Exercises: Simple Linear Regression

Valentina Zangirolami - valentina.zangirolami@unimib.it

November, 2023

(Referring to the theoretical parts: 1, 2, 3, 6)

1 Mother and Daughter heights data

Let consider a sample of data with $n = 11$ observations (Table 1) with two variables:

- **mother's height** x (independent variable);
- **daughter's height** y (dependent variable).

Table 1: Mother and Daughter heights data: data are expressed in centimeters.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
x	153.7	156.7	173.5	157.0	161.8	140.7	179.8	150.9	154.4	162.3	166.6
y	163.1	159.5	169.4	158.0	164.3	150.0	170.3	158.9	161.5	160.8	160.6

We would like to find out if there exists a relationship between these two variables.

Exercise 1.1
Starting from the data (in Table 1), write the equation of the simple linear regression model. Compute $\bar{x}$, $\bar{y}$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$ and, then, find the estimates of the linear model parameters.

Given the above two variables, we can use the simple linear regression model as follows

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad , \quad i = 1, \dots, 11$$

where (β_1, β_2) are unknown parameters and $\mu_i = \mathbb{E}[Y_i \mid X_i = x_i]$ $= \beta_1 + \beta_2 x_i$ is the deterministic component. The error term ε_i is the stochastic component.

The simple linear model is based on three assumptions:

1) Absence of systematic error $\Rightarrow E[\varepsilon_i]=0$
which implies $E[Y_i]=\beta_1+\beta_2 x_i$ (linearity)

2) Homoscedasticity of the errors $\rightarrow Var(\varepsilon_i)=\sigma^2>0$
which implies $Var(Y_i)=\sigma^2>0$ (homoscedasticity of the response)

3) Uncorrelated errors $\Rightarrow Cov(\varepsilon_i,\varepsilon_k)=0$ for $i\neq k$
which implies $Cov(Y_i,Y_k)=0$ for $i=k$

The unknown quantities are $\theta=(\beta_1,\beta_2,\sigma^2)$ and the parameter space is $\Theta=\mathbb{R}^2\times\mathbb{R}^+$

From notes part 2, we can derive the equations of OLS estimates:

ⓐ $\hat{\beta}_1=\bar{y}-\hat{\beta}_2\bar{x}$

ⓑ $$\hat{\beta}_2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$$

Where $\bar{x}$ and $\bar{y}$ are the sample means.

Im this case,

$$\bar{x}=\frac{153.7+156.7+173.5+157+161.8+140.7+179.8+150.9+154.4+162.3+166.6}{11}=159.76$$

Same procedure for $\bar{y}=\sum_{i=1}^{11}y_i$ and $\bar{y}=161.49$

$\sum_{i=1}^{11}x_iy_i=284335$ and $\sum_{i=1}^{n}x_i^2=281941$

Using ⓑ, $$\hat{\beta}_2=\frac{284335-11\cdot159.76\cdot161.49}{281941-11\cdot(159.76)^2}=0.45473$$

