First name:

Last name:

Student ID number:

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

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

d) 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.

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? What prediction has the largest uncertainty? Why?

g) It is then introduced an additional variable 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 of 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 model is preferable between models A and B.

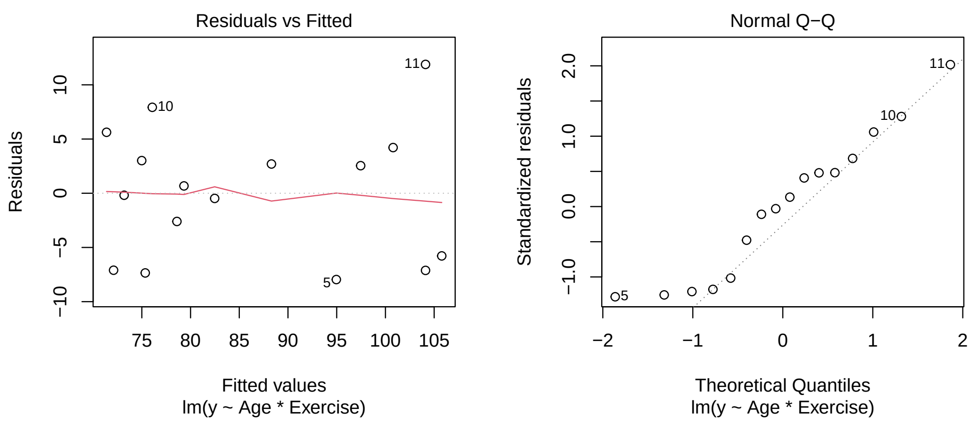
j) Specify a new model ("model C") which assumes that there is an interaction effect between age and physical activity. Write the model formulation.

k) The output of fitting model C to the data is the following:

	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 of the regression function for women who do exercise regularly, and the one for those who do not exercise. Make a (reasonable) sketch of the two lines.

l) The figure below shows two plots regarding model C. Explain what they represent and interpret them.



Exercise 2

In a study about the hiring process of a company, it is of interest to study the relationship between the outcome of a job interview (hired or not), and the age and gender of the individuals. In particular, the outcome takes value 1 if the person has been hired and 0 otherwise; the age variable is expressed in years; and the gender variable takes value 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 corresponding to such fitted model.

b) What is the role of the link function in the specified model? Would it be possible to use the identity function instead? Why?

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 on the response variable? Comment on the output.

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

		p						
		0.90	0.95	0.975	0.99	0.995	0.9975	0.999
standard Normal	z_p	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902
t with 18 d.o.f	$t_{18,p}$	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105
t with 17 d.o.f	$t_{17,p}$	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458
t with 16 d.o.f	$t_{16,p}$	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520	3.6862
t with 15 d.o.f	$t_{15,p}$	1.3406	1.7531	2.1314	2.6025	2.9467	3.2860	3.7328
t with 14 d.o.f	$t_{14,p}$	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874
χ^2 with 1 d.o.f	$\chi^2_{1,p}$	2.7055	3.8415	5.0239	6.6349	7.8794	9.1406	10.8276
χ^2 with 2 d.o.f	$\chi^2_{2,p}$	4.6052	5.9915	7.3778	9.2103	10.5966	11.9829	13.8155
χ^2 with 3 d.o.f	$\chi^2_{3,p}$	6.2514	7.8147	9.3484	11.3449	12.8382	14.3203	16.2662
χ^2 with 4 d.o.f	$\chi^2_{4,p}$	7.7794	9.4877	11.1433	13.2767	14.8603	16.4239	18.4668

Table 1: Some quantiles of Gaussian, Student's T and chi-squared distribution: $p = \mathbb{P}(X \leq q_p)$. Columns correspond to probabilities p. Rows correspond to different distributions, in particular, for the T and χ^2, each row corresponds to different degrees of freedom (d.o.f.).

	p						
	0.9000	0.9500	0.9750	0.9900	0.9950	0.9975	0.9990
$f_{1,17;p}$	3.0262	4.4513	6.0420	8.3997	10.3842	12.5525	15.7222
$f_{2,17;p}$	2.6446	3.5915	4.6189	6.1121	7.3536	8.7006	10.6584
$f_{3,17;p}$	2.4374	3.1968	4.0112	5.1850	6.1556	7.2053	8.7269
$f_{1,16;p}$	3.0481	4.4940	6.1151	8.5310	10.5755	12.8201	16.1202
$f_{2,16;p}$	2.6682	3.6337	4.6867	6.2262	7.5138	8.9179	10.9710
$f_{3,16;p}$	2.4618	3.2389	4.0768	5.2922	6.3034	7.4027	9.0059
$f_{1,15;p}$	3.0732	4.5431	6.1995	8.6831	10.7980	13.1328	16.5874
$f_{2,15;p}$	2.6952	3.6823	4.7650	6.3589	7.7008	9.1726	11.3391
$f_{3,15;p}$	2.4898	3.2874	4.1528	5.4170	6.4760	7.6343	9.3353
$f_{1,14;p}$	3.1022	4.6001	6.2979	8.8616	11.0602	13.5026	17.1434
$f_{2,14;p}$	2.7265	3.7389	4.8567	6.5149	7.9216	9.4748	11.7789
$f_{3,14;p}$	2.5222	3.3439	4.2417	5.5639	6.6804	7.9097	9.7294
$f_{1,13;p}$	3.1362	4.6672	6.4143	9.0738	11.3735	13.9468	17.8154
$f_{2,13;p}$	2.7632	3.8056	4.9653	6.7010	8.1865	9.8392	12.3127
$f_{3,13;p}$	2.5603	3.4105	4.3472	5.7394	6.9258	8.2424	10.2089
$f_{1,12;p}$	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

Table 2: Some quantiles of the F distribution: $p = \mathbb{P}(X \leq f_{df_1,df_2;p})$. Columns correspond to probabilities p. Rows correspond to different distributions, in particular, each row corresponds to different degrees of freedom (df).