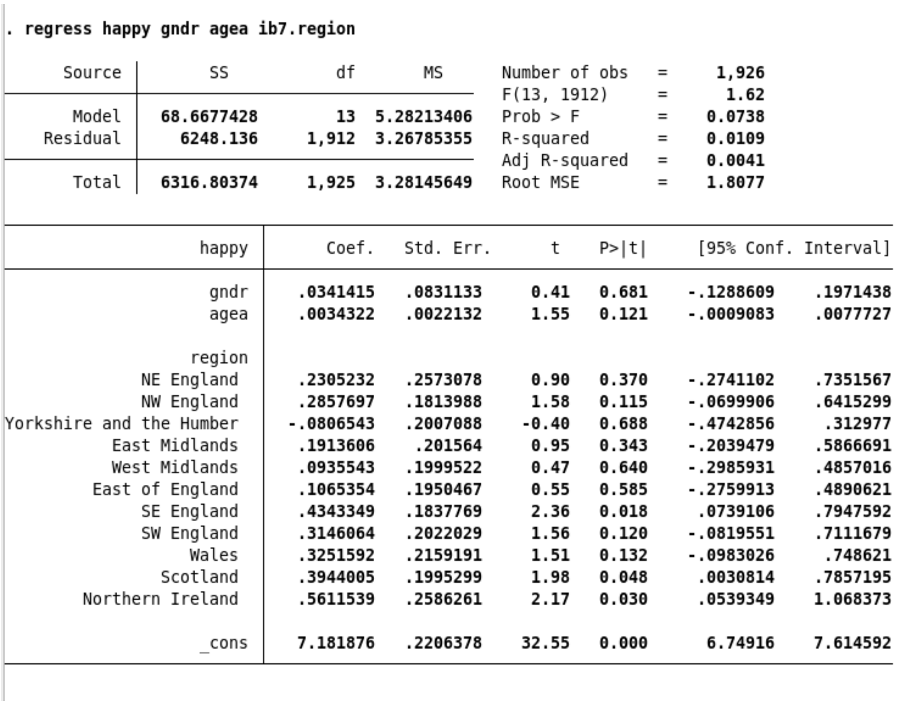
With the `table` command and the `contents` option we can explore gender and age at the same time. The tabulation above shows the mean (average) age for males and females. Tabulating means (not only frequencies or percentages) could sometimes be very useful. If you were to ask the data "are males happier than females?" you would probably want to start by looking at the average level of happiness:

```
table gndr, contents(freq mean happy) row
```

and you would find that female respondents are ever so slightly happier than male respondents on average (7.66 versus 7.6).

Now that we have explored a few variables let's run some regressions. Start with:

```
regress happy gndr agea ib7.region
```



```
. regress happy gndr agea ib7.region

      Source |       SS           df       MS      Number of obs   =     1,926
-------------+----------------------------------   F(13, 1912)     =      1.62
       Model |  68.6677428        13  5.28213406   Prob > F        =    0.0738
    Residual |    6248.136     1,912  3.26785355   R-squared       =    0.0109
-------------+----------------------------------   Adj R-squared   =    0.0041
       Total |  6316.80374     1,925  3.28145649   Root MSE        =    1.8077

------------------------------------------------------------------------------------------
                    happy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------------+---------------------------------------------------------------
                     gndr |   .0341415   .0831133     0.41   0.681    -.1288609    .1971438
                     agea |   .0034322   .0022132     1.55   0.121    -.0009083    .0077727
                          |
                   region |
               NE England |   .2305232   .2573078     0.90   0.370    -.2741102    .7351567
               NW England |   .2857697   .1813988     1.58   0.115    -.0699906    .6415299
Yorkshire and the Humber  |  -.0806543   .2007088    -0.40   0.688    -.4742856     .312977
            East Midlands |   .1913606    .201564     0.95   0.343    -.2039479    .5866691
            West Midlands |   .0935543   .1999522     0.47   0.640    -.2985931    .4857016
          East of England |   .1065354   .1950467     0.55   0.585    -.2759913    .4890621
               SE England |   .4343349   .1837769     2.36   0.018     .0739106    .7947592
               SW England |   .3146064   .2022029     1.56   0.120    -.0819551    .7111679
                    Wales |   .3251592   .2159191     1.51   0.132    -.0983026     .748621
                 Scotland |   .3944005   .1995299     1.98   0.048     .0030814    .7857195
         Northern Ireland |   .5611539   .2586261     2.17   0.030     .0539349    1.068373
                          |
                    _cons |   7.181876   .2206378    32.55   0.000      6.74916    7.614592
------------------------------------------------------------------------------------------
```

Figure 1: Multiple linear regression model with happiness as outcome.