Using ⓐ, $\hat{\beta}_1=161.49-0.45473\cdot159.76=88.842$

Exercise 1.2

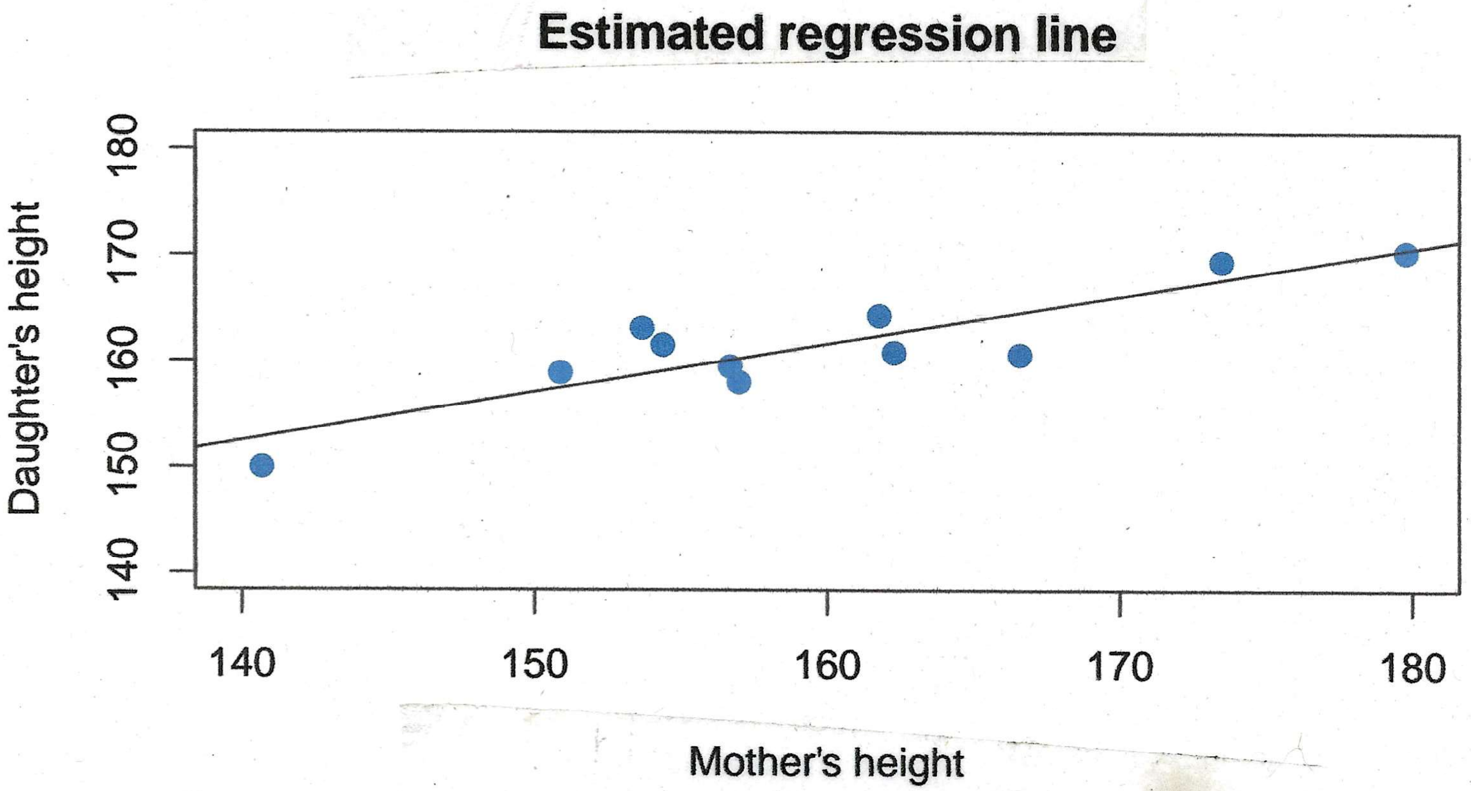
Given the results of the previous exercise, compute the fitted values for each i. Make a plot involving the observations of the couple (y, x) and the estimated regression line.

For each observation, the fitted values are obtained using the estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ (from ex. 1.1)

i	fitted values
1	$\hat{y}_1 = 88.842 + 0.45473 \cdot 153.7 = 158.734$
2	$\hat{y}_2 = 88.842 + 0.45473 \cdot 156.7 = 160.098$
3	$\hat{y}_3 = 88.842 + 0.45473 \cdot 173.5 = 167.738$
⋮	same procedure
	"
11	

$\hat{y}_4 = 160.235$ $\hat{y}_5 = 162.417$ $\hat{y}_6 = 152.823$ $\hat{y}_7 = 170.602$

$\hat{y}_8 = 157.461$ $\hat{y}_9 = 159.052$ $\hat{y}_{10} = 162.645$ $\hat{y}_{11} = 164.6$



Exercise 1.3

Compute the residuals (e_i) and the unbiased estimate (s^2) of the variance σ^2. Then, find the estimates of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$. (You can also use the notation of the notes part 6: $\sum_{i=1}^{n} e_i^2 = \ldots$)

To estimate the variance σ^2, we can use its estimator which corresponds to $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

However, this estimator is biased and we can considere the following unbiased estimator

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{n}{n-2} \hat{\sigma}^2$$

To find estimates, we can exploit the simplified equation for $\hat{\sigma}^2$.

In notes part 6, you saw that the sum of residuals can be expressed as

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and then holds

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_2^2 \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 =$$

$$= S_y^2 - \hat{\beta}_2^2 S_x^2 \quad \textcircled{1}$$

where S_y^2 and S_x^2 are the sample variances.

From $\textcircled{1}$, we have

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \bar{y}^2 - \hat{\beta}_2^2 \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right) =$$

$$= 26107 - (161.49)^2 - (0.45473)^2 \cdot (25631 - 159.76^2) = 5.7$$

and then we can compute s^2:

$$s^2 = \frac{11}{9} 5.7 = 6.97$$

Now, it is easy to find the estimated variances of $\hat{\beta}_1$ and $\hat{\beta}_2$.

From notes part 3, we can say that the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ correspond to

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \text{and} \quad Var(\hat{\beta}_1) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

and, then, replacing σ^2 with s^2

$$\hat{Var}(\hat{\beta}_2) = \frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{6.97}{1167} = 0.00595$$

$$\hat{Var}(\hat{\beta}_1) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) = 6.97\left(\frac{1}{11} + \frac{159.76^2}{1172}\right) = 1524$$

Exercise 1.4

Compute the total sum of squares (SST), the residual sum of squares (SSE) and the regression sum of squares (SSR). Then, find the coefficient of determination R^2. Compute the correlation coefficient r_{xy} and its squared. What happens in this case?

