TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Quantitative Methods III - Practice 1
Simple Linear Regression

Prof. Lorenzo Cavallo: lorenzo.cavallo.480084@uniroma2.eu
Prof. Marianna Brunetti: marianna.brunetti@uniroma2.it

Exercise *Investments in Research and Development* (X, million EUR) and *Number of patents* (Y) are collected for 5 companies. The distribution of X and Y is as follows:

X	Y
5	7
6	8
7	8
9	10
9	9

1. Draw the scatterplot of the distribution: what kind of relationship do you expect?
2. Estimate the parameters of the regression line and explain their meaning.
3. Add the regression line to the graph produced in point (1).
4. Predict the number of patents for a company that invests 11 million euros in Research and Development.
5. Calculate the best goodness-of-fit indices of the estimated line to the observed data and comment on the results.
6. Calculate the correlation coefficient.

7. Construct the 95% confidence interval for the *slope* of the regression line (it is assumed that the errors are normal distributed).

8. We want to test the hypothesis that the number of patents (Y) increases or decreases linearly as investments in Research and Development (X) increase. What is the hypothesis system?

9. Calculate the p-value for the test described in the previous point.

10. Compute the p-value for a two-sided hypothesis test with $H_0 : \beta_1 = 1$. Reject H_0 at the 5% significance level? What if we assume a significance level of 10%?

11. Construct a 90% confidence interval for the $y - intercept$ of the regression line (assuming normality of errors).

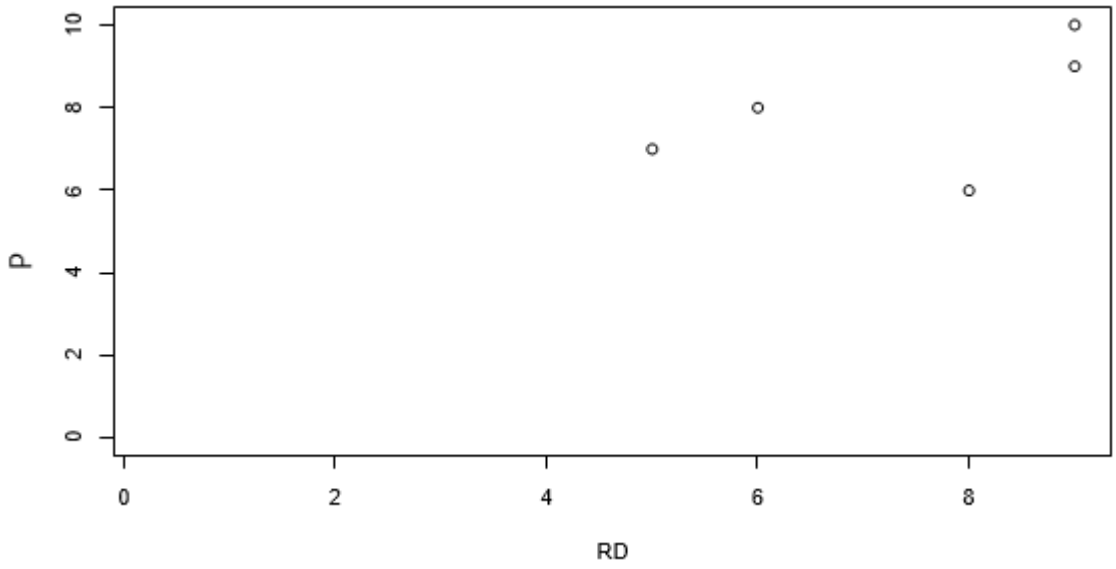
12. Suppose that by repeating the analysis on a sample of $n = 100$ observations, the following notes are obtained:

```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.7715   6.519
x            0.4766      0.1032          1.19e-05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.87 on 98 degrees of freedom
Multiple R-squared:  0.1786,    Adjusted R-squared:  0.1702
F-statistic: 21.31 on 1 and 98 DF,  p-value: 1.185e-05
```

Fill in the missing elements of the regression output (highlighted in yellow) and derive the 95% confidence intervals for both estimated coefficients.

