First name:

Last name:

Student ID number:

Statistical Modelling
Exam 22/02/2024

Exercise 1

At admission to a college an entrance test is administered to 20 randomly selected students. The study aims to determine whether a student's grade point average (GPA) at the end of the first year (y) can be predicted from the entrance test score (x). Assume that a Gaussian linear model $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ is fitted. The observed values of the entrance test and of the GPA for the students are:

unit	1	2	3	4	5	6	7	8	9	10
x	5.50	4.80	4.70	3.90	4.50	6.20	6.00	5.20	4.70	4.30
y	3.10	2.30	3.00	1.90	2.50	3.70	3.40	2.60	2.80	1.60
unit	11	12	13	14	15	16	17	18	19	20
x	4.90	5.40	5.00	6.30	4.60	4.30	5.00	5.90	4.10	4.70
y	2.00	2.90	2.30	3.20	1.80	1.40	2.00	3.80	2.20	1.50

From the data it is possible to compute the following useful quantities:

$$\sum_{i=1}^{20} x_i = 100.00 \qquad \sum_{i=1}^{20} y_i = 50 \qquad \sum_{i=1}^{20} x_i^2 = 509.12 \qquad \sum_{i=1}^{20} y_i^2 = 134.84 \qquad \sum_{i=1}^{20} x_i y_i = 257.66$$

Moreover, from fitting the model we know that

$$\hat{var}(\hat{\beta}_1(Y)) = 0.7268^2 \qquad \hat{var}(\hat{\beta}_2(Y)) = 0.1440^2$$

$$\sum_{i=1}^{20} (y_i - \hat{y}_i)^2 = 3.4063$$

a) State the assumptions on ε_i, $i = 1, \ldots, 20$.

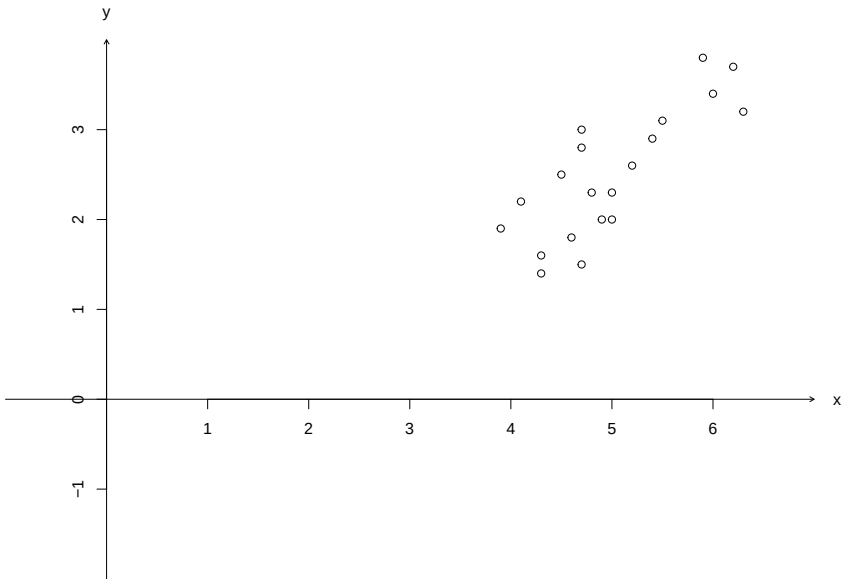
b) Obtain the maximum likelihood estimates of (β_1, β_2) and interpret them.

c) Write the expression of the estimated regression function and plot it (in the figure below).

d) Obtain a 0.99 confidence interval for β_2. Does it include zero? Why might one be interested in whether the confidence interval includes zero?

e) Is β_2 statistically significant using a significance level of 1%?

f) Two new students "A" and "B" take the test. The result of the entrance test of student "A" is 5.0, the result of the entrance test of student "B" is 6.5. Imagine you want to estimate their mean GPA using a 0.95 confidence interval. What student do you expect to have the wider confidence interval? Why?



Exercise 2

The `titanic` dataset is a collection of data about 714 passengers, and the goal is to predict the survival (`Survival`: 1 if the passenger survived, 0 if they did not) based on some personal characteristics. In particular, here we consider the ticket class (`Class`: 1 = first, 0 = second or third; dummy), the gender (`Gender`: man = 1, woman = 0; dummy), and the age (`Age`, in years). Fitting a logistic regression model in R produces the following summary:

	Estimate	Std. Error	z value	Pr(>|z|)
`(Intercept)`	1.5003	0.2462	6.09	0.0000
`Class`	2.0103	0.2479	8.11	0.0000
`Gender`	-2.5473	0.2017	??	0.0000
`Age`	-0.0299	0.0074	-4.06	??
Null deviance:	964.52			
Residual deviance:	675.14			

a) Write the corresponding theoretical model and the expression of the estimated model.

b) Write the likelihood and log-likelihood function.

c) Complete the missing values in the table. For Pr(>|z|) of `Age`, write an approximate value. What variables are statistically significant?

d) Provide an estimate of the odds for a woman aged 30 with a ticket of first class (denote this individual as "A"). How do you expect this value to change if you consider a person with the same characteristics but aged 31 (denote this individual as "B")?

e) Provide the interpretation of the coefficient associated with the `Class` variable. Given the estimate of this coefficient, what is the effect of this covariate on the survival probability?

f) Perform a test $H_0 : \beta_{\text{class}} = 0$ vs $H_1 : \beta_{\text{class}} < 0$.

g) Define the "null deviance" and "residual deviance" in the output.

h) Perform a test about 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 20 df	$t_{20,p}$	1.3253	1.7247	2.0860	2.5280	2.8453	3.1534	3.5518
t with 19 df	$t_{19,p}$	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794
t with 18 df	$t_{18,p}$	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105
t with 17 df	$t_{17,p}$	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458
t with 16 df	$t_{16,p}$	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520	3.6862
t with 15 df	$t_{15,p}$	1.3406	1.7531	2.1314	2.6025	2.9467	3.2860	3.7328
t with 14 df	$t_{14,p}$	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874
χ^2 with 1 df	$\chi^2_{1,p}$	2.7055	3.8415	5.0239	6.6349	7.8794	9.1406	10.8276
χ^2 with 2 df	$\chi^2_{2,p}$	4.6052	5.9915	7.3778	9.2103	10.5966	11.9829	13.8155
χ^2 with 3 df	$\chi^2_{3,p}$	6.2514	7.8147	9.3484	11.3449	12.8382	14.3203	16.2662
χ^2 with 4 df	$\chi^2_{4,p}$	7.7794	9.4877	11.1433	13.2767	14.8603	16.4239	18.4668
χ^2 with 5 df	$\chi^2_{5,p}$	9.2364	11.0705	12.8325	15.0863	16.7496	18.3856	20.5150

Table 1: Some quantiles of Gaussian, Student's t, and χ^2 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 (df).