JBS Quantitative Research Methods Module MPO1

Michaelmas 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 2
Linear Regression in R; Good Practice Guide

Contents

Script-files

Script files are collections of R commands you wish to execute in sequence and comments that you add in order to have a better understanding of what is going on. They may become especially helpful in complex tasks and to keep track of your actions, to fix errors and to keep documentation on how and why you did what you did.

Concept. Script-files are text-files in which instructions to R are defined by the user in a pre-defined sequence. The commands are set and executed in the same order as they are defined in the text. As such Script-files are programs, but beware, as in R this word (program) is reserved for a very specific kind of program.

The principle is just that you place a series of commands you intend to execute in a text file and the operating system reads through the file and executes the commands. With a script file you can modify and replicate your commands, perhaps on new or modified datasets.

Any text-editor may be used to write these, but we start by using the one defined already in R for this purpose.

Run a script-file. You have several ways to do it once a script file is open:

a) Highlight and press Ctrl+R to run the selected part.
b) Execute the commands line by line by just pressing Ctrl+R.
c) Another way is to save the script-file and run it from the command line or from another script-file by typing: **source("[your directory]/Scriptname.R")**

Exercise 1. Run Script-file from command line

a) Open a new Script-file
Write:

- **print("hello, world")**

Press Ctrl+R to run it.

b) Save it in your project area as hello.R
Change working directory to your project area in the command window:

- **setwd("[your directory]")**

Then type:

- **source("hello.R")**

c) Create a new script-file, keep it open throughout the session and use it to record the (correct) commands you run for each exercise.

Main assumptions of the Classic Linear Model

$$Y_i = \beta_1 + \beta_2 \cdot X_i + u_i$$

First, a bit of terminology:

Table 1: Regression terminology

Y	X	U	β	$\hat{\beta}$ **or simply** b	$\widehat{Y}$
Dependent variable	Independent variable	Error term	Parameter	Estimator of β (or estimate of β, if we have carried out the calculations from the sample)	Estimate of Y (or estimator)
Explained variable	Explanatory variable	Distur-bance	Coefficient		
Regressand	Regressor		**unknown**		

β_1 : intercept parameter;
β_2 : slope parameter.

Exercise 2. 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) Review the contents and regress employment growth on GDP growth. Provide an interpretation of the results.
b) Visually inspect data and regression line.
c) Are the coefficients significant?
d) Is there any other interesting test you may wish to carry out?
e) Is the fit good?
f) Build a confidence interval for the slope.
g) How would you interpret the column after the t?
h) Explain the remaining of the statistics reported when you run the model

Answers:

a)

- **str(growth)**
- **lm2 <- lm(empgrow ~ GDPgrow, data=growth); summary(lm2)**

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.

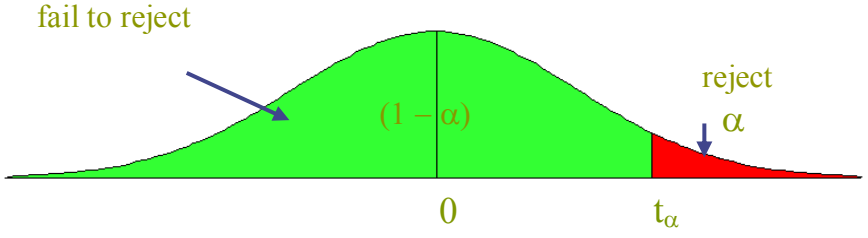
c) Let's start with the slope. We may perform two types of tests:

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 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



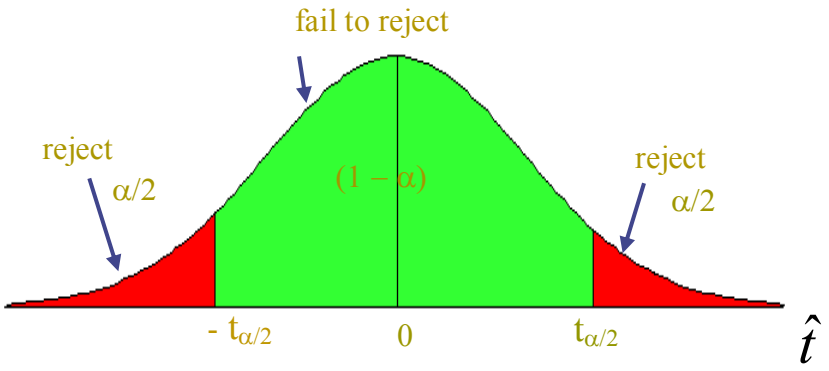
[1] The probability of rejecting H_0 when it is true.

