JBS Masters in Finance Econometrics Module

Michaelmas 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 1
Descriptive Statistics and Linear Regression in R

Contents

A) Introduction

Note: Commands you need to execute in R are in bold type, with an arrow prefixed

R is an open source statistical software package for use on Linux, Mac and Windows machines. Precompiled binary distributions are available for download at http://cran.r-project.org. For Windows, use: **base** distribution, R version 2.11.1.

Manuals. Several manuals are provided with the software. Useful manuals can be accessed online: http://cran.r-project.org/manuals.html

The ***help*** command (to be explored today) should be your first port of call. http://search.r-project.org provides a powerful search for R functions and mailing-list archives on the web.

This session introduces the R interface (see part B**);**
the menu bar (see part C**);**
the tool bar (see part D**);**
and script files, which allow you to define a sequence of instructions and execute them sequentially.

Exercise I-1. Find help; R commander; install/load packages

There should be an R icon at the bottom of the Quicklaunch section of the desktop sidebar (blue letter "R"). Click on it.

B) The R interface

In *Linux*, the R distribution comes with no pre-installed graphical user interface (GUI). After successful installation, you can run the program directly in the console by typing the upper case letter 'R'. The *Windows* distribution comes with a Java GUI. R opens up with the console window.

To begin with, it probably is convenient to install the 'R Commander' package from one of several R mirrors. Type the following command in the command line:

➢ **install.packages("Rcmdr")**

Choose a mirror (e.g. Bristol or London) and click the 'OK' button. After the package has been unpacked, you can load it in the active workspace by typing:

➢ **library(Rcmdr)** or **require(Rcmdr)**

The R Commander has two windows. You can use the **Script** window to write commands and send them to the console by pressing the 'Submit' button or the hot key 'Ctrl+R'. The **Output** window reproduces the command and gives the corresponding output. Try:

➢ **1+1**

Also, notice the *menu-bar* at the top, starting with **File, Edit**, … **Help**.

Type in at the script window:

➢ **memory.size(2000)**

This is the memory that will now be available for you to work in R (the maximum memory that your datasets may take up).

Case-sensitive. Commands in R are always given in lowercase letters. R is case sensitive, so variable names should be typed exactly the way they were created. Each command has, associated with it, a help file, which can be consulted using the help command. For example:

- **?memory.size**
- **help.search("memory size")**

C) The menu-bar

There are **several ways** to give **commands** to R.

1) Via the menu bar at the top of the screen.
2) You can specify commands in the command line.
3) You may create a text document (a script) with a series of commands and get them to be executed sequentially as a program.

We will learn all the above. The last one is used most common because it is the most flexible. To start we use the *menu bar*.

Exercise I-2. Load a dataset and save it with a different name

Data is best read into R using the 'Comma Separated Value' format (file extension .csv). To convert an Excel (.xls) file to .csv, simply save the dataset as .csv in the spreadsheet editor of your choice. To load a dataset in R from the menu bar:

- **Data/Import data/from text file…/**

Enter 'eaef2' as the name of the dataset, select 'Commas' as the field separator, click 'OK' and then select the dataset.

The drive which contains the data is JBSroot on 'PROTON (ntdomain)' which has the alias V. If you explore that drive, you will be able to find the folder: V:\Public\MP01\data

Open the CSV dataset: eaef.csv

Convince yourself that the data was successfully imported:

- **ls()**

Save the dataset in your own file space, from the menu bar:

- **Data/Active data set/Save active data set… [your network directory for this module's lab sessions] \eaef2.RData**

If you do not have an account in the JBS network, you may use a temporary folder that you create in C:\. Copy the files you create in this folder into your own USB by the end of the session.

If you are not a student from the Judge, provide me with your name and crsid for the use of the JBS IT Services, who will create an account for you.

You can load multiple datasets into R. **A dataset is in the form of a matrix** with variables arranged in columns and a different observation in each row. To refer to a variable in R you must indicate which dataset it belongs to by typing dataset**$**variable. Using the **attach**(dataset) and **detach**(dataset) options, you can simply refer to a variable in the attached dataset by typing its name.

Clear the R workspace:

- **rm(list=ls())**

and open your newly saved file.

- **Data/Load data set/eaef2.RData**

Variables in dataset eaef.csv

Personal variables

AGE	C	age in 1994	
S	C	years of schooling (highest grade completed as of 1994)	
MALE	D	sex of respondent (1 if male, 0 if female)	
ETHBLACK	D	ethnicity:	black
ETHHISP	D		hispanic
		score on a component of the *ASVAB* battery (scaled with mean 50, standard deviation 10)	
ASVAB2	C		arithmetic reasoning
ASVAB3	C		word knowledge
ASVAB4	C		paragraph comprehension
ASVABC	C	composite of *ASVAB2* (with double weight), *ASVAB3* and *ASVAB4*	
CHILDREN	C	number of children in the household	
YOUNGEST	C	age of youngest child	
CHILDL06	C	presence of a child age < 6 in the household	
CHILDL16	C	presence of a child age < 16, but no child age < 6, in the household	
MARISTAT	T	marital status, coded as: 1 never married; 2 married, spouse present; 3 other	
MARRIED	D	married (*MARISTAT*=2)	

Work-related variables

EARNINGS	C	current hourly earnings in $ reported at 1994 interview
WORKING	D	working (has recorded earnings)
EMPSTAT	T	employment status, coded as: 1 employed; 2 unemployed; 3 out of the labor force

For more comments regarding loading data **see Appendix 1**.