Solutions

1. The following graph shows the scatterplot of the distribution. A positive relationship is therefore expected between *Investments in Research and Development* (RD) and the *Number of patents* (P).



2. The simple regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

with Y dependent variable and X independent variable.
The least squares estimators of the regression line parameters are:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

The first step is to calculate the means of X e Y:

$$\bar{x} = \frac{5+6+7+9+9}{5} = 7.2 \qquad \text{e} \qquad \bar{y} = \frac{7+8+8+10+9}{5} = 8.4$$

The further calculations necessary to obtain the parameter estimates are shown in the following table, on the basis of which we obtain:

$$\hat{\beta}_1 = \frac{7.6}{12.8} = 0.59 \qquad \text{and} \qquad \hat{\beta}_0 = 8.4 - (0.59 \times 7.2) = 4.14$$

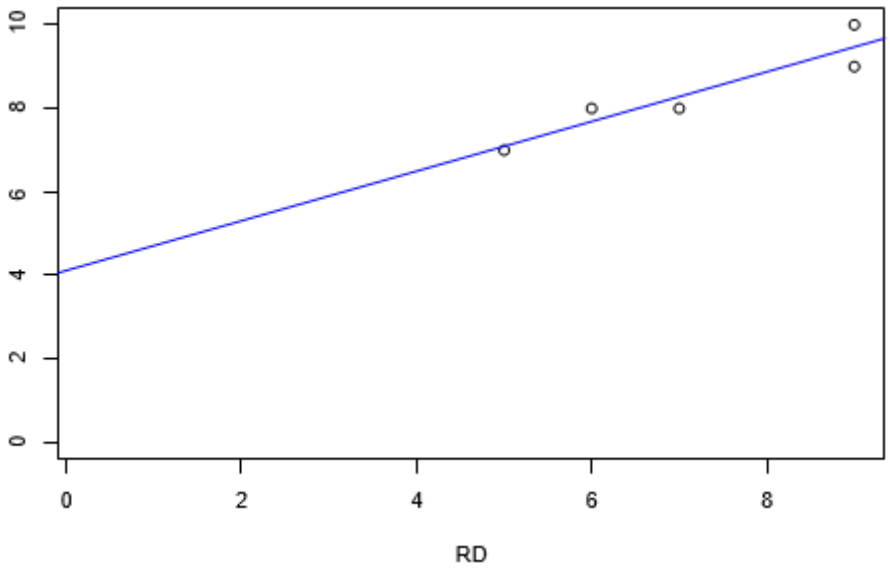
X_i	Y_i	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
5	7	$(-2.2) \cdot (-1.4) = 3.08$	4.84
6	8	$(-1.2) \cdot (-0.4) = 0.48$	1.44
7	8	$(-0.2) \cdot (-0.4) = 0.08$	0.04
9	10	$(+1.8) \cdot (+1.6) = 2.88$	3.24
9	9	$(+1.8) \cdot (+0.6) = 1.08$	3.24
		7.60	12.80

Therefore, the estimated regression line is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 4.14 + 0.59 X_i$$

and we conclude that the linear relationship between X and Y is increasing: if X (*Investments in RD*) increases by one unit (1 million euro) the value of Y (*Number of patents*) increases by 0.59.

3. In the following graph, the estimated regression line is added to the scatterplot. The initially hypothesized positive relationship is confirmed.



4.
$$\hat{Y}(11) = 4.14 + 0.59 \times 11 = 10.63$$
The number of patents for a company with an investment value in RD equal to 11 is 10.63 (> 10).
Following the same methodology, it is possible to obtain the predicted values $\hat{Y}_i$ of the model, shown in the following table.

5. The goodness-of-fit index generally used is the R^2. There are also the measures of the standard deviation of the residuals, the SER and the $RMSE$.