(NOTES PART 6)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and SST = SSR + SSE holds.

SST = 306.809   SSR = 240.5685   SSE = 66.24

As you already saw during the theoretical lectures, the coefficient of determination R^2 is the proportion of variability of the dependent variable that is predicted by the covariate:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \in [0,1]$$

$$R^2 = \frac{240.5685}{306.809} = 0.78$$

Thus, the model explains 78% of the total variability of the response.

Correlation coefficient

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = 0.885$$

$$r_{xy}^2 = (0.885)^2 = 0.78$$

=> In the case of the simple linear regression model,

$R^2 = r_{xy}^2$

Exercises: Gaussian Simple Linear Regression Part I

Valentina Zangirolami - valentina.zangirolami@unimib.it

November, 2023

(Referring to the theoretical parts: 4, 5, 6, 7)
(The results obtained in the previous practical part - simple linear regression - can be useful)

1 Mother and Daughter heights data

Let consider a sample of data with $n = 11$ observations (Table 1) with two variables:

- **mother's height** x (independent variable);
- **daughter's height** y (dependent variable).

Table 1: Mother and Daughter heights data: data are expressed in centimeters.

	1	2	3	4	5	6	7	8	9	10	11
x	153.7	156.7	173.5	157.0	161.8	140.7	179.8	150.9	154.4	162.3	166.6
y	163.1	159.5	169.4	158.0	164.3	150.0	170.3	158.9	161.5	160.8	160.6

We would like to find out if there exists a relationship between these two variables.

Exercise 1.5
Starting from the data (in Table 1), write the equation of the gaussian simple linear regression model together with the associated assumptions. Explain the difference from simple linear regression (make a comparison between the assumptions of ex. 1.1 and this case).

Given the above two variables, we can use the simple gaussian linear regression model as follows

$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad , \quad i = 1, \dots, 11$

where $\varepsilon_i \sim N(0, \sigma^2)$

The simple gaussian linear model is based on four assumptions:

1) Absence of systematic error $E[\varepsilon_i]=0$

2) Homoscedasticity of the errors $Var(\varepsilon_i)=\sigma^2>0$

3) Uncorrelated errors $Cov(\varepsilon_i,\varepsilon_k)=0$ for $i\neq k$

4) Gaussian distribution of the errors

Although the assumptions are similar to the previous case (ex. 11), gaussian linear regression required an additional assumption: gaussian distribution of the errors.

Exercise 1.6

Let consider the following system of hypothesis

$$\begin{cases} \text{H0: } \beta_2 = 1 \\ \text{H1: } \beta_2 \neq 1 \end{cases}$$

Compute the t-test and the p-value (use the t-table that you can find below).

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318

(NOTES PART 5)

To make inference about β_1 and β_2, you can use pivotal quantities as

$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{V}(\hat{\beta}_2)}} \quad , \text{where } \hat{V}(\hat{\beta}_2) = \frac{s^2}{\sigma^2} V(\hat{\beta}_2)$$

Since the exact distribution of $\hat{\beta}_1(Y)$ and $\hat{\beta}_2(Y)$ and $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$, the t-test corresponds to

$$t_2 = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{V}(\hat{\beta}_2)}} \overset{H_0}{\sim} t_{n-2}$$

In this case,

$$t_2^{oss} = \frac{\hat{\beta}_2 - 1}{\sqrt{\hat{V}(\hat{\beta}_2)}} = \frac{0.45473 - 1}{\sqrt{0.00595}} = -7.069$$

And the p-value is

$$\alpha^{oss} = 2\min\{\mathbb{P}(t_{n-2} \leq t_2^{oss}), \mathbb{P}(t_{n-2} \geq t_2^{oss})\} =$$

$$= 2\min\{\mathbb{P}(t_9 \leq -7.069), \mathbb{P}(t_9 \geq -7.069)\} =$$

$$= 2\min\{2.931 \times 10^{-5}, 1\} = 5.862 \times 10^{-5}$$

(In this case, I computed the exact p-value through the use of R softwar you can use the command pt(q=-7.069, df=9, lower.tail=TRUE) to find 2.931×10^{-5}

α^{oss} is below all the conventional levels of significance (i.e. $\alpha^{oss} < 0.1, \alpha^{oss} < 0.05, \alpha^{oss} < 0.$ which means we reject the null hypothesis.