Exercise I-3. Observe the displays in different windows and create a script file

Observe **the main transformations** that occurred in the different windows.

In the **'Script window'**, all the commands that we executed are stored in sequence.

In the **'Output window'** we observe commands AND results written in different colours. In particular, note that the commands that have been executed are presented in red, prefixed with a '>'.

Note the structure of the commands given:

```
eaef2 <- read.table("C:/…/eaef.csv", header=TRUE, sep=",",
na.strings="NA", dec=".", strip.white=TRUE)
```

Each dataset loaded must be given a name (here for example: ‘eaef2’) as in R there can be multiple dataset in use at a time. The above command reads the table stored in the file C:/…/eaef.csv into the current workspace. The options specified tell R that (1) the first line of the file contains names of the variables, (2) commas are field separator characters, (3) “NA” is the string that is to be interpreted as Not Applicable (NA) value, (4) “.” is the character used in the file for decimal points, and (5) leading and trailing white space should be stripped from character fields.

Note that when the last item in the output window is a red ‘>’ all by itself, the program is ready to receive a fresh command.

As a program, R functions by manipulating variables. The **notion of a variable** in R corresponds directly to the notion of a variable in statistics. Thus each variable has a certain number of observations associated with it (frequently all variables will have the same number of observations). Each observation corresponds to what we think of as a data point.

Now create a **Script-file**:

- **File\ New Script** or: press **Ctrl+O**

Save the new script in your own file space to document this session:

- **File\ Save as…** or: press **Ctrl+S**

Copy all commands used so far in the R Commander into the script file.

Note the advantage of using script compared to mouse and menu: it allows you to document your results and helps others to replicate them. Your script could look like this example:

```
# ---------------------------------------------------------------------------------
# Econometrics Module
# Lab Session 1
# ---------------------------------------------------------------------------------

# --- Ex 1: find help, R Commander, install/load packages -----------------

?memory.size                       # help if command is known
help.search("memory size")        # help if command is not known

install.packages("Rcmdr")          # install package Rcmdr
library(Rcmdr)                     # load package


# --- Ex 2: Load a dataset and save it with a different name ---------------

eaef2 <- read.table("C:/…/eaef.csv", header=TRUE, sep=",", na.strings="NA",
dec=".", strip.white=TRUE)         # read dataset from .csv file

ls()                               # display active objects in workspace

save("eaef2", file="C:/…/eaef2.RData")   # save active object eaef2

rm(eaef2)                          # clear object eaef2 from workspace
rm(list=ls())                  # clear workspace
load("C:/…/eaef2.RData")       # load object eaef2
```

The new Script-file is used in the same way as the ‘Script window’ in the R Commander. Commands are sent to the console (equivalent to the Rcmdr’s ‘Output window’) using:

- **Ctrl+R**

You can switch between console and script using

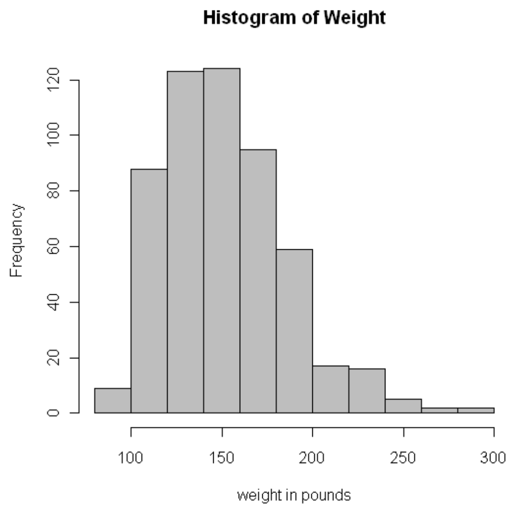
- **Ctrl+Tab**

Exercise 1: Summary statistics

Create useful summary statistics for the variables in the dataset eaef2. To facilitate this task, use the attach command. The command below looks at the dataset eaef2 and reads the names of the variables. It then sets each one as a variable in its own right.

- **attach(eaef2)**

- **hist(weight)**



summary(weight)
mean(weight)
median(weight)
quantile(weight)

sd(weight)

max(weight)
min(weight)
range(weight)
sum(weight)
length(weight)

- **library(timeDate)**

skewness(weight, method="moment")
kurtosis(weight, method="moment")

boxplot(weight)

Exercise 2: Calculate proportions of observations presenting a certain characteristic

a) What is the proportion of observations with 3 siblings?

