Statistical Modelling
Exam 24/09/2024

Exercise 1

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

Additionally, the following summary statistics are provided:

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

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

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

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

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 line.

b) Interpret the coefficient associated with the age variable.

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

d) Provide the definition of the residuals. Obtain the value of the residual for the 8th observation. What is the value of the sum of the residuals for the fitted model? Justify your answer.

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

f) 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? Which prediction has the largest uncertainty? Why?

g) An additional variable is then introduced, indicating whether the woman regularly exercises or not (1: yes; 0: no). Formulate an appropriate Gaussian linear model ("Model B") to study how muscle mass depends on age and physical activity.

h) The residual sum of squares of model B is equal to $SSE_B = 466.593$. Compute the coefficient of determination R^2 for Model B. Did you expect the R^2 of Model B to be larger or smaller than the R^2 of Model A? Why?

i) Conduct a statistical test (level $\alpha = 0.05$) to evaluate which of Models A and B is preferable.

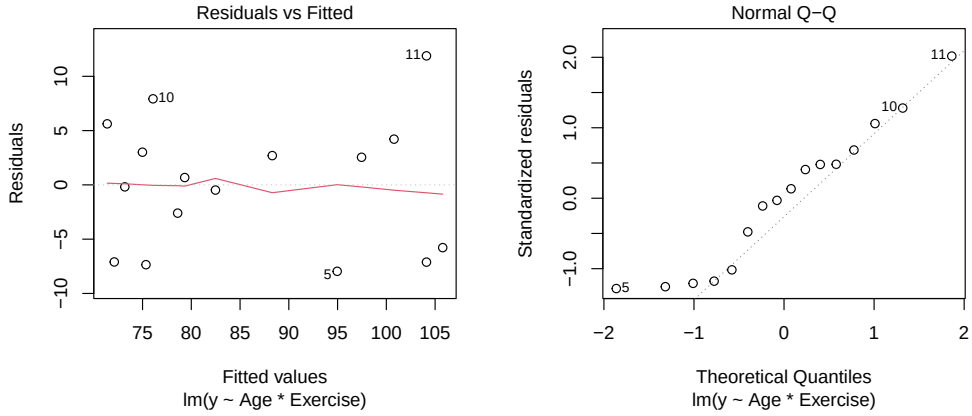
j) Specify a new model ("Model C") that includes an interaction effect between age and physical activity. Write the model formulation.

k) The output of fitting Model C to the data is as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	127.4640	10.0232	12.72	0.0000
Age	-0.7441	0.1567	-4.75	0.0005
Exercise	39.6505	24.5576	1.61	0.1324
Age:Exercise	-0.4713	0.4494	-1.05	0.3150

Write the expression for the regression function for women who exercise regularly and for those who do not. Provide a reasonable sketch of the two lines.

l) The figure below shows two plots for Model C. Explain what they represent and provide an interpretation.



Exercise 2

In a study of the hiring process of a company, the relationship between the outcome of a job interview (hired or not) and the age and gender of the individuals is of interest. Specifically, the outcome equals 1 if the person has been hired and 0 otherwise, the age variable is expressed in years, and the gender variable equals 1 if the individual is a man and 0 otherwise. Fitting a logistic regression produces the following result:

	Estimate	Std. Error	z value	Pr(>|z|)
`Intercept`	-2.0871	??	-7.695	??
`Gender 1`	0.2076	??	??	0.101
`Age`	-0.0394	0.0053	??	??

Null deviance:	136.66
Residual deviance:	114.61

a) Write the model formulation and assumptions for the fitted model.

b) What is the role of the link function in this model? Could the identity function be used instead? Why or why not?

c) Compute the missing values in the output (for the p-values, provide an approximation or a lower/upper bound). Which variables appear to have a significant effect? Comment on the results.

d) Perform a test of level $\alpha = 0.01$ to evaluate the overall significance of the model.