# **BQA WEEK 2**

## **EXERCISE 3-6:**

## **BIVARIATE PARAMETRIC METHODS**

**+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

DESCRIPTION: This worksheet is actually a collection of four exercises that are similar by their association as parametric methods. I have collated them in this one document to improve the fluidity of the session:

* Exercise 3 will allow you to examine the different ways of producing a correlation coefficient in Stata, and how it compares with the line of best fit.
* Exercise 4 covers the difference between one-sample and a two-sample t-tests, and details the commands for producing multiple t-tests based on the categories of a sorting variable.
* Exercise 5 covers ANOVA and uses as its example how life satisfaction varies with living arrangements.
* Exercise 6 covers OLS regression; it is designed to allow you to assess the equivalence in outputs between ANOVA and OLS regression.

FILES: **BQA\_Lab\_v2.dta**

### STEPS:

1. Open Stata and the BQA\_Lab\_v2 dataset.
2. Open a new do.file titled ‘Exercise 3-6’ so you can save and replicate your work
3. Complete each exercise, and use the answer document and do.file to assess your answers

### COMMANDS:

* Housekeeping: **lab list**
* Modifying variables: **replace**
* Analysing data: **corr, pwcorr, ttest, oneway, pwmean, regress**

### 3. CORRELATION COEFFICIENTS

#### 3.1 CORRELATION MATRIX

Your task is to create a correlation matrix of all the variables in the data set relating to trust – there are 8; find them by typing **lookfor trust**

All these variables contain system missing values, which will need taking care of before you are ready to start (see the missing values by typing **lab list [label name]**. Missing values for all the variables are all values greater than 10. Take care of missing values by

1. Copy each variable into a new one with a similar name (eg **gen ppltrst2 = ppltrst**)
2. Replace the coded missing values with system missing values   
   (eg **replace ppltrst2 = . If ppltrst2 >= 11**)
3. Apply value labels to the new variables (you can copy the existing value labels, since all you have done is re-code the missing values) **lab val ppltrst2 ppltrst**

*This is pretty laborious, so please feel free to cut and paste from the do-file.*

Now, you are ready to produce your correlation matrix

**corr ppltrst2 – trstun2** will produce a correlation matrix

**pwcorr ppltrst2 – trstun2, sig** will also give p-values for every cell

|  |  |
| --- | --- |
| **One variable in the list appears to be less strongly correlated with the others. Which is it?** |  |
| **Two groups of variables exhibit a particularly strong relationship with each other. What are those groups of variables, and why do you think they are particularly strongly correlated?** |  |

#### FINDING THE SLOPE OF A SCATTERPLOT USING OLS

Last week, we plotted scatterplots using four variables aggregated at country level: LS (from LifeSat), IMPRICH, IMPCREATE, GETBY

We’re now going to learn how to calculate the slopes of these lines using OLS. Create a scatterplot, including a line of best fit, between LS and GETBY

**twoway (scatter LS GETBY, mlabel(country)) (lfit LS GETBY)**

|  |  |
| --- | --- |
| **Look at the line of best fit, and write down its approximate slope, (the increase in the Y variable associated with a one-unit increase in the X variable)** |  |
| **Now calculate the slope formally using OLS regression , along with a p-value (the syntax is regress LS GETBY)** |  |

|  |  |
| --- | --- |
| **Now do the same for the relationship between IMPRICH and IMPCREATE: write down your estimate of the slope by looking at the scatterplot:** |  |
| **And calculate it formally using OLS regression, along with a p-value** |  |

**Extension Exercise**: Please only do this in class if you are steaming ahead with the work - (if you are working more slowly, please skip this and return to it later).

The four variables LS, IMPRICH, IMPCREATE and GETBY are country means. The current task is to replicate this analysis at the *individual* level. You can use the variable LifeSat, which has no coded missing values; the others (imprich, ipcrtiv and hincfel) contain coded missing values AND are coded the “wrong” way, with low values indicating high levels of importance and high levels of comfort with income. When you re-code them you will have to sort out missing values and reverse-code so that high levels of the variables relate to high levels of importance/comfort.

|  |  |
| --- | --- |
| **Question: how are these findings similar, and how are they different, to what we found last week at an aggregate country level?** |  |

#### 3.3 RELATIONSHIP BETWEEN CORRELATION COEFFICIENT AND EFFECT SIZE

In the lecture, we learned that the correlation coefficient is NOT the same as the effect size (that is, the slope of the line of best fit), except in the case when variables have been standardized. Let’s see this for ourselves:

The variables **GovSat** and **progressive** have no missing values. Let’s treat **GovSat** as the dependent variable, and **progressive** as the explanatory variable. Create a new variable which contains standardized values of **GovSat**, and another new variable containing standardized values of **progressive**.