Use **length()**, the condition operator **[ ]** and the logical assertion **==** to look-up the size of this subsample:

- **length(siblings[siblings==3])**

Look-up the size of the total sample:

- **length(siblings)**

We have 118 observations with 3 siblings and 540 observations in total. So the proportion of observations is 118/540.

b) What is the proportion of observations with weight less than 120?

```
length(weight[weight<120])        # and so on…
```

Exercise 3. Analysis of the frequency of discrete variables

a) Tabulate the count, percentage, cumulative count and cumulative percentage for every given number of siblings (from 0 to 13).

```
table(siblings)                     # count
t <- table(siblings)                # percentage
round( t/sum(t), 4 )*100
cumsum(t)                           # cumulative count
round( cumsum(t/sum(t)), 4)*100     # cumulative percentage
```

			Cumulative	Cumulative
Value	Count	Percent	Count	Percent
0	24	4.44	24	4.44
1	95	17.59	119	22.04
2	140	25.93	259	47.96
3	118	21.85	377	69.81
4	68	12.59	445	82.41
5	41	7.59	486	90.00
6	19	3.52	505	93.52
7	14	2.59	519	96.11
8	10	1.85	529	97.96
9	3	0.56	532	98.52
10	4	0.74	536	99.26
11	2	0.37	538	99.63
12	1	0.19	539	99.81
13	1	0.19	540	100.00
Total	540	100.00	540	100.00

b) What is the proportion of observations with 3 siblings?

Look-up 4th row, 2nd column in the matrix above (21.85%).

Exercise 4. Frequency of a combination of discrete variables (two-way tables)

The command **cut()** allows you to categorise a continuous variable into several levels. Produce a variable agecut with four levels based on the quartiles of age.

```
eaef2$agecut <- cut(age, breaks=quantile(age))
```

Cross-tabulate the new variable agecut and siblings.

```
attach(eaef2)
table(agecut, siblings)
```

Exercise 5. Analysis of a variable conditioned on a discrete variable

a) Evaluate the mean earnings conditioned on the four age-groups defined above.

```
➢ by(data=earnings, INDICES=agecut, FUN=mean)
```

b) Plot histograms for earnings conditioned on the four age-groups.

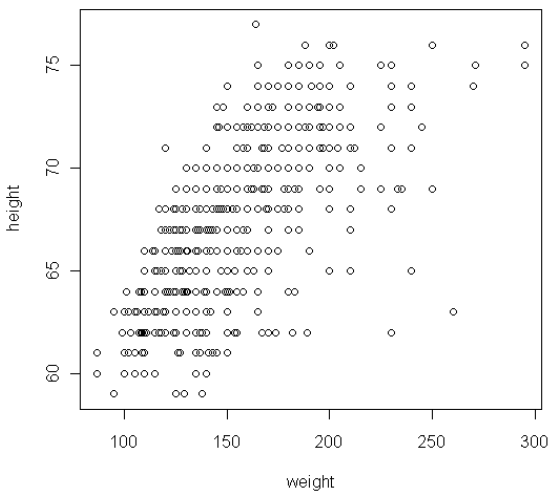
```
➢ par(mfrow=c(2,2))
➢ by(data = earnings, INDICES = agecut, FUN = hist)
```

Exercise 6. Graphs

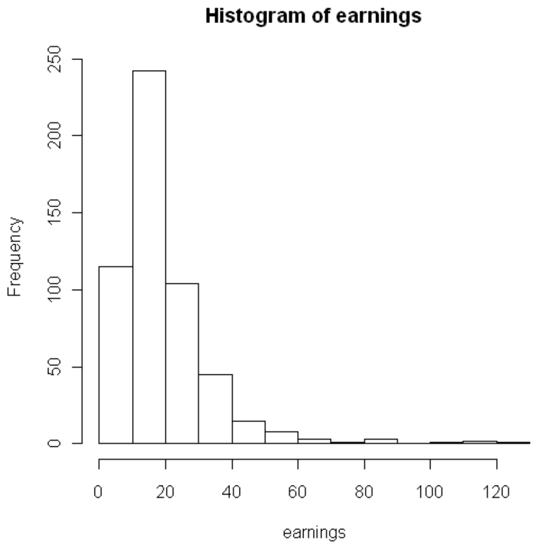
a) Produce a scatter plot of height against weight.

b) Produce a histogram of earnings.

```
➢ plot(height ~ weight)
```



```
➢ hist(earnings)
```



Exercise 7. Generate linear transformation

a) Generate a linear transformation to obtain post tax-benefit earnings

```
➢ eaef2$ptearnings <- 2 + (earnings-2)*0.8
```

b) Generate a log-transformation of earnings.

```
➢ eaef2$earnings <- log(earnings)
```

Exercise 8. T-tests

We use the normal distribution when: i) the population is known to follow a normal distribution with known population variance; or ii) when the shape of the population distribution is not known but the size of the sample is bigger than 30. A company wants to assess their new measures to control claim costs.

Costs used to be 60 pounds. They wish to test if they have been reduced. They are willing to work with a 5% level of significance. They extract a sample of 26 observations:

45	49	62	40	43	61	48	53	67	63
78	64	48	54	51	56	63	69	58	51
58	59	56	57	38	76				

a) Test the hypothesis that they benefited from the new procedures.

b) Perform this test and calculate the confidence interval at 99% significance level. (This part, to be done in Lab-Session 1).