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 $P(|t_{b_2}| \leq (t \ \ from \ \ tables_{0.975})) = 0.95$. Basically, we're saying that the probability that our estimator b_2 differ from the parameter β_2 by a small number, $(t \ \ from \ \ tables_{0.975}) = t_{0.975}$, is very big, 0.95. It only remains to operate:

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

$$P\left(-t_{0.975} \le \left(b_2 - \beta_2\right) \Big/ se(b_2) \le t_{0.975}\right) = 0.95$$
$$P\left(-b_2 - se(b_2)\cdot t_{0.975} \le \beta_2 \le -b_2 + se(b_2)\cdot t_{0.975}\right) = 0.95$$
$$P\left(b_2 - se(b_2)\cdot t_{0.975} \le \beta_2 \le b_2 + se(b_2)\cdot t_{0.975}\right) = 0.95$$

In our case this is:

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

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!**

h) Now, what about the information presented below the model coefficients? It indicates, under SS (sum squares) the sum of squares of the model (or explained sum of squares): SSE = $\sum_i \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2$; the sum of squares of the residuals: SSR = $\sum_i e_i^2$; and of the dependent variable (total sum of squares): SST = $\sum_i \left(Y_i - \bar{Y}\right)^2$. Note that it can be proved that

SSE + SSR = SST. This can be used to calculate R^2 as the proportion of the SST that is explained by SSE:

R^2 = SSE/SST = 1 – SSE/SST, and it'll be between 0 and 1. The degrees of freedom for the model is always 1, the degrees of freedom for the residuals are n-K (K being the number of betas in the model, in this case 2,[3] and total degrees of freedom is equal to n-1). Under MS (mean squares) you'll find the result of dividing the first column by the second. So 'mean square residuals' corresponds to the estimate for the variance of the residuals, in this case, 0.4403. On the left hand side, you have the number of observations that were considered (always check), F is a test of the model's goodness of fit. We'll talk about this in the multivariate case as now it does not add any information to the t-test of the betas. The adjusted R-squared, as you know, takes into account the amount of variables included in the model. We'll talk about this also when

[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.

[3] Note that you also estimate the variance of the disturbances, but you don't count this parameter for this purpose.

we deal with the multivariate case. Finally, the root MSE is the root of the mean squared errors.

Exercise 3. 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.
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

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

- **lm3a <- 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**
- **lm3b <- 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 4. 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:

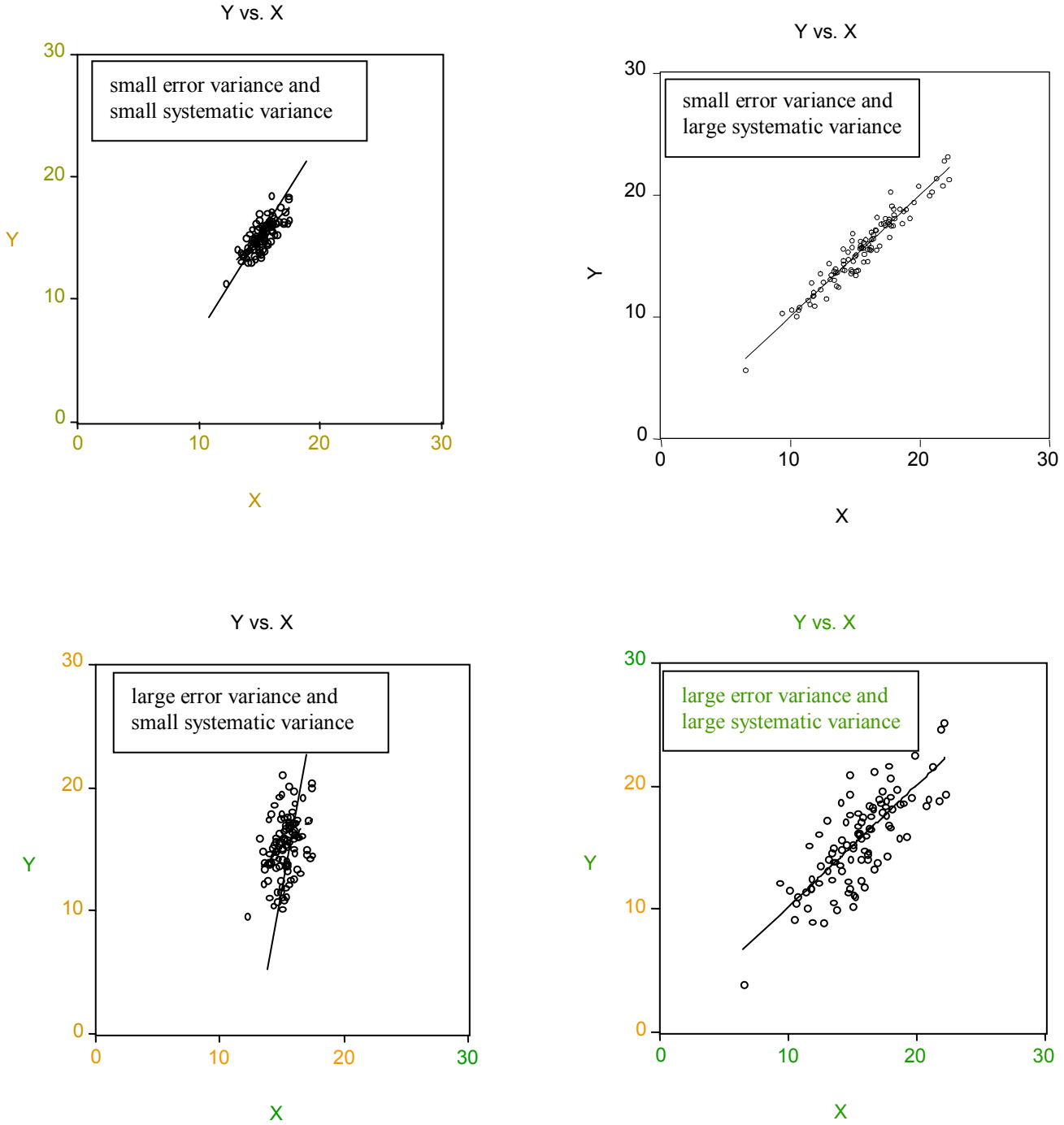
$$Y_i = \beta_1 + \beta_2 \cdot X_i + u_i$$

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

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

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

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 5. 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 $WM = 454W$ and $HM = 2.54H$.

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

Remember:

$$If \quad E(Y) = \mu_Y \quad and \quad Z = \lambda \cdot Y \Rightarrow E(Z) = E(\lambda \cdot Y) = \lambda \cdot \mu_Y$$
$$\text{cov}(X,Y) = E\left[(X - \mu_X)\cdot(Y - \mu_Y)\right] \Rightarrow$$
$$\Rightarrow \text{cov}(X, \lambda \cdot Y) = E\left[(X - \mu_X)\cdot(\lambda \cdot Y - \lambda \cdot \mu_Y)\right] = \lambda \cdot E\left[(X - \mu_X)\cdot(Y - \mu_Y)\right] = \lambda \cdot \text{cov}(X,Y)$$

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

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

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

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

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 6. 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) $price = -19.32 + 0.128 \cdot sqrft + 15.20 \cdot bdrms + u$

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.

$$\Delta price = 0.128 \cdot (140) = 17.92, \text{(or: \$17.920)}$$

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 7. Confidence intervals for regression coefficients (1)

Use oilprice1.csv. This is an example that can be found expanded in the very good book by Murray (2006). It has to do with a trial, where the judge has to decide if the price differential charged to oil suppliers to a pipe because of differences in the quality of oil is fair. The quality is measured in API degrees (higher with greater quality). Up to the trial the implicit premise of the rule applied by the carriers is \$0.15 per API degree of oil. The challengers want a price of about 3 and 5 cents. In this data set you have information on the crude oil's quality and price per barrel.

a) Observe by a scatter-plot if quality has any impact on prices.

b) Use regression analysis to quantify this relationship.

c) Construct a 95% confidence interval for how much the price of barrel changes when API increases in one degree.

d) Is the price charged fair?

e) Perform the same tests in cents. For this create a new variable multiplying the price by 100. Does the fit of the regression change? What happens to the coefficients estimated?