Exercise 1.7

Let consider the following system of hypothesis

$$\begin{cases} \text{H0: } R^2 = 0 \\ \text{H1: } R^2 > 0 \end{cases}$$

Compute the F-test and the p-value (you can find the F-table here `https://faculty.washington.edu/heagerty/Books/Biostatistics/TABLES/F-Tables/`). In this case, does an equivalent test exists? Specify the hypothesis, the test statistic and compute the p-value.

(NOTES PART 7)

$$F = \frac{R^2}{1-R^2}(m-2) = \frac{SSR}{SSE}(m-2) \overset{Ho}{\sim} F_{1,m-2}$$

In the simple LM, you exploited the relationship with ~~te~~ T-test by proving

$$F = T^2 \overset{Ho}{\sim} F_{1,m-2}$$

where

$$T^2 = \frac{\hat{\beta}_2^2}{\hat{Var}(\hat{\beta}_2)} \overset{Ho}{\sim} F_{1,9}$$ under the system of hypothesis $\begin{cases} Ho: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$

Then,

$$\alpha^{oss} = \mathbb{P}_{Ho}(F \geq f^{obs}) = \mathbb{P}_{Ho}(T^2 \geq (t_{obs})^2) = \mathbb{P}(F_{1,9} \geq 34.7528) \overset{\circledast}{=} 0.0002304$$

$$t_{obs}^{2} = \frac{(0.45473)^2}{0.00595} = \frac{0.20678}{0.00595} = 34.7528$$

⊛ (You can use the R command 1-pf(34.7528, 1,9) to obtain the exact p-value

α^{oss} is below all the conventional levels of significance, ~~the~~. which means we reject the null hypothesis.

During the lectures, you proved $\mathbb{P}_{Ho}(F > f^{obs}) = \mathbb{P}_{Ho}(T^2 > (t_{obs})^2) = 2\mathbb{P}_{Ho}(T > |t_{obs}|)$

where $T \sim t_{m-2}$

An equivalent test is

$$\begin{cases} \beta_2 = 0 \\ \beta_2 \neq 0 \end{cases}$$

$$T = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} \overset{H_0}{\sim} t_{n-2}$$

$t_{oss} = -7.069$

$\alpha^{oss} = 2 \min \{ \mathbb{P}(t_{n-2} \leq t_{oss}), \mathbb{P}(t_{n-2} \geq t_{oss}) \} =$

$\overset{!}{=} 2 \min \{ 2.931 \times 10^{-5}, 1 \} = 5.862 \times 10^{-5}$ (see ex 1.6)

Exercise 1.8

Provide the confidence intervals for β_r, $r = 1, 2$ at level $1 - \alpha = 0.95$.

A confidence imterval for β_r can be obtained by

$$1-\alpha = \mathbb{P}\left(\hat{\beta}_r - t_{m-2;1-\alpha/2}\sqrt{\hat{V}(\hat{\beta}_r)} < \beta_r < \hat{\beta}_r + t_{m-2;1-\alpha/2}\sqrt{\hat{V}(\hat{\beta}_r)}\right)$$

~~Given~~ If $1-\alpha = 0.95$, then $\alpha = 0.05$ and $1 - \frac{\alpha}{2} = 0.975$

Using the t-table (you can find it in the ex1.6), $t_{9;0.975} = 2.262$

and then we can write our confidence interval (also using the estimated quantities from previous exercises).

confidence interval for $\hat{\beta}_1$:

$(88.842 - 2.262 \cdot \sqrt{152.423}, 88.842 + 2.262\sqrt{152.423}) = (60.915, 116.769)$

Confidence interval for $\hat{\beta}_2$:

$(0.45473 - 2.262 \cdot \sqrt{0.00595}, 0.45473 + 2.262\sqrt{0.00595}) = (0.280, 0.629)$

Exercise 1.9

During the theoretical lectures, you exploited the relationship between R^2 and the t-test to prove the equivalence among two statistical tests in the case of simple linear regression. Provide the formula and verify that it holds with the data. (In the exercise 1.4, you have already computed the R^2 and the components of ~~its decomposition)~~ deviance decomposition)

(NOTES PART 7)

$$t_2^2 = \frac{R^2}{1-R^2}(n-2) \quad *$$

$$t_2^2 = \frac{\hat{\beta}_2^2}{\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}}{\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{s^2}$$

$$\overset{\circledast}{=} \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}e_i^2/(n-2)} = \frac{SSR}{SSE/(n-2)} = \frac{R^2}{1-R^2}(n-2)$$

$$\circledast \quad \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_1-\hat{\beta}_2 x_i-\bar{y})^2 = \sum_{i=1}^{n}(\bar{y}-\hat{\beta}_2\bar{x}+\hat{\beta}_2 x_i-\bar{y})^2$$

$$= \hat{\beta}_2\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Then, we can also write

$$R^2 = \frac{t_2^2}{n-2+t_2^2} \approx 0.79$$

Exercises: Simple Gaussian Linear Regression Part II

Valentina Zangirolami - valentina.zangirolami@unimib.it

November, 2023

(Thanks to Dr. Roberto Ascari for providing these exercises)
(Referring to the theoretical parts: 4, 5, 6, 7)

2 Computer repair data

A computer repair company is interested in knowing the relationship between the duration of interventions (measured in minutes) and the number of electronic components to be replaced or repaired. Therefore, a simple linear regression model was considered to explain the duration in minutes of interventions (y) as a function of the number of units (x) to be replaced.