Answer:

a) First, just put the data in vertical form with Excel:

Ho: μ=60; Ha: μ<60

The test is one-tailed one so. We decide the significance level at 5%. The test statistic is the t-student, because the sample is small and we don’t know about the variance of the population.

The computed value for t is $t_{25} = \frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{56.42-60}{10.04/\sqrt{26}} = -1.818$. The critical value is -1.708, so we reject the null. Note that if we had decided to test it at 1% significance level, the result would have been the converse!

b) For this part, first highlight the data and copy them into your clipboard (Ctrl+C). Then read the data from the clipboard into R:

- **data <- scan("clipboard")**

Second, the confidence interval at 99% level:

- **mean(data) + c(-1,1)*sd(data)*qnorm(0.99)**

The test of hypothesis is performed using:

- **t.test(data, mu=60, alternative="less", conf.level=0.99)**

Observe the outcome: note that different results arise according to the type of test suggested (one-tail on the left, one-tail on the right, two-tail). Note that the confidence interval is built for a two-tail test.

Exercise 9. The linear model

Load the dataset growth.csv.

This dataset shows the average rates of growth of GDP and employment for 25 OECD countries for the period 1988-1997. It was taken from Dougherty's book. Mexico is not included because it is an outlier, as employment increases dramatically after the implementation of NAFTA. The reason is that individuals who worked in the informal sector (and therefore were not included in the series) moved into the formal sector with the arrival of US manufacturing companies.

a) Check the content and regress employment growth on GDP growth. Provide and interpretation of the results.

b) Visually inspect data and regression line.

c) Are the coefficients significant?

d) Is it there any other interesting test to run?
e) Is it a good fit?
f) Build the interval of confidence for the slope yourself.
g) How would you interpret the column after the t?

Answers:

a)

```
➢ str(growth)
➢ lm9 <- lm(empgrow ~ GDPgrow, data=growth); summary(lm9)
```

What is the regression line saying? In the first column the table gives the name of the regressor, in the second it gives its estimate. The regression implies that a 1 percent increase in the growth of GDP generates a 0.48 percent increase in the rate of growth of employment. Should the investigator expect increments of the same magnitude in growth rate of employment and that of GDP? According to these results she shouldn't, technical progress is clearly making GDP grow more than employment.

The intercept suggests that, if GDP is static (growth = 0), employment will have a negative growth rate of 0.55 percent per year (maybe technical change saves labour). In some slow-growing countries employment growth has actually been negative, and this could be the reason for this result on the intercept.

b)

```
➢ plot(empgrow ~ GDPgrow, data=growth)
➢ abline(lm, col="red")
```

It is evident that the true relationship is in fact nonlinear. Probably a function of a different form for the explanatory variable would be more suitable. We will study this issue in the next session.

As modelers, we are interested in testing whether GDP growth has or hasn't had an impact on employment (and therefore shouldn't be included in the model). For this purpose, we define H0: $\beta_2 = 0$ (so it has no influence). Then we fix the maximum probability we allow for the error of type I[1] (the level of significance) and the critical region is defined so that the error of type 2 is minimized. For this purpose we need the tables.

c) Let's start with the slope. We may perform two types of tests:

i) We may think that it is meant to be positive, as in the long run both variables should be positively correlated. The test would be:

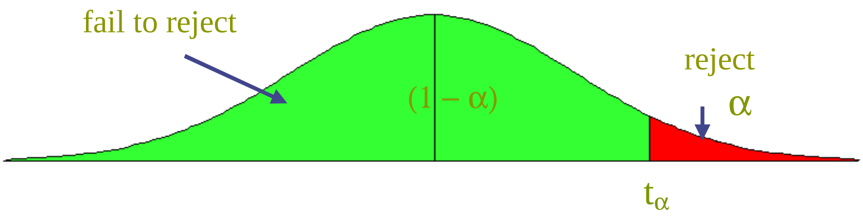
H_0: $\beta_2 = 0$; Ha: $\beta_2 > 0$

This is a one tailed test. We have to define the level of significance (α, let's say we fix it at 5%) and then look for the t-value from the tables at the point up to where the t-Student cumulates 0.95 of probability. In our case the value we look for is 1.714, as we have 23 degrees of freedom. The t-statistic is given in the fourth column. In this case the t-statistic is 5.75 (as such, higher than the t from the tables) and, therefore, we

[1] The probability of rejecting H_0 when it is true.

reject the null. Why's that? Remember: $t_b = {}^{b-\beta}/_{se(b)}$. Commonly, in our examples of test of hypothesis we're thinking that if $\beta = 0$; then, the distribution of ${}^{b}/_{se(b)}$ is a t-student with *n-K* degrees of freedom. Now if t_b does not fall within the bigger zone in Figure 1, then we *decide* that the null hypothesis was a wrong one, and we reject it.

