BS2280 - Econometrics
R Workshop 2 - Introduction to Regression Analysis

In this computer lab we will have to achieve the following tasks/learning outcomes:

- import .Rdata data and view the data
- generate scatter plot and add regression line
- calculate covariance and correlation
- run regression

Preparing your workspace

Before you do each task, you need to prepare your workspace first.

Step 1. Create a folder called RWorkshop2 under folder BS2280

Step 2. Go to Blackboard week 4 R Workshop 2 and download datafile: crime.Rdata, save it in the RWorkshop2 folder you created in step 1

Step 3. Open Rstudio and set working directory

Menu bar → Click Session → Set Working Directory → Choose Directory → Select RWorkshop2 folder you created in step 1

Step 4. Create an R script

!! If you forget how to prepare your workspace, please review R Workshop 0 first.

For each task, replace the missing part XXXX to make these codes work.

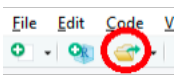
Task 1. Open crime.Rdata.

Before you start working on the crime dataset, ensure that you have prepared the workspace.

From R Workshop 1, we know the native data format of R is .Rdata. In this Workshop 2, we are going to use crime.Rdata so we do not need to activate package "readxl" this time.

To open the R datafile crime.Rdata in your folder R Workshop 2,

Option 1. Click on



and select the crime.Rdata dataset.

Option 2. Use the command line:

```
load("~/Desktop/BS2280/R Workshop2/crime.rdata")
# this is my working path, you should change it to your working path
```

I recommend to check briefly summary descriptive statistics for the variables crimes and officers (see R Workshop 1) to get an idea about the data characteristics. Note that the summary descriptive statistics do not provide any information whether there is a relationship between the two variables or not.

```
XXXX(crime$crimes)          # get summary descriptive statistics for variable crimes
XXXX(crime$crimes)               # calculate standard deviation for variable crimes
summary(XXXX$XXXX)        # get summary descriptive statistics for variable officers
sd(XXXX$XXXX)             # calculate standard deviation for variable officers
```

Task 2. Generate a scatter plot with number of crimes on the y-axis and the number of police officers on the x-axis.

The analyse the relationship between to variables, I recommend to always plot the data first. A good graph type is scatter plot.

You will get the following graph:

```
plot(crime$XXXX~crime$XXXX, main = "Relationship between number of
police officers and crime")
```

Now we look at the codes:

plot(): this function gives a scatter plot.

Note that the call to plot uses formula notation, to specify "officers on crime". If you use ',' instead of '~', you would apply the coordinate vector form (X, Y). I will use the formula notation in this module, as it fits better to linear regression analysis.

crime$officers~crime$crimes: pick variable officers from dataset crime and put it on Y axis, pick variable crimes from dataset crime and put it on X axis

main = : give this plot a title "Relationship between number of police officers and crime"

Task 3. Calculate the Covariance and Correlation Coefficient of number of crimes and number of police officers and comment on their values.

A scatterplot is a good start for identifying relationships between two variables, but it is not sufficient to identify accurately how strong the relationship is between crimes and officers. There are two numerical statistics, that provide more information about the relationship between two variables: The Covariance cov() and the Correlation Coefficient cor().

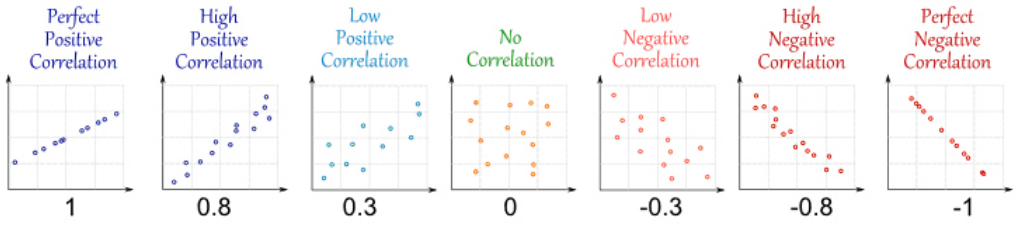
To produce a Covariance matrix, use the following command:

```
XXXX(crime$officers,crime$crimes)
```

The result is: 18212436. The covariance appears with 18,000,000 to be very large, however, the covariance does not provide any information on how strong this relationship is. It only reveals that there is a positive relationship between the number of police officers and the number of crimes committed.

Instead of using the covariance, we can use a 'standardised' covariance - the Correlation Coefficient. To calculate the correlation matrix, we only have to adjust slightly the covariance command.

```
XXXX(crime$officers,crime$crimes)
```



The correlation coefficient for officers and crimes is 0.91. We conclude that there is a strong positive relationship between our two variables.

Task 4. Regress the number of police officers on crimes and comment on

a. the sign and size of the regression coefficients

Running regression in R is simple.

```
officersfit <- lm(XXXX~XXXX,data=crime)
officersfit
```

Look at the codes first.

officersfit: name this regression model as officersfit

lm(): lm() activates the regression function for linear model

officers: the dependent variable

crimes: the independent variable

data=crime: specify variables "officers" and "crimes" are from dataset "crime"

We can avoid repetitively including the crime$ prefix by using the argument 'data = NAME OF DATASET'. The tilde (~) is telling R that we regress officers on crimes.

The estimated regression model is:

$\widehat{officers} = -5.4183 + 0.0238\widehat{crimes}$

The output is produced and we should be careful with the interpretation. The intercept states that if we have zero crimes within a city, the number of police officers would be -5.4. In this case, the intercept is meaningless. Always be careful when interpreting an intercept when it does not lie within the data range.

We do not have any city that has actually a crime rate of zero. The slope coefficient states that for every additional crime, we observe on average 0.024 more police officers.

To use more user-friendly numbers, we can also infer that for every 1,000 additional crimes committed within a city, 24 more police officers are employed.

b. the goodness of fits of the estimated model

R^2 is the measure that provides information on the overall goodness of fit of the model. The R^2 value can be found in the summary output table of the regression object officersfit:

```
summary(XXXX)
```

In this case it is 0.83. This means that 83% of the variation in police officers can be explained with the variation in number of crimes committed. Our estimated model has a good degree of explanatory power.

Task 5. Add a regression line to the scatter plot you created in Task 2.

To add a regression line to the plot, we have to use the regression object officersfit and add it to the previous scatterplot.

abline(): this function adds the regression line to the plot

```
plot(crime$officers~crime$crimes, main = "Relationship between number of
police officers and crime")
XXXX(officersfit)
```

Further Exercise

If you are confident with all the exercises we have completed based on the crime dataset, download the data set called EAWE21.dta from the module page on Blackboard and save it. The dataset is a subset of the Educational Attainment and Wage Equations dataset used in Dougherty (2016) available from https://global.oup.com/uk/orc/busecon/economics/dougherty5e/student/datasets/eawe/ For this exercise we are interested in two variables:

EXP	...	Total out-of-school work experience (years) as of the 2002 interview
EARNINGS	...	Current hourly earnings in $ reported at the 2002 interviews

1. Open the data. This dataset was stored in Stata format .dta, you will have to import it first.
2. Calculate summary statistics (mean, median, minimum, maximum) for the variables EXP and EARNINGS
3. Draw the scatter graph of EARNINGS on EXP and add a regression line
4. Calculate the covariance and correlation between earnings and exp and comment on the values

5. Regress EARNINGS on EXP and comment on

a. the sign and size of the regression coefficients
b. the goodness of fits of the estimated model