Remember that linear regression requires continuous variables as your dependent variable. In the model above, variables "happy" and "agea" are considered continuous, whereas "gndr" and "region" are categorical. It is important to tell Stata that it needs to treat specific variables as categorical, for your model will need $X - 1$ *dummy variables* (where X is the total number of categories in a variable). Take "region" as an example: this variables has twelve categories (twelve regions), therefore your model will need eleven dummy variables and one *base group*. In the command above, Stata has been instructed to treat "London" as base group (`ib7.` does that, if you wanted the second category to be the base group instead, you can type `ib2.` before the variable name).

The value of R^2 tells you the proportion of variance in your dependent variable ("happy") that is explained by the model. In this case, we can see that this model does not really explain much, only 1.1 percent of variance or information. Moreover, if we look at the p-value for the F-test of overall significance (reported as $Prob > F$ in Stata) we can see that it is not statistically significant itself. This is telling us that our model does not fit any better than a model with no predictors (also called *intercept-only*).

Most coefficients are not statistically significant at the 0.05 level (we already knew this from the F-test of overall significance). A non-statistically significant coefficient is telling you that we cannot be sure the coefficient (i.e., the value of the parameter) is actually different from zero. If you look at the confidence interval of a non-statistically significant coefficient, you will see that the interval includes zero in its range. This means you should not interpret that coefficient. However, do acknowledge in your papers and reports that you have non-statistically significant coefficients by simply saying so.

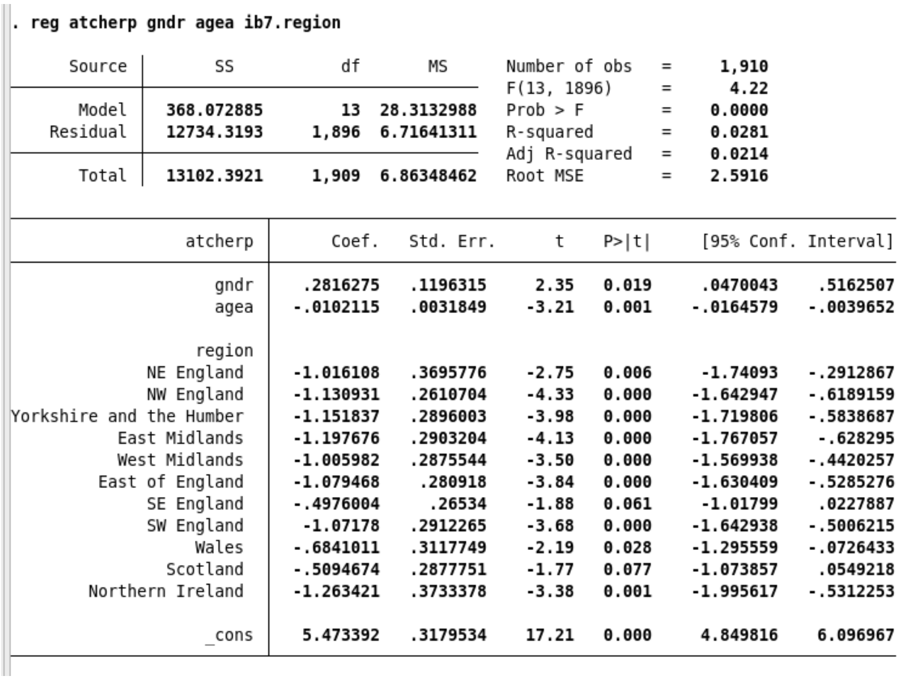
Let's now run the following regression model:

```
regress atcherp gndr agea ib7.region
```

Notice that this new model loses 16 observations (not many) when compared to our previous "happiness model" (the observations go from 1,926 to 1,910). However, now our F-test is statistically significant ($Prob > F = 0.0000$), meaning that our model fits better than an empty or intercept-only model. The model explains about 2.8 percent of information in the dependent variable "atcherp" (a 0-to-10 scale of emotional attachment to Europe).