Figure 1. One tailed test

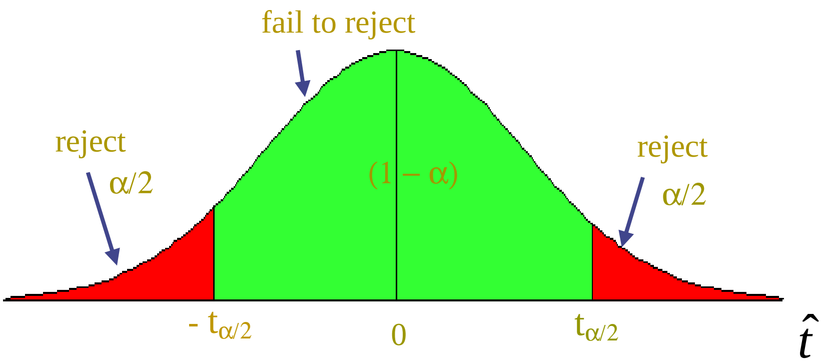


ii) Instead we may have no idea of which the sign of the slope may be. The test would be, in that case, a two-tail one:

H$_0$: $\beta_2 = 0$; Ha: $\beta_2 \neq 0$

This is a two-tailed test (we don't say that the statistic has to be much larger than the null to reject, but just that it has to be far from it: much larger or much smaller). Again, we have to define the level of significance (α) and then look for the t-value from the tables. Now we want 5% aggregated in both tails, so we need to look for $t_{\alpha/2} = t_{0.25}$ or the value up to which the t-Student cumulates 0.975 of probability. In our case, this is 2.069, as we have 23 degrees of freedom. In this case the t-statistic is 5.75 (as such, higher than the t from the tables) and, therefore, we reject the null.

Figure 2. Two tailed test



d) In fact it would've been interesting to test H$_0$: $\beta_2 = 1$; Ha: $\beta_2 \neq 1$, that is, if employment grows as fast as GDP or if labour-saving technical progress makes that it is less than 1. t-statistic would be then (1- 0.4897)/0.08511= 6.13, which is higher than 2.069, and so we reject the null.

e) With an R^2 of 0.59 it seems quite a good fit, especially considering that there's only one regressor.

f) We will build the interval of confidence only for the two-tailed test at a 5% level of significance. The question in Figure 2 is what is the interval such that β_2 falls in it with a 95% of confidence? We don't have a table of β_2's distribution, but remember that: $t_{b_2} = {b_2 - \beta_2}/{se(b_2)}$, and this distribution is described in tables. We can use, therefore, this in order to build the interval of confidence. In fact we want to compute the interval in which P(|t_{b_2} | ≤ 0.975 t-quantile) = 0.95. Basically, we're saying that the probability that our estimator b_2 differs from the parameter β_2 by a small number, 0.975 t-quantile (= $t_{0.975}$), is very big, 0.95. It only remains to operate:

$$P(|t_{b_2}| \leq t_{0.975}) = 0.95$$

$$P\left(\left|{b_2 - \beta_2}/{se(b_2)}\right| \leq t_{0.975}\right) = 0.95$$

$$P\left(-t_{0.975} \leq {b_2 - \beta_2}/{se(b_2)} \leq t_{0.975}\right) = 0.95$$

$$P(-b_2 + se(b_2)\cdot t_{0.975} \leq \beta_2 \leq -b_2 + se(b_2)\cdot t_{0.975}) = 0.95$$

$$P(b_2 - se(b_2)\cdot t_{0.975} \leq \beta_2 \leq b_2 + se(b_2)\cdot t_{0.975}) = 0.95$$

In our case this is:

$$0.489737 - 0.0851184 \cdot 2.069 < \beta < 0.489737 + 0.0851184 \cdot 2.069$$

And this is what we have in the last two columns.

DIY with the intercept. Note that the null is not rejected at 5% significance level when we consider two-tailed tests (critical value being 2.07, do: **qt(p=1-0.025, df=23)** or: **qt(p=0.025, df=23, lower.tail=F)**); but it is rejected for one-tailed tests (critical value being 1.71, do: **qt(p=1-0.05, df=23)**).

g) In the fifth column the p-value is reported. This informs us about how much probability is cumulated in both tails. That is the probability of having obtained the t-statistic that we did obtain, or others higher if the null hypothesis is true. If the p-value is less than 0.05 then $\hat{t}$ has fallen in the darker probability zone (the critical region), and we reject the null hypothesis at a 5% level of significance. In this case this is what happens with the intercept.[2] **Note that p-values and confidence intervals are computed for a two-tailed test!**

Exercise 10. The linear model with quadratic terms

Use housing.csv. For many years it has been conjectured that households spent a constant share of their incomes in housing.

a) Estimate a model to test this, using total expenditure as a proxy for total income.