Let us first recall the formula for R^2, where with $\hat{u}_i$ we indicate the residuals:
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$
The calculations are shown in the following table:

X_i	Y_i	$(Y_i - \bar{Y})^2$	$\hat{Y}_i$	$u_i = (\hat{Y}_i - Y_i)$	u_i^2	$(\hat{Y}_i - \bar{Y})^2$
5	7	1.96	7.09	0.09	0.01	1.71
6	8	0.16	7.69	-0.31	0.10	0.51
7	8	0.16	8.28	0.28	0.08	0.01
9	10	2.56	9.47	-0.53	0.28	1.14
9	9	0.36	9.47	0.47	0.22	1.14
		5.20		0.00	0.69	4.51

hence,
$$R^2 = \frac{4.51}{5.20} = 1 - \frac{0.69}{5.20} = 0.86$$
The value assumed by the index denotes a good adaptation of the regression line to the observed values. In fact, 86% of the Y variability is explained by the regression line.

The SER and the $RMSE$ are respectively equal to:
$$SER = \sqrt{\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2}} \qquad \text{and} \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n}}$$

hence,

$$SER = \sqrt{\frac{0.69}{5-2}} = 0.48 \qquad \text{and} \qquad RMSE = \sqrt{\frac{0.69}{5}} = 0.37$$

6. In the regression model, the goodness-of-fit index R^2 (coefficient of determination) is equal to the square of the correlation coefficient ρ_{XY}. Therefore, it is possible to calculate this relationship in an inverse way to obtain that:

$$\rho_{XY} = \sqrt{R^2} = \sqrt{0.86} = \pm 0.93$$

Since the covariance between the variables is positive, we conclude that

$$\rho_{XY} = 0.93$$

.

7. With n large enough, the $(1-\alpha)\%$ confidence interval for the slope of the regression line is:

$$[\hat{\beta}_1 \pm Z_{1-\alpha/2} SE(\hat{\beta}_1)] \quad \text{where} \quad SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} = \sqrt{\frac{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}}{n \times V(X_i)}}$$

In our case $n < 30$. However, thanks to the assumption of normality of the errors, we can proceed in a similar way, with the only foresight to use the t-Student distribution instead of the Normal distribution.

$$[\hat{\beta}_1 \pm t_{\alpha/2}^{n-2} SE(\hat{\beta}_1)]$$

Let's start by calculating $SE(\hat{\beta}_1)$. From the previous point,

$$\sum_{i=1}^{n} \hat{u}_i^2 = 0.69 \quad \text{and} \quad n \times V(X_i) = 12.8.$$

Hence,

$$V(\hat{\beta}_1) = \frac{\frac{0.69}{5-2}}{12.80} = 0.0180 \quad \Rightarrow \quad SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} = \sqrt{0.018} = 0.134$$

The required confidence level is 95%, with $\alpha = 5\%$ and $t^{n-2}_{\alpha/2} = t^3_{0.025} = 3.182$. So, the confidence interval is equal to:

$$[\hat{\beta}_1 \pm t^{n-2}_{\alpha/2} SE(\hat{\beta}_1)] = [0.59 \pm 3.182 \times 0.134] = [0.168; 1.020]$$

8. Testing the hypothesis that an increase of X, *Investments in Research and Development*, changes Y, the *Number of patents*, is equivalent to testing the following system of hypotheses:

$$\begin{cases} H_0: \ \beta_1 = 0 \\ H_1: \ \beta_1 \neq 0. \end{cases}$$

9. To calculate the p-value ofthe test in the previous point, it is first necessary to calculate the t-ratio (i.e. the observed value of the test statistic):

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.59}{0.134} = 4.437$$

In a two-sided test, the p-value is defined as:

$$2P(t_3 > 4.437)$$

The value 4.437 is between $t_{3,0.025} = 3.182$ and $t_{3,0.01} = 4.54$, so the p-value is between 2% and 5%.
At the 5% level, the null hypothesis is rejected and it is concluded that investments in RD have an effect on the number of patents.

Note that the confidence interval at the 95% confidence level does not contain zero.