It is of paramount importance to bear in mind that the interpretation of coefficients of categorical variables is totally different from the interpretation of coefficients of continuous variables. Below the coefficients of the variables "agea" (continuous) and "gndr" (categorical) are interpreted for you:

- "agea": keeping the rest of variables constant, for each one-unit increase in a respondent's age we expect a decrease of 0.01 points in emotional attachment to Europe.
- "gndr": keeping the rest of variables constant, female respondents present an emotional attachment to Europe 0.281 points higher than that of male respondents.



```
. reg atcherp gndr agea ib7.region

      Source |       SS           df       MS      Number of obs   =     1,910
-------------+----------------------------------   F(13, 1896)     =      4.22
       Model |  368.072885        13  28.3132988   Prob > F        =    0.0000
    Residual |  12734.3193     1,896  6.71641311   R-squared       =    0.0281
-------------+----------------------------------   Adj R-squared   =    0.0214
       Total |  13102.3921     1,909  6.86348462   Root MSE        =    2.5916

------------------------------------------------------------------------------------------
                 atcherp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------+----------------------------------------------------------------
                    gndr |   .2816275   .1196315     2.35   0.019     .0470043    .5162507
                    agea |  -.0102115   .0031849    -3.21   0.001    -.0164579   -.0039652
                         |
                  region |
              NE England |  -1.016108   .3695776    -2.75   0.006     -1.74093   -.2912867
              NW England |  -1.130931   .2610704    -4.33   0.000    -1.642947   -.6189159
Yorkshire and the Humber |  -1.151837   .2896003    -3.98   0.000    -1.719806   -.5838687
           East Midlands |  -1.197676   .2903204    -4.13   0.000    -1.767057    -.628295
           West Midlands |  -1.005982   .2875544    -3.50   0.000    -1.569938   -.4420257
         East of England |  -1.079468    .280918    -3.84   0.000    -1.630409   -.5285276
              SE England |  -.4976004     .26534    -1.88   0.061     -1.01799    .0227887
              SW England |   -1.07178   .2912265    -3.68   0.000    -1.642938   -.5006215
                   Wales |  -.6841011   .3117749    -2.19   0.028    -1.295559   -.0726433
                Scotland |  -.5094674   .2877751    -1.77   0.077    -1.073857    .0549218
        Northern Ireland |  -1.263421   .3733378    -3.38   0.001    -1.995617   -.5312253
                         |
                   _cons |   5.473392   .3179534    17.21   0.000     4.849816    6.096967
------------------------------------------------------------------------------------------
```

Figure 2: Multiple linear regression model with emotional attachment to Europe as outcome.

Firstly, notice that when interpreting regression coefficients in multiple regression *you must* keep constant the variables you are not interpreting. Secondly, when interpreting continuous variables we speak of “one-unit increase” if the coefficient is positive or of “one-unit decrease” if the coefficient is negative. However, when interpreting categorical variables (dummy variables) we compare groups to the base group. In the example above, there are only two categories: males and females. Because the variable is coded 1 "male” and 2 “female”, Stata takes by default the first value, “male”, as base group. The coefficient of “gndr”, therefore, is really the coefficient of female respondents in comparison to that of male respondents. Here too a positive coefficient means an increase, but it is an increase over the base group’s effect on the dependent variable. To clarify this point further consider the variable “region” below (remember the base group was “London”):

- “NE England”: keeping the rest of variables constant, respondents from the North East (England) are 1 point less emotionally attached to Europe than respondents from London.
- “Scotland”: keeping the rest of variables constant, respondents from Scotland are 0.51 points less emotionally attached to Europe than respondents from London.

Because all regional categories (i.e., all the dummy variables in the model) present a negative coefficient, we know that London is the region most emotionally attached to Europe. The interpretation of all these coefficients will therefore be in "negative" terms (less than the base group). Needless to say that if we did not know that "London" is the base group we could not interpret these figures. Also, notice that Scotland's coefficient is not statistically significant. What you already know about statistical significance also applies here in the same fashion: because $p > 0.05$ in the case of the category "Scotland", we cannot say for sure the coefficient is different for zero (and we do not interpret it). Finally, the closer the coefficient of a dummy variable is to zero the more similar that category is to the base group.