[2] This is what we want, to reject the null. Otherwise, in principle, our model would not be explaining the dependent variable. The rule of thumb is: a low value for the p-value indicates that our model is in good health.

b) Is a quadratic form more appropriate?

a) We have data of *expenditures on housing* and *incomes*. We want to test whether housing/income is constant. So we may want to estimate

$housing_i = \beta_1 + \beta_2 \cdot income_i + u_i$

```
lm10a <- lm(housing ~ total, data=house)
```

If $\beta_1 = 0$ then the share would have been constant. This hypothesis is rejected.

b) We try now $housing_i = \beta_1 + \beta_2 \cdot income_i + \beta_3 \cdot income^2_i + u_i$

```
house$totalsq <- house$total^2
lm10b <- lm(housing ~ total + totalsq, data=house)
```

β_3 is not significant. Since we don't have a theoretical background which would define the polynomial form to apply, we drop the squared component: share is increasing with income.

Exercise 11. Extrapolation and accuracy of least squares

Load the eaef.csv dataset. Learn a bit about the dataset by using **str()**. Is it possible to explain the weight of the students measured in pounds (*weight*) with their height measured in inches (*height*)? Provide an interpretation of the coefficients.

Answer:

The regression implies that, for every extra inch of height, an individual tends to weigh an extra 5.56 pounds.

Note the negative value of the intercept. This would suggest that an individual with no height would weigh –221 lbs (pounds). Of course this has no meaning and raises an important issue: if you don't have observations close at both sides of the ordinates (no X negative) or even no X close to 0, then you may find no reasonable intercepts.

Accuracy of least squares:

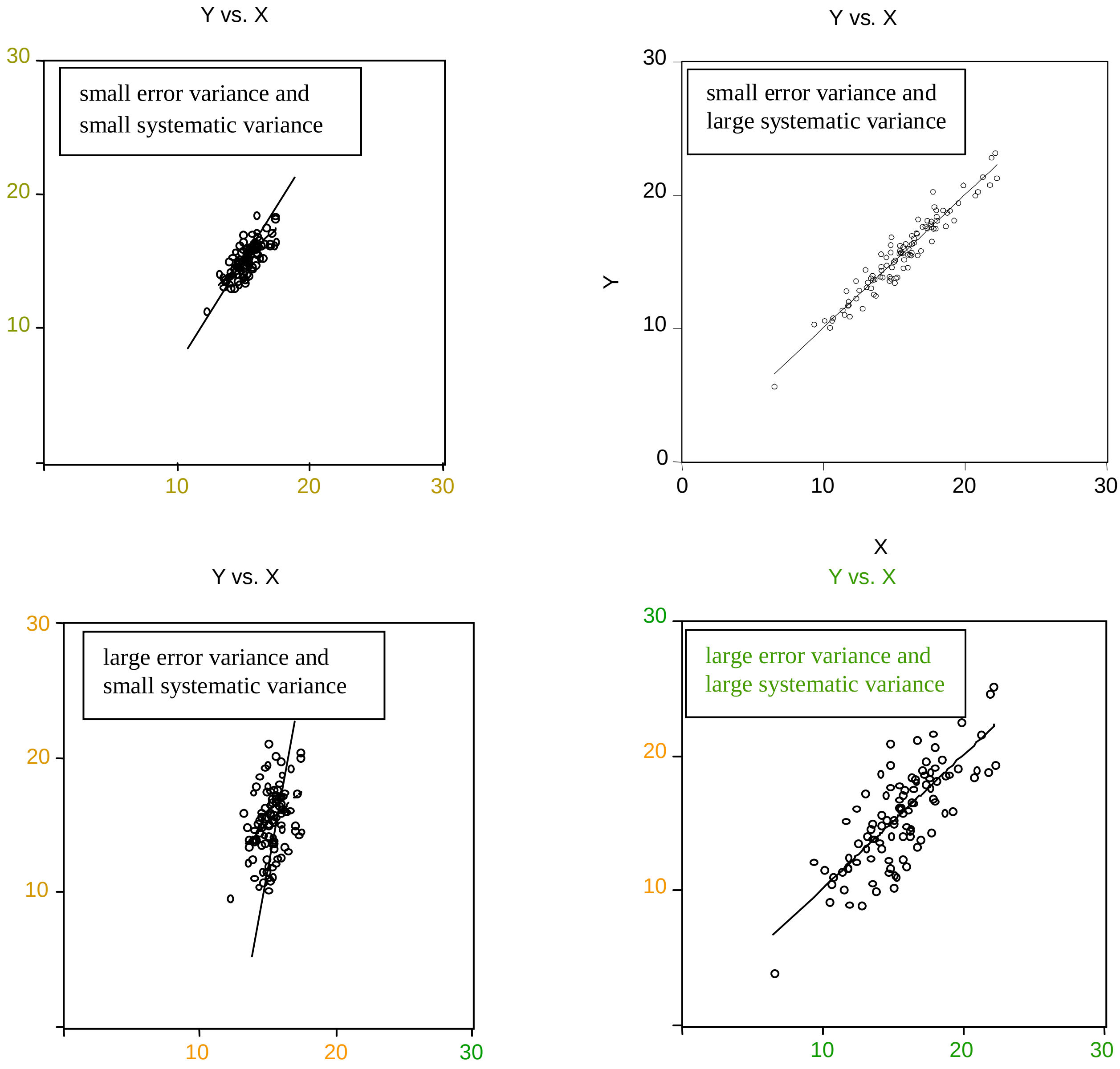
Remember in the simple linear regression model: $Y_i = \beta_1 + \beta_2 \cdot X_i + u_i$