Answers:

a)

- **str()**
- **plot(price ~ api, data=oilprice1)** # there's a strong relationship

b)

- **lm7 <- lm(price ~ api, data= oilprice1)**
- **summary(lm7)**

c)

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

Lookup 0.975-quantile of t-distribution with n-2 degrees of freedom:

- **qt(p=0.975, df=13-2)** # = 2.201

$$0.09493 - 0.00827*2.201 \le \beta_2 \le 0.09493 + 0.00827*2.201$$

$$0.08 \le \beta_2 \le 0.11$$

d) The results in the last part indicate that the price that the market pays is statistically within the interval 9 cents and 11 cents per API degree, more would be unfair. The price suggested by the challenger is too low.

e) This is another example of Exercise 5. The only change is in the scale of the results. R^2 does not change.

Exercise 8. 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:

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

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

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

From this result the second individual derives

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

and concludes: $\hat{earnings} = -167.7 + 13.7 \cdot schooling$

a) Explain why this equation is different from that fitted by the first individual. Is 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

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

which is

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

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

Exercise 9. 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

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$RSS = \sum_{i=1}^{n}(Y_i - b_1)^2 = \sum_{i=1}^{n}(Y_i^2 - 2b_1Y_i + b_1^2)$$
$$= \sum_{i=1}^{n}Y_i^2 + \sum_{i=1}^{n}(-2b_1Y_i) + nb_1^2 = \sum_{i=1}^{n}Y_i^2 - 2b_1\sum_{i=1}^{n}Y_i + nb_1^2$$

The first-order condition for a minimum is:

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

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

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 10. Confidence intervals for regression coefficients (2)

A researcher hypothesizes that years of *schooling*, may be related to the number of *siblings* (brothers and sisters), 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 she has 60 individuals. What should she report? (*Note: this exercise may be repeated at home with real data using eaef.csv).*

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.

Recommendations on good practice in R programming

In the first session we went through the menu bar in the *R Commander* in order for you to feel comfortable that any time you are lost you can find the command that you're looking for up there. This is not the most common practice in R, it is too slow. Today we review important concepts of a good programming practice in R. The key message is that we should always be able to repeat exactly all the steps that we are doing whenever is necessary in the future. Imagine that you produce a report in which you had certain datasets as your source, you deleted some observations, made some calculations, went wrong somewhere then did something else and arrived at your final results. Would you rely on your memory to reproduce this a month later when a referee (or an editor or an examiner or a client) asks you for details? And, pedestrian, but… in three months time, will you be able to reproduce all the work that you have done when you have to present your analysis for MPO1? Hence the need for a good method of keeping your work so that you can trace back all the steps.

The most basic rule:

Be prepared to repeat all steps of your work at any time in the future.

Rule 1. Every project will have its own folder

In this one we'll save the original data source (which we will never modify), and the rest of the files that we will be creating with this one, with our instructions and with the output that we will be creating.

Exercise 1. Change directory

Change directory:

- **setwd("projectpath")**

R looks for files in the directory that was given by default. Usually C:/Programme/R/R-2.10.1/bin/. You may not want this. If you're working on a project you may want to refer to it at the beginning so that you may save and load files without mentioning the whole path. **setwd()** is used for this purpose. It changes the path defined as default.

Rule 2. Always add good documentation to your script-files

Explain as much as you can every time that you have to deal with a complex task. All annotation that may be useful later when you come back to this piece of work should be included. This could be extended to labeling correctly variables and data when necessary.

You can add documentation to your script-file by adding "comments" to yourself. This can be done in several ways. We use # to open a comment. This line then will not be read as a command.