The syntax for creating a standardized variable, which we learned in FIAS, is   
**egen stdGovSat = std(GovSat) (**Note that we can call the new variable whatever we like).

Now, fill in the first two columns of the table below, using the **corr** and **regress** commands.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Unstandardized variables**  **(GovSat and progressive)** | **Standardized**  **(stdGovSat and stdprogressive)** | **Standardized, taking complete care of missing values** |
| **Correlation coefficient (use corr)** |  |  |  |
| **Slope (use regress)** |  |  |  |

You’ll notice that in the second column, the two figures are closer, but they are not identical. This is because of missing values. If we standardize a variable, the entire distribution has a mean of 0 and a standard deviation of 1. However, when we analyze that variable together with a second variable, any observations which are missing for *either* variable are excluded from the analysis, and therefore the distribution *for the variables used in the regression* may not have a mean of exactly 0 or a SD of exactly 1. The answer is to standardize over the observations that will be used in the analysis.

See if you can work out how to do this; if not, please consult the do-file supplied. *Don’t worry if you don’t understand the commands for the “intelligent” standardization. Just ensure you know that the correlation coefficient is equal to the slope of the line of best fit if the variables are standardized, and almost always not, otherwise.*

### 4. T-TESTS

#### 4.1 ONE SAMPLE T-TESTS

The syntax for conducting a one-sample t-test in Stata is

**ttest [variable] = [the value you are testing]**

For example,

**ttest LifeSat = 59.5**

provides a test under which the null hypothesis is that the mean of LifeSat is equal to 59.5

The top of the output gives summary statistics; below that, the null hypothesis is set out, and below that, three alternative hypotheses: that the actual mean is greater than, equal to, or less than 59.5.

Below these three hypotheses are the weight of the distribution in question.

Try this with a few different (continuous) variables in the data set, and a few different cut-off points.

#### 4.2 TWO-SAMPLE T-TEST

T-tests are more interesting when they are used to compare samples. At the start of this exercise, we saw that men appeared to have higher levels of life satisfaction than women. We can investigate this formally by typing:

**ttest LifeSat if female != ., by(female)**

The “by” option tells Stata that we are performing this test between the different categories of the variable “female”. Check that you understand all the sections of this command, then look at the output. This time, the differences are expressed in terms of differences between the groups. The difference is highly statistically significant Pr(T > t) = 0.0000

Modify the “if” statement to perform this gender test for people living in Spain and then for people in the UK. What do you find?

|  |
| --- |
|  |

We want to assess whether men or women score more highly on the “progressive values” scale.

To look at the whole sample, the syntax is **ttest progressive, by(female)**

We could look at all countries separately **bysort country: ttest progressive, by(female)**

To look at how progressive values vary with education [note: this variable name contains an underscore]

**bysort country: ttest progressive, by(ed\_d)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Countries in which people with degrees score significantly *lower* on the progressive values scale than people with lower levels of education** | **Countries in which people with degrees score on average *lower* on the progressive values scale than people with lower levels of education, but the difference is insignificant** | **Countries in which people with degrees score on average *higher* on the progressive values scale than people with lower levels of education, but the difference is insignificant** | **Countries in which people with degrees score significantly *higher* on the progressive values scale than people with lower levels of education** |
|  |  |  |  |

### 5. ANOVA

Let’s explore how life satisfaction varies with living arrangements – in this case, how many adults live in your household. This information is recorded in the variable **nadults** but the variable has a few problems – there are a few observations coded zero, and a right-hand tail of large households. Let’s confine our analysis to those living in households containing one to six adults. Create a new variable **nadults2** which contains the values of nadults for households with one to six adults, and a system missing value otherwise.

* *Check in the supplied do-file if you are unsure how to do this.*

Ignore the pre-estimation tests for the moment – run them later if you are interested.

* Now, run the ANOVA by typing **oneway LifeSat nadults2, tabulate**
* Follow it up with **pwmean LifeSat, over(nadults2)**

### 6. OLS

Also run an OLS regression, using the syntax **regress LifeSat i.nadults2**

Enter both sets of results into the table below. They should be identical (or equivalent). You will have to use a post-estimation test to work out the final cell in the table for OLS, and remember that the sample means *per se* are not reported for OLS – rather, you can calculate them from the constant term and the other coefficients.

|  |  |  |
| --- | --- | --- |
|  | **ANOVA** | **OLS** |
| **F-test on whole distribution** |  |  |
| **p-value on F-test** |  |  |
| **Sample mean over those living in single-adult households** |  |  |
| **Sample mean over those living in two-adult households** |  |  |
| **Difference between those living in 1 and 2 adult households** |  |  |
| **CI or p-value of that difference** |  |  |
| **Difference between those living in 4- and 5- adult households** |  |  |
| **CI or p-value of that difference** |  |  |