Exercises: Simple Gaussian Linear Model

Exercise 1, exam 24/09/2024

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected four women from each 10-year age group, beginning at age 40 and ending at age 79, and recorded their muscle mass index.
The observed values of age (x) and muscle mass (y) are:

unit	1	2	3	4	5	6	7	8
x	71	64	43	67	56	73	68	56
y	82	91	100	68	87	73	78	80
unit	9	10	11	12	13	14	15	16
x	76	65	45	58	45	53	49	78
y	65	84	116	76	97	100	105	77

Moreover, it is known that

$$\sum_{i=1}^{16} x_i = 967 \qquad \sum_{i=1}^{16} y_i = 1379$$

$$s_x^2 = 131.0625 \qquad s_y^2 = 202.2958 \qquad s_{xy} = \frac{1}{15}\sum_{i=1}^{16}(x_i - \bar{x})(y_i - \bar{y}) = -134.1542$$

where s_x^2 and s_y^2 are the unbiased estimates of the sample variances of x and y, respectively; and $\bar{x}$ and $\bar{y}$ are the sample means.
Assume that the following Gaussian linear model is appropriate:

$$\text{Model A:} \quad Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The estimates of the variances of the estimators are

$$\hat{var}(\hat{\beta}_1) = 133.63 \qquad \hat{var}(\hat{\beta}_2) = 0.03542$$

while the unbiased estimate of the variance σ^2 is

$$s^2 = 69.62.$$

Answer the following questions:

a) Write the expression of the estimated regression function.

b) Derive and explain the interpretation of the coefficient associated with the age variable.

c) Perform a statistical test to test the hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 < 0$.

d) Derive a 95% confidence interval of the age coefficient. Can you say anything about the significance of the coefficient?

e) Two new women "A" and "B" enter the study. Woman A is 38 while woman B is 60 years old. What is their predicted muscle mass according to the fitted model? What prediction has the largest uncertainty? Why?

f) Provide the definition of residuals. Obtain the value of the residual for the 8th observation. What is the value of the sum of the residuals for the specified model? Explain why.

g) Obtain the coefficient of determination R^2 and interpret it.

Exercise 2: Mother and Daughter heights data

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

- mother's height x_i, $i = 1, \ldots, n$ (independent variable);
- daughter's height y_i, $i = 1, \ldots, n$ (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 want to study the relationship between the two variables. Answer the following:

(a) Starting from the data in Table 1, write the equation of the Gaussian simple linear regression model and the associated assumptions.

(b) Knowing that

$$\bar{x} = 159.76, \qquad \bar{y} = 161.49, \qquad \sum_{i=1}^{n} x_i^2 = 281940.6$$

$$\sum_{i=1}^{n} y_i^2 = 287179.3, \qquad \sum_{i=1}^{n} x_i y_i = 284335.1,$$

compute the maximum likelihood estimates $(\hat{\beta}_1, \hat{\beta}_2)$

(c) Compute the unbiased estimate of the variance s^2.

(d) Perform a statistical test to test the following system of hypothesis

$$\begin{cases} \mathrm{H}_0 : \beta_2 = 1 \\ \mathrm{H}_1 : \beta_2 \neq 1 \end{cases}$$

(e) Obtain the confidence intervals for $\beta_r, r = 1, 2$ using a confidence level $1 - \alpha = 0.95$.

(f) 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 $\rho_{X,Y}$ and its squared. What happens in this case?

(g) Perform a statistical test to test the following system of hypothesis

$$\begin{cases} \mathrm{H}_0 : R^2 = 0 \\ \mathrm{H}_1 : R^2 > 0 \end{cases}$$

and compute the p-value. In the model under study, is there an equivalent test? Specify the set of hypotheses and the test statistic. What is the p-value following this procedure?

(h) We have seen the equivalence between the test about the significance of β_2 and the test about R^2 in the simple linear model. Provide the formula and verify that it holds with the data.

Exercise 3: Computer repair data

A computer repair company is interested in understanding the relationship between the number of electronic components to repair and the duration of the intervention (in minutes). Therefore, a simple linear regression model was fitted to study the duration (y) as a function of the number of repaired units (x).

A sample of $n = 14$ interventions provided the following data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 95.768, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 6,$$

$$\sum_{i=1}^{14}(y_i - \bar{y})^2 = 31108.357, \qquad \sum_{i=1}^{14}(x_i - \bar{x})^2 = 114$$

Moreover, the fitted model provides a coefficient of determination $R^2 = 0.984$.

(a) Compute the maximum likelihood estimates of β_1 and β_2. Then, write the equation of the estimated regression line.

(b) Find the estimate of the variance σ^2 using the decomposition of the total sum of squares. Through an adequate test, test the significance of the overall model using a 5% significance level.

(c) The estimates of the standard errors of the estimators $(\hat{B}_1, \hat{B}_2)$ are

$$\sqrt{\widehat{Var}(\hat{B}_1)} = 4.014, \qquad \sqrt{\widehat{Var}(\hat{B}_2)} = 0.604.$$

Through a valid test (using a 5% significance level), verify whether the coefficients β_1 and β_2 are significant.