Exercise 2. Comments

Try first:

- **# this is a comment**

To add longer comments you can then try:

- **if(2==3){ Comment #1**
  This
  is
  a
  comment.
  }

Rule 3. A clear format for your programs: copy and paste the following every time you create a script-file. (then edit it as appropriate)

Exercise 3a. THE HEADING. Start a Script-file as follows:

- **setwd("project folder")**

```
# ------------------------------------------------------------------------------------------------
# This Script-file does the following:
# Inputs:    fill in
# Outputs:  fill in
#
# Created by:          (you)      Date:
#
```

Exercise 3b. THE TEXT-BODY. Complete your Script-file as follows:

Whenever tasks are complex it is useful to include Sub-headings by adding

```
# ------------------------------------------------------------------------------------------------
#
#  Step 1: summary statistics
#
```

…..

```
# ------------------------------------------------------------------------------------------------
#
#  Step 2: regression analysis
#
```

And so on.

Exercise 3c. Save this as a Script-file session2.R in your folder

In practice this would be the folder that you defined for your specific project:

- **File\Save as… session2.R**

Rule 4. Never write very long scrip-files

It is difficult to follow what you're doing if you write a script-file over several pages. If you have a long task it is better to split the task up in several script-files. A good recommendation is to do it in the following way. You'll have three kinds of script-files:

Those needed for *creating the database*, those needed for *analyzing it*, and a *master* script-file, which will provide the ordering in which the other script-files will run.

Example: you want to study income per household according to household characteristics. It may be a long task which you split up into: creating new variables with characteristics for children; creating new variables with characteristics for parents; creating new variables with household characteristics; analyzing the income distribution; analyzing household characteristics; a master script-file sequencing the order of the others.

They will be:

- cr_parentschar
- cr_childrenchar
- cr_hhincomes
- an_incomedis
- an_hhchar
- master

Here is a rule I learnt in a course given by R to name these files. By sticking to it I saved many problems. *The order is important in those script-files named with prefix cr_*, as you may create a variable in cr_parentschar that will be needed in cr_childrenchar (for example if you create a variable with age-bands for all individuals). *It is not important in those files with prefix an_*. That doesn't mean that they will not generate variables. But these variables should not be passed further on. That is no an_-file should depend on variables created by another an_-file.

This is how master script-files may appear:

```
# ------------------------------------------------------------------------
# Master Script-file
# ------------------------------------------------------------------------

source("cr_parentschar")
source("cr_childrenchar")
source("cr_hhincomes")
source("an_incomedis")
source("an_hhchar")
```

It will be called master.R, and you may just run all the script-files in sequence by the command:

- **source("master.R")**

Rule 5. The original data-files are sacred. Do not over-write or save over them!!

You should keep them always safe from being overwritten once you load them. You'll need to be able to go back to them any time you need to. In case you managed to overwrite your script-file, have a look at the folder C:/Programmes/R/R-2.10.1/bin. R keeps you a log-file called .Rhistory that contains all comments ever sent to the console for sessions that were

saved with q("yes") when leaving R. Every time you save the current workspace by typing q("yes") to exit R, the file gets appended.

Rule 6. Keep track of all your changes.

I recommend you to avoid using **edit()** command. You may do the same with commands in the command line and you will have the record of what you've done.

Rule 7. Don't include experiments in your script-file.

Do any experimentation by giving instructions interactively, from the command line, and only include the successful commands in the script-file. Otherwise, interesting output will be mixed with useless output.

Of course you may, on occasions, want to be a sinner and work quick and dirty. That's human nature, but try to stick to the rules whenever you feel what you do is important.

Rule 8. Try to keep a neat environment in your folders.

It helps, for instance, to give similar names to your files across projects, like the recommendation given in rule 5 in order to manage your script-files. For the original source datasets I usually give names starting with base: basecompanies, baseincomes. You may want to develop your own convention. It is really useful when you start with several sources and you finish with several outputs. Usually, there also are lots of files that may be transitorily created by your script-file and which you don't need to keep as an output.

These should be erased, so that the folder doesn't get messy because of these. However, you may want to leave them in the first tries of your script-file in order to check that every step runs fine. When you're happy that the program is ok, you may decide to erase in order to finish with a neat folder.

Books employed in this session:

- Dougherty, Christopher (2002). *Introduction to Econometrics*. Oxford University Press.
- Murray, Michael (2006), Econometrics, A Modern Introduction. Pearson.
- STATA Manuals
- Wooldridge, Jeffrey (2000), *Introductory Econometrics*. A Modern Approach. Thomson-South Western.