10. In this case the t-ratio is:

$$t = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{0.59 - 1}{0.134} = -3.036$$

For a two-sided test, the p-value is defined as:

$$2P(t_3 > 3.036)$$

The value 3.0367 is between $t_{3,0.05} = 2.353$ and $t_{3,0.025} = 3.182$ so the p-value is between 10% and 5%. At the 5% level we do not reject the null hypothesis

and conclude that investments in RD have no effect on the number of patents. The same decision is made at the 1% level. Conversely, if a 10% confidence level were assumed, the null hypothesis would be rejected.

11. The confidence interval for the y-intercept is constructed in a similar way to that already seen for the slope.
With n large enough,

$$[\hat{\beta}_0 \pm Z_{1-\alpha/2} SE(\hat{\beta}_0)] \quad \text{where} \quad SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)} = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{n \times V(X_i)}\right) \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}}$$

Also in this case, $n < 30$, but thanks to the assumption of normality of the errors, we can proceed in a similar way, being careful only to use the t-Student distribution instead of the Normal distribution.

$$[\hat{\beta}_0 \pm t_{\alpha/2}^{n-2} SE(\hat{\beta}_0)]$$

Let's start from the calculation of $SE(\hat{\beta}_0)$.
We know:

$$\bar{x} = 7.2 \quad , \quad \sum_{i=1}^{n} \hat{u}_i^2 = 0.69 \quad \text{and} \quad n \times V(X_i) = 12.8.$$

Hence,

$$V(\hat{\beta}_0) = \left(\frac{1}{5} + \frac{7.2^2}{12.8}\right) \frac{0.69}{5-2} = 0.974 \quad \Rightarrow \quad SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)} = \sqrt{0.974} = 0.987$$

With $\alpha = 10\%$ and $t_{\alpha/2}^{n-2} = t_{0.05}^{3} = 2.353$, the 90% confidence interval is:

$$[\hat{\beta}_0 \pm t_{\alpha/2}^{n-2} SE(\hat{\beta}_0)] = [4.14 \pm 2.353 \times 0.987] = [1.802; 6.448]$$

12. It is possible to derive the t-ratio for β_1 as

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.4766}{0.1032} = 4.616$$

In the same way we can calculate β_0 as

$$t = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \Rightarrow \hat{\beta}_0 = t \times SE(\hat{\beta}_0) = 6.519 \times 2.7715 = 18.0687$$

With reference to the y-intercept it is possible to calculate the p-value as

$$2(1 - \Phi(|t|)) = 2 \times (1 - \Phi(|6.52|))2 \times (1 - 1) \simeq 0$$

Finally, since for both estimated coefficients the p-value is close to zero, it is possible to associate the significance code *** to both.

The regression output is:

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5061  -9.0745   0.8533   8.4412  30.9705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.0687     2.7715   6.519 3.08e-09 ***
x             0.4766     0.1032   4.616 1.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.87 on 98 degrees of freedom
Multiple R-squared:  0.1786,    Adjusted R-squared:  0.1702
F-statistic: 21.31 on 1 and 98 DF,  p-value: 1.185e-05
```

In this case, n is large enough, so the confidence intervals for the intercept and slope of the regression line are constructed as:

$$[\hat{\beta}_0 \pm Z_{1-\alpha/2} SE(\hat{\beta}_0)] \quad \text{and} \quad [\hat{\beta}_1 \pm Z_{1-\alpha/2} SE(\hat{\beta}_1)]$$

It follows that, setting the confidence level at 95%, $Z_{1-\alpha/2} = 1,96$ and the two required intervals are:

$$[\hat{\beta}_0 \pm Z_{1-\alpha/2} SE(\hat{\beta}_0)] = [18.0687 \pm 1.96 \times 2.7715] = [12.6; 23.50]$$

and

$$[\hat{\beta}_1 \pm Z_{1-\alpha/2} SE(\hat{\beta}_1)] = [0.4766 \pm 1.96 \times 0.1032] = [0.27; 0.68].$$