$b_2 = \frac{\sum(X_i-\bar{X})\cdot(Y_i-\bar{Y})}{s_x^2}$, where $s_x^2 \equiv \sum(x_i-\bar{x})^2$

$$b_2 = \frac{cov(X,Y)}{Var(X)}$$

$$Var(b_2) = \frac{\sigma_u^2}{Var(X)}$$

With the last formulae we may see how the precision of estimator for $\hat{\beta}_2$ (its variance) varies with the variances of the errors and of the Xs (called systematic variance in the figure below). In the Figure presented in next page there are four possible cases. Note that for a good fit not only a small variance of the errors is needed but also large variation in the regressors. It is basically case 2 where the variance of the errors is low and the systematic variance is high. In the diagrams below, this corresponds to the lower left hand side figure.



The standard deviation of x in the right diagrams is 3 times as large as in the left ones, and the standard deviation of the error terms in the lower diagrams is 3 times as large as in the upper ones.

Exercise 12. Estimates for changing units of measurement

a) Consider what slope coefficient would have been in Exercise 4 if *weight* had been measured in *grams*. Consider what changes would have occurred to the original slope coefficient if *height* were measured in metric units, i.e. *cm*.

b) Confirm these conclusions by creating the new variables in R and comparing the estimated parameters. What happens with the slope? (*Note*: one pound is 454 grams, and one inch is 2.54 cm.)

Answer:

a) Let the weight and height be *W* and *H* in *imperial units* and *WM* and *HM* in *metric units*. Then *WM* = 454*W* and *HM* = 2.54*H*.

$$b_2 = \frac{cov(H,W)}{Var(H)}$$

Remember:

If $E(Y) = \mu_Y$ and $Z = \lambda \cdot Y \Rightarrow E(Z) = E(\lambda \cdot Y) = \lambda \cdot \mu_Y$

$$cov(X,Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)]$$

$$\Rightarrow cov(X, \lambda \cdot Y) = E[(X - \mu_X)(\lambda Y - \lambda \mu_Y)]$$

$$= \lambda \cdot E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \lambda \cdot cov(X,Y)$$

We apply this property two lines below. The slope coefficient for the regression with weight measured in grams, b_2^G, is given by

$$b_2^G = \frac{cov(H,WM)}{Var(H)} = \frac{cov(H, 454 \cdot W)}{Var(H)} = 454 \cdot \frac{cov(H,W)}{Var(H)} = 454 \cdot b_2$$

The slope coefficient for the regression with height measured in centimeters, $\hat{\beta}_1^{CM}$, is given by

$$b_2^{CM} = \frac{cov(HM,W)}{Var(HM)} = \frac{cov(2.54 \cdot H, W)}{Var(2.54 \cdot H)} = \frac{2.54 \cdot cov(H,W)}{2.54^2 \cdot Var(H)} = \frac{1}{2.54} b_2$$

In other words, if we change scale in the Y, multiplying it by a factor κ, then the estimate for the slope will also be multiplied κ. On the other hand, if we change scale in the X, multiplying it by a factor γ, then the estimate for the slope will be divided by γ.

b)

```
eaef$weight_grams <- eaef$weight*454
eaef$height_metric <- eaef$height * 2.54

lm(weight_grams ~ height, data=eaef)
5.562*454 # =2525.148

lm(weight ~ height_metric, data=eaef)
5.562496/2.54 # =2.189959
```

Exercise 13. Multiple linear regression

Use hprice1.csv and familiarize yourself with the dataset to estimate the model

$$price = \beta_0 + \beta_1 \cdot sqrft + \beta_2 \cdot bdrms + u$$

where *price* is the house price measured in thousands of dollars.

a) Write out the results. What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

b) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size?

c) What percentage of the variation in price is explained by square footage and number of bedrooms?

d) The first house in the sample has *sqrft* = 2,438 and *bedrms* = 4. Find the predicted selling price for this house from the OLS regression line.

e) The actual selling price of the first house in the sample was $300.000 (*price=300*). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

Answer:

a) *price = -19.32 + 0.128 · sqrft + 15.20 · bdrms + u*

The estimated increase in price, given square feet size is 0, β_2= 15.20. Note that it is insignificant, so the increase in price due to an increase in 1 bedroom is not statistically different from zero.

b) Now the increase in price is higher because the house is larger.

$\Delta price = 0.128 \cdot (140) = 17.92.$ (or: $17.920)

c) That's equal to the R^2, 63.2%.

d) The predicted price is –19.32 + .128(2,438) + 0(4) = 292.74, or $292,740.

e) If the actual selling price was $300,000, the buyer overpaid by some margin. But, of course, there are many other features of a house (unobserved by us) that affect price, and we have not controlled for these.