A sample of 14 interventions provided the following data: $\bar{y} = 95.768$, $\bar{x} = 6$, $\sum_{i=1}^{14}(y_i - \bar{y})^2 = 31108.357$, and $\sum_{i=1}^{14}(x_i - \bar{x})^2 = 114$. The model provides a coefficient of determination $R^2 = 0.984$.

Exercise 2.1

Starting from the data, compute the maximum likelihood estimates of β_1 and β_2. Then, write the equation of the estimated linear regression model.

We know

$$\hat{\beta}_2 = \frac{S_{xy}}{S_x^2} = \frac{\frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{m}(x_i-\bar{x})^2} = \frac{Cod(x,y)}{Dev(x)}$$

and we know the following relationship (which holds for any simple linear regression)

$$R^2 = r_{xy}^2 = \left(\frac{S_{xy}}{S_x S_y}\right)^2$$

where r_{xy} is the correlation coefficient.

$$S_{xy} = r_{xy}\sqrt{S_x^2 S_y^2} = \sqrt{R^2}\sqrt{S_x^2 S_y^2} = \sqrt{0.984}\sqrt{\frac{114}{14}\,\frac{31108.357}{14}} = 13343$$

$$Cod(x,y) = S_{xy}\cdot m = 133.43\cdot 14 = \overset{1}{1}868.05$$

Hence, $\hat{\beta}_2 = \frac{1868.05}{114} = 16.386$

$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 95.786 - (16.386 \cdot 6) = -2.532$

Then, the estimated model corresponds to $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i = -2.532 + 16.386 x_i$

Exercise 2.2

Find the estimate for the variance σ^2 using the decomposition of the total sum of squares. Through a valid test, verify the goodness of fit at 5% confidence level.

The unbiased estimate s^2 of σ^2 can be computed by using

$$s^2 = \frac{SSE}{n-2}$$

where SSE is the residual deviance.

Using the deviance decomposition, we know

$$SST = SSR + SSE$$

where $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ and $R^2 = \frac{SSR}{SST}$

Then,

$$SSR = SST \cdot R^2 = 31108.357 \cdot 0.984 = 30610.623$$

$$SSE = SST - SSR = 31108.357 - 30610.623 = 497.734$$

and finally we can compute

$$s^2 = \frac{497.734}{14-2} = 41.478$$

We can verify or check the goodness of fit using the F-test with the following system of hypothesis

$$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 > 0 \end{cases}$$

and then

$$F^{obs} = \frac{SSR/1}{SSE/(n-2)} = \frac{30610.623}{41.478} = 737.999$$

The p-value is

$$\alpha^{obs} = \mathbb{P}(F_{1,12} > 737.999) = 2\,3.810952 \cdot 10^{-12} < 0.05$$

$\Rightarrow$ we reject H0 : as we reject the hypothesis $R^2 = 0$ then our model has a value of $R^2 > 0$

Exercise 2.3

Given the standard errors (S.E.) of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$, which correspond to $\sqrt{\hat{Var}(\hat{\beta}_1)} = 4.014$ and $\sqrt{\hat{Var}(\hat{\beta}_2)} = 0.604$. Through a valid test (at 5 % confidence level), verify if the coefficients β_1 and β_2 are significant (you can use the following t-table for computing p-values).

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073

INFERENCE ON β_1

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

It is a two-side test. The t-statistic corresponds to

$$t^{obs} = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} = \frac{-2.532}{4.014} = -0.631 \quad \left(\text{we know } T = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \overset{H_0}{\sim} t_{m-2} \right)$$

Then, the p-value is

$$\alpha^{obs} = \mathbb{P}(|T_{m-2}| > |t^{obs}|) = 2 \cdot \min\{ \mathbb{P}(T_{12} \leq -0.631), \mathbb{P}(T_{12} > -0.631)\} =$$

$$= 0.5399 > 0.05$$

$\Rightarrow$ We cannot reject Ho : β_1 is not significant

(Let's check the t-table : $t_{12;0.25} = 0.695$ then $\alpha^{obs} \approx 2 \cdot 0.25 = 0.5$)

Inference on β_2

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases}$$

The test statistic corresponds to

$$t^{obs} = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} = \frac{16.386}{0.604} = 27.129$$

The p-value is

$$\mathbb{P}(|T_{n-2}| > |t^{obs}|) = 2 \cdot \min\{\mathbb{P}(T_{12} \leq 27.129), \mathbb{P}(T_{12} > 27.129)\} =$$

$$= 3.873092 \cdot 10^{-18} < 0.05$$

We reject H_0 : β_2 is significant

Exercise 2.4

Given the ex. 2.2, is there any statistical test in the exercise 2.3 that might be unnecessary?