(d) Is there any statistical test performed in (c) that was unnecessary, given the results of (a) and (b)?

Exercise 4: Bacteria mortality data

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

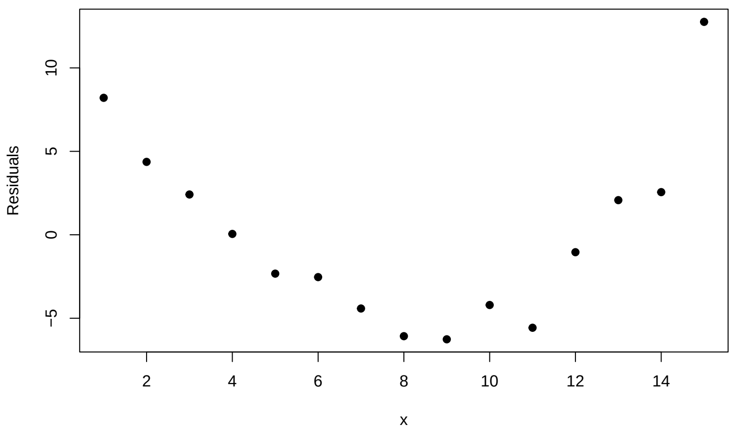
	Estimate	Std. Error	t value	p-value
(Intercept)	49.162	22.76	?	?
Exposure	-19.46	?	-7.79	< 0.0001

Moreover, it is known that $n = 15$, $R^2 = 0.823$ and $s = 41.83$.

(a) Complete the missing values in the table.

(b) Through a valid statistical test, evaluate the significance of the overall model.

(c) Given the following plot of the residuals against x, what can you say about the model assumptions? Were they reasonable?



	p						
	0.90	0.95	0.975	0.99	0.995	0.9975	0.999
z_p	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902

Table 2: Some quantiles of the Gaussian distribution: $p = \mathbb{P}(Z \leq z_p)$. Columns correspond to probabilities p.

	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$t_{1;p}$	3.0777	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088
$t_{3;p}$	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145
$t_{9;p}$	1.383	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968
$t_{10;p}$	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437
$t_{12;p}$	1.3562	1.7823	2.1788	2.681	3.0545	3.4284	3.9296
$t_{14;p}$	1.345	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874
$t_{16;p}$	1.3368	1.7459	2.1199	2.5835	2.9208	3.252	3.6862

Table 3: Some quantiles of the t distribution: $p = \mathbb{P}(T \leq t_{\alpha;p})$ with $T \sim t_\alpha$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom α.

	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$f_{1,8;p}$	3.4579	5.3177	7.5709	11.2586	14.6882	18.7797	25.4148
$f_{2,8;p}$	3.1131	4.459	6.0595	8.6491	11.0424	13.8885	18.4937
$f_{3,8;p}$	2.9238	4.0662	5.416	7.591	9.5965	11.9786	15.8295
$f_{4,8;p}$	2.8064	3.8379	5.0526	7.0061	8.8051	10.9407	14.3916
$f_{1,9;p}$	3.3603	5.1174	7.2093	10.5614	13.6136	17.1876	22.8571
$f_{2,9;p}$	3.0065	4.2565	5.7147	8.0215	10.1067	12.5392	16.3871
$f_{3,9;p}$	2.8129	3.8625	5.0781	6.9919	8.7171	10.7265	13.9018
$f_{4,9;p}$	2.6927	3.6331	4.7181	6.4221	7.9559	9.7411	12.5603
$f_{1,10;p}$	3.285	4.9646	6.9367	10.0443	12.8265	16.0363	21.0396
$f_{2,10;p}$	2.9245	4.1028	5.4564	7.5594	9.427	11.5723	14.9054
$f_{3,10;p}$	2.7277	3.7083	4.8256	6.5523	8.0807	9.8334	12.5527
$f_{4,10;p}$	2.6053	3.478	4.4683	5.9943	7.3428	8.8876	11.2828
$f_{1,11;p}$	3.2252	4.8443	6.7241	9.646	12.2263	15.1674	19.6868
$f_{2,11;p}$	2.8595	3.9823	5.2559	7.2057	8.9122	10.848	13.8116
$f_{3,11;p}$	2.6602	3.5874	4.63	6.2167	7.6004	9.1668	11.5611
$f_{4,11;p}$	2.5362	3.3567	4.2751	5.6683	6.8809	8.2521	10.3461
$f_{1,12;p}$	3.1765	4.7472	6.5538	9.3302	11.7542	14.4896	18.6433
$f_{2,12;p}$	2.8068	3.8853	5.0959	6.9266	8.5096	10.2865	12.9737
$f_{3,12;p}$	2.6055	3.4903	4.4742	5.9525	7.2258	8.6517	10.8042
$f_{4,12;p}$	2.4801	3.2592	4.1212	5.412	6.5211	7.7618	9.6327

Table 4: Some quantiles of the F distribution: $p = \mathbb{P}(F \leq f_{df_1,df_2;p})$ with $F \sim F_{df_1,df_2}$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom (df_1, df_2).