Exercise 14. Reversal of regressor and regressand

Load eaef.csv. The theory indicates that earnings are determined by schooling. Two individuals model this problem. The first individual does it correctly and obtains the following result:

$$\widehat{earnings} = -12.6 + 2.37 \cdot schooling$$

The second individual, instead, first regresses *schooling* on *EARNINGS*, obtaining the following result:

$$\widehat{schooling} = 12.24 + 0.073 \cdot earnings$$

From this result the second individual derives

$$earnings = \frac{(-12.24 + schooling)}{0.073}$$

and concludes:

$$\widehat{earnings} = -167.7 + 13.7 \cdot schooling$$

a) Explain why this equation is different from that fitted by the first individual. Is it only one of them correct.

b) Under which circumstances would both individuals get the same results?

Answer:

a) The slope coefficient for any estimation is equal to Cov(Y,X)/Var(X).

The first individual calculated the slope coefficient as Cov(*earnings*, *schooling*)/Var(*schooling*). This is what this exercise was asking.

The slope in the second strategy corresponds to: Cov(*earnings*, *schooling*)/Var(*earnings*). The second applicant, then, revises the equation, and in an attempt to estimate the parameter in the model recommended by theory uses the inverse of this to estimate the parameter on schooling in the original model. Therefore, she is effectively using the expression Var(*earnings*)/Cov(*earnings*, *schooling*). Obviously the two individuals are using different estimators and therefore in general will obtain different results.

b) The estimates in fact turn out to be identical when

$$\frac{cov(earnings,schooling)}{Var(schooling)} = \frac{Var(earnings)}{cov(earnings,schooling)},$$

which is

$$\frac{[cov(earnings,schooling)]^2}{Var(schooling) \cdot Var(earnings)} = 1\,,$$

In other words, both strategies produce the same results only when the correlation coefficient is equal to plus or minus one.

Exercise 15. Regression against a constant (optional)

What happens if we only include the constant as a regressor? **a)** Examine this by estimating a model for weight using eaef. **b)** Demonstrate algebraically.

Answer:

a)

```
➢ lm(weight ~ 1, data=eaef)
➢ summary(eaef$weight)
```

b)

The model that we're estimating is: $Y_i = \beta_1 + u_i$
We need to calculate the corresponding sum of square errors and then minimize them. First, then, we calculate the errors:
Let the fitted model be: $\hat{Y}_i = b_1$
Then e_i, the error in observation i, is given by

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1$$

and the sum of square errors, or residual sum of squares (RSS), is given by

$$\begin{aligned} RSS &= \textstyle\sum_{i=1}^{n} e_i^2 \\ &= \textstyle\sum_{i=1}^{n} (Y_i - b_1)^2 \;= (Y_i^2 - 2b_1 Y_i + b_1^2) \\ &= \textstyle\sum_{i=1}^{n} Y_i^2 + \sum_{i=1}^{n}(-2b_1 Y_i) + nb_1^2 = \sum_{i=1}^{n} Y_i^2 - 2b_1 \sum_{i=1}^{n} Y_i + nb_1^2 \end{aligned}$$

The first-order condition for a minimum is:

$$\frac{dRSS}{db_1} = -2 \cdot \textstyle\sum_{i=1}^{n} Y_i + 2 \cdot n \cdot b_1 = 0$$

Hence: $b_1 = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}$

The second derivative of *RSS*, 2n, is positive, confirming that we have found a minimum.

In sum, if *Y* is a random variable with unknown population mean β_1, we have shown that the sample mean of *Y* is equal the least squares estimator (and, therefore, the BLUE estimator) of b_1 in the model $Y_i = \beta_1 + u_i$.

Exercise 16. Confidence intervals for regression coefficients

A researcher hypothesizes that years of schooling, *schooling*, may be related to the number of siblings (brothers and sisters), *siblings,* according to the relationship

$$schooling = \beta_1 + \beta_2 \cdot siblings + u$$

She tests the null hypothesis H_0: $\beta_2 = 0$ against the alternative hypothesis H_1: $\beta_2 \neq 0$ at the 5 percent and 1 percent levels. Assume he has 60 individuals. What should she report? (*Note: this exercise may be repeated at home with real data using eaef.dta).*

1. if $b_2 = –0.20$, s.e.(b_2) = 0.07?
2. if $b_2 = –0.12$, s.e.(b_2) = 0.07?
3. if $b_2 = 0.06$, s.e.(b_2) = 0.07?
4. if $b_2 = 0.20$, s.e.(b_2) = 0.07?

Answer:

There are 58 degrees of freedom, and hence the critical values of *t* at the 5 percent and 1 percent levels are 2.001 and 2.663 respectively.

```
➢ qt(p=0.025, df=60-2, lower.tail=F)
➢ qt(p=0.05, df=60-2, lower.tail=F)
```

1. The *t* statistic is -2.86. Reject H_0 at the 1 percent level.
2. $t = –1.71$. Do not reject at the 5 percent level.
3. $t = 0.86$. Do not reject at the 5 percent level.
4. $t = 2.86$. Reject H_0 at the 1 percent level.