Given the answer at the exercise 2.2, we could avoid one of the two test of exercise 2.3. Im particular, the test for $\hat{\beta}_2$ was unnecessary.

As you can verify at the exercise 1.7, in the case of simple linear regression we know that the hypothesis $H_0: R^2 = 0$ and $H_0: \beta_2 = 0$ are equivalent.

Indeed, $\underbrace{(27.129)^2}_{(t^{obs})^2} = 735.98 \approx \underbrace{737.999}_{(F^{obs})}$

(as we already proved in the exercise 1.7

3 Bacteria mortality data

Suppose we want to analyze bacterial mortality (y) as a function of radiation exposure (x). The output of a linear regression of y as a function of x is partially summarized in the table below:

Table 1: Output of a linear regression.

Variable	Coefficients	S.E.	T-value	P-value
Constant	49.162	22.76	2.16	0.05
Exposure (x)	-19.46	2.498	-7.79	<0.0001

where $n = 15$, $R^2 = 0.823$ and $\hat{\sigma} = 41.83$.

Exercise 3.1

Complete the Table 1 writing the equations you should use.

INFERENCE ON β_1

- t-value

$$t_1^{obs} = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} = \frac{49.162}{22.76} = 2.16$$

[system of hypothesis
$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$
]

- p-value

$$\alpha^{obs} = P_{H_0}(|T_{13}| > |t_1^{obs}|) = 2 \min\{P(T_{13} < 2.16), P(T_{13} > 2.16)\} = 0.05$$

[By looking the t table in the exercise 2.3, we have $t_{13;0.025} = 2.16$

~~Therea~~Hence, $\alpha^{obs} = 2 \cdot 0.025 = 0.05$]

- 1% SIGNIFICANCE LEVEL: $\alpha^{obs} > 0.01$ We cannot reject H_0: β_1 is not significant
- 10% SIGNIFICANCE LEVEL: $\alpha^{obs} < 0.05$ We can reject H_0: β_1 is significant

INFERENCE ON β_2

- Estimated standard error of β_2

$$\hat{SE}(\hat{\beta}_2) = \sqrt{\hat{Var}(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{t_2^{obs}} = \frac{-19.46}{-7.79} = 2.498$$

Exercise 3.2

Through a valid statistical test, evaluate the goodness of fit of the model.

To evaluate the goodness of fit, we can use the F-test.

System of hypothesis

$$\begin{cases} H_0: R^2 = 0 \\ \\ H_1: R^2 > 0 \end{cases}$$

Test statistic

$$F = \frac{SSR/1}{SSE/(m-2)} \overset{H_0}{\sim} F_{1,m-2}$$

To find f^{obs}, we need to calculate SSR and SSE.

$$SSE = (m-2)s^2 = (41.83)^2 \cdot 13 = 22746.7357$$

$$SST = \frac{SSE}{1-R^2} = \frac{22746.7357}{0.177} = 128512.6311$$

$$SSR = SST - SSE = 128512.6311 - 22746.7357 = 105765.8954$$

Hence,

$$f^{obs} = \frac{105765.8954}{22746.7357} = 60.446$$

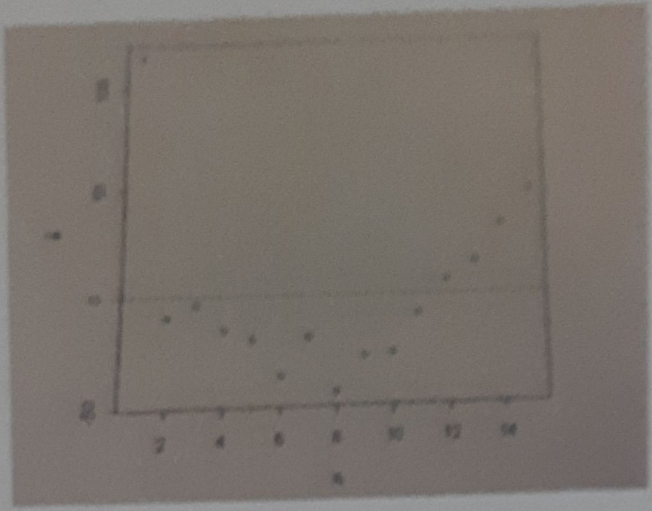
and the p-value is

$$\alpha^{obs} = \mathbb{P}(F_{1,13} > 60.446) = 3.052945 \cdot 10^{-6}$$

$\alpha^{obs} < 0.01 \Rightarrow$ We reject Ho

Exercise 3.3

Discuss about the hypothesis related to the model, looking the following residual plot.



