Introduction to Statistics
lab 3

Johan A. Elkink
jos.elkink@ucd.ie

24 September 2018

Downloading the data

For this third lab we will use teaching data from Dougherty (2011), which is publicly available, with the full data description in Dougherty (2012). You can access the data from the course teaching data page[1] and is listed as "education & earnings". See the instructions for Lab 1 on how to open this file, which is also in Stata format. Remember to always save the file first, then open the statistical package of your choice, then open the file – double-clicking the file name does not work. Make sure you also open the description file, so that you have a list of variables in the data set, with brief descriptions.

For R users, I will assume you opened the data using "educ" as the data set name, i.e. `educ <- read.dta(...)`.

Private sector and degrees

Investigate whether those with a degree (i.e. **degree** equals 1) are more or less likely than those without a degree (**degree** equals 0) to work in the private sector (i.e. **privateSector** equals 1) using cross-tabulation (see Lab 1). Make sure you calculate the correct percentages before you draw conclusions. What do you conclude?

Earnings by gender

In Lab 2 we produced single boxplots, but boxplots can also be very useful to compare the distribution of a variable across the categories of another variable, i.e. to look at the relationship between a scale and a categorical variable. For example, looking at the distribution of earnings for males and females:

[1] `http://www.joselkink.net/data.php`

SPSS: `EXAMINE earnings BY female /PLOT BOXPLOT.`

R: `boxplot(earnings ~ female, data = educ)`

Stata: `graph box earnings, over(female)`

What do you conclude about the relationship between gender and hourly earnings?

Repeat using the logarithmic transformation of the **earnings** variable (see section "Transforming the variable" in Lab 2).

Earnings by education

Produce a scatter plot of the **earnings** variable by the **schooling** variable (see section "Extra" in Lab 2). Which one is the dependent and which the independent variable? Repeat using the logarithmic transform of **earnings**. What do you conclude about the relationship between the two variables?

Calculate the covariance and correlation between **earnings** and **schooling**.

SPSS: `CORRELATIONS /VARIABLES earnings schooling /STATISTICS XPROD.`

R: `cor(educ$earnings, educ$schooling)` and `cov(educ$earnings, educ$schooling)` – note that this assumes there is no missing data. Otherwise you would need to use `cov(educ$earnings, educ$schooling, use = "complete.obs")`.

Stata: `correlate earnings schooling, covariance`

Repeat for the logarithmically transformed **earnings** variable.

How do the correlation coefficients relate to the scatter plots?

Subsetting data

Sometimes it is useful to look at only a subset of the data. For example, if we want to select only females, we could (don't do this now!) use:

SPSS: `FILTER BY female.` – note that this only works for selecting all cases on a variable that are not zero, in this case, by filtering out all men. If you want to filter the opposite way around, you need to construct a new variable first:
`COMPUTE male = 1 - female.` – this inverts the dummy variable **female**.
`FILTER BY male.`
You can undo the filter using: `FILTER OFF.`

R: `educFemale <- subset(educ, female == 1)` – note that this creates a new data set, called "educFemale". Beware in subsequent commands whether you're using "educ" or "educFemale".

Stata: `keep if female == 1` – note that this leads to a loss of data (i.e. you might have to re-open the file to get the unselected data back). This can be avoided by using the subset only with respect to a specific command, e.g.:

```
correlate earnings schooling if female == 1, covariance
```

Now use the above to plot the relationship between the logarithm of earnings by years of education (**schooling**), separately for blacks and for non-blacks in the data (**ethblack**). Calculate the correlation coefficient for the two groups separately. How does it compare to the overall correlation? What does this say about the relationship between schooling and earnings for the two groups?

Extra

Use Google and/or help files to produce a scatter plot of the log of earnings by schooling, colored by gender (**female**).

References

Dougherty, Christopher. 2011. *Introduction to econometrics*. Oxford: Oxford University Press.

Dougherty, Christopher. 2012. "EC220 - Introduction to Econometrics: Education and earnings cross-section data sets." Teaching resource.
URL: *http://learningresources.lse.ac.uk/148/1/Education%20and%20earnings%20cross-section%20data%20sets%20manual%20(EC220).pdf*