At the beginning, we have to consider that the number of observations is not enough.

However, we can observe that the sum of residuals should be around zero (some of them are negative and others positive).

The plot highlights a systematic behavior (see notes part 8), which is characterized by a sort of dependence among errors.

Indeed, the plot suggests a non-linear relationship. The problem may be the model chosen.

Perhaps, involving transformations on covariates, we can solve the issues.

4 Grades data

In 2011 among 62 adolescents, the variables x "daily hours spent on average to video games" and y "average report card grade" were observed. We proposed a gaussian simple linear model with y, as response variable, and obtained the following estimates: $\hat{\beta}_1 = 7.4$, $\hat{\beta}_2 = -0.48$, $SSE = 223$, $\frac{Var(\hat{\beta}_1)}{\sigma^2} = 3.43$ and $\frac{Var(\hat{\beta}_2)}{\sigma^2} = 0.07$.

Exercise 4.1

Compute the OLS estimates for β_1 and β_2 and provide an explanation.

The exercise provides the maximum likelihood estimates $\hat{\beta}_1 = 7.4$ and $\hat{\beta}_2 = -0.48$.

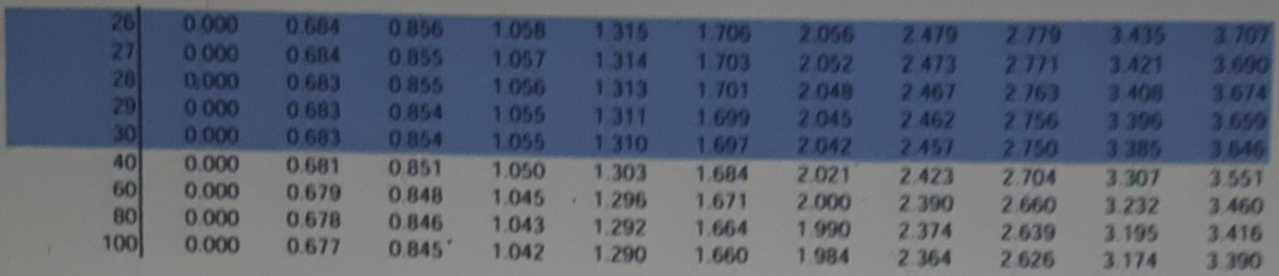
With the assumption of normality for the regression error term, OLS (Ordinary least square) corresponds to maximum likelihood (ML) estimation.

Hence, it implies

$$\hat{\beta}_1^{OLS} = \hat{\beta}_1^{ML} \quad \text{and} \quad \hat{\beta}_2^{OLS} = \hat{\beta}_2^{ML}$$

Exercise 4.2

Through a valid test, use p-values to evaluate if the coefficients are significant (you can use the t-table below for computing p-values). Then, evaluate the goodness of fit using p-values.



26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390

INFERENCE ON β_1

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

It is a two-side test.

TEST STATISTIC:

$$T_1 = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} \overset{H_0}{\sim} T_{m-2}$$

$$t_1^{obs} = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{S^2 \cdot 3.43}} \overset{(*)}{=} \frac{7.4}{\sqrt{3.7164 \cdot 3.43}} = 2.0726$$

$$(*)\ S^2 = \frac{SSE}{m-2} = \frac{223}{60} = 3.7167$$

$$\alpha^{obs} = \mathbb{P}_{H_0}(|T_{m-2}| > |t_1^{obs}|) = 2\min\{\mathbb{P}(T_{m-2} \leq 2.0726), \mathbb{P}(T_{m-2} > 2.0726)\} =$$

$= 0.0425$ (From the t-table, we could conclude p-value was around 0.05)

Indeed, $t_{60;0.025} = 2$. Then, $\alpha^{obs} \approx 2 \cdot 0.025 = 0.05$

- $\alpha^{obs} < 0.05 \Rightarrow$ we can reject H_0

INFERENCE ON β_2

$$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$$

$$t_2^{obs} = \frac{\hat{\beta}_2}{\hat{SE}(\hat{\beta}_2)} = \frac{-0.48}{\sqrt{3.7167 \cdot 0.07}} = -0.9411$$

p-value:

$$\alpha^{obs} = \mathbb{P}(|T_{n-2}| \geq -0.9411) = 2\min\{\mathbb{P}(T_{60} \leq -0.9411), \mathbb{P}(T_{60} > -0.9411)\} = 0.35$$

From the table, we can observe $t_{60;0.20} = 0.848$ and $t_{60;0.15} = 1.045$

Hence, $0.4 < \alpha^{obs} < 0.3$

- For each value in the interval $(0.3, 0.4)$, we can't reject H_0 at 1%, 5% and 10% significance level.

INFERENCE ON R^2

We can use the following equations to find SSR, SSE and then the f^{obs} (for performing the F-test).

We know that

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{n S_x^2}$$

From the text of this exercise, we have $\frac{Var(\hat{\beta}_2)}{\sigma^2} = \frac{1}{n S_x^2} = 0.07$

and we know that $\hat{Var}(\hat{\beta}_2) = \frac{S^2}{n S_x^2}$

- Exploiting the relationship between the t-test and the F-test (N.B. We can do it JUST FOR THE SIMPLE LINEAR REGRESSION), we know that

$$t_2^2 = f^{obs}$$

which means

$$\frac{\hat{\beta}_2^2}{\hat{Var}(\hat{\beta}_2)} = \frac{SSR}{SSE/(n-2)}$$

- We know $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = n\hat{\sigma}^2$,

where $\hat{\sigma}^2$ is the biased estimator of σ^2. If we use $S^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i$, we're using the unbiased estimator for σ^2

Hence,

$$SSR = \hat{\beta}_2^2 \cdot \frac{SSE}{m-2} \cdot \frac{1}{\widehat{Var}(\hat{\beta}_2)} = \hat{\beta}_2^2 \cdot \frac{m\hat{\sigma}^2}{(m-2)} \cdot \frac{mS_x^2}{S^2} = \hat{\beta}_2^2 \, mS_x^2 \, \frac{\cancel{m\hat{\sigma}^2}}{\cancel{(m-2)}} \, \frac{\cancel{(m-2)}}{\cancel{m\hat{\sigma}^2}} =$$

$$= (-0.48)^2 (0.07)^{-1} = 3.291.$$

$$f^{obs} = \frac{SSR/1}{SSE/(m-2)} = \frac{3.291}{223/60} = 0.885$$

The p-value is

$$\alpha^{obs} = \mathbb{P}(F_{1,60} > 0.885) = 0.3506$$

Indeed, $f_{1,60;\,0.65} = 0.88$ then $\alpha^{obs} \approx 0.35$

We cannot reject H_0 (corresponds to the t-test on β_2)

Finally, if you want to find the SST

$$SST = SSR + SSE = 3.291 + 223 = 226.291$$

5 Additional exercise

A linear regression model was estimated on 82 units. Complete the tables below and specify the hypothesis, the test statistic and p-value for inference.

Table 2: Analysis of variance.

Deviance	Sum of squares	d.o.f.	F	p-value
(SSR) Regression	3589.6	1	10.21	0.0019
(SSE) ~~Regression~~ Residual	28126.15	-	-	-
(SST) Total	31715.75	-	-	-

where d.o.f. means degree of freedom.

Table 3: Output of a linear regression.

Variable	Coefficients	S.E.	T-value	P-value
Constant	12.7	15.488	0.82	0.4142
x_1	-19.3	6.04	-3.195	0.002

$\boxed{1° \text{ TABLE}}$

- p-value for the F-test

$\alpha^{obs} = \mathbb{P}(F_{1,80} > 10.21) = 0.0014$ (We can reject Ho at 10%, 5% and 1% significance level)

- Find SSE

$$f^{obs} = 10.21 = \frac{SSR/1}{SSE/(n-2)} = \frac{3589.6}{SSE/80}$$

$$SSE = \frac{3589.6 \cdot 80}{10.21} = 28126.15$$

$SST = SSR + SSE = 31715.75$

$\boxed{2° \text{ TABLE}}$

- INTERCEPT

$$t_1^{obs} = 0.82 = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \Rightarrow \hat{SE}(\hat{\beta}_1) = \frac{\hat{\beta}_1}{0.82} = \frac{12.7}{0.82} = 15.488$$

p-value

$\alpha^{obs} = 2 \cdot \mathbb{P}(T_{80} \leq 0.82) = 0.4147$

(Indeed, $t_{80;0.2} = 0.846$ and hence $\alpha^{obs} \approx 0.4$)

We cannot reject Ho at 1%, 5% and 10% significance levels.

- inference on β_1

→ t-value

Given the p-value $\alpha^{obs} = 0.002 = 2 \cdot \mathbb{P}(T_{80} \leq q^{*})$

$\Rightarrow \mathbb{P}(T_{80} \leq q^{*}) = 0.001$

We can use the t table in the exercise 4.2, where $t_{80;0.001} = -3.195$

(The sign must be equal to the sign of the coefficient)

→ Standard error of β_2

$$\hat{SE}(\hat{\beta}_2) = \frac{-19.3}{-3.195} = 6.04$$