First name:

Last name:

Student ID number:

Statistical Modelling
Exam 03/09/2024

Exercise 1

Consider the following models, for $i = 1, \dots, n$

1. $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 \log_{10} x_{i,3} + \beta_4 x_{i,4}^2 + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$ independent.
2. $Y_i = \dfrac{\beta_1 + \beta_2 x_{i,2}}{\beta_3 x_{i,1}} + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$ independent.
3. $\log(Y_i) = \dfrac{\beta_2 x_{i,1} + \beta_3 \log(x_{i,3})}{x_{i,2}} + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$ independent.
4. $Y_i = \beta_1 \, {x_{i,2}}^{\beta_2} \exp\{\varepsilon_i\}$ and $\varepsilon_i \sim N(0, 1)$ independent.

Answer the following questions:

a) For each model, indicate whether it is a linear regression model. If it is not, explain why and whether it can be expressed in the form $Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i$ by a suitable transformation and write explicitly such transformation.

b) Consider model 4 appropriately transformed, denoting with Y^*, $x^*_{i,2}$, (β_1^*, β_2^*) and ε_i^* the transformed quantities. Express it in the matrix form $\underline{Y}^* = X^* \underline{\beta}^* + \underline{\varepsilon}^*$, explicitly stating $\underline{Y}^*$ (and its distribution), X^*, $\underline{\beta}^*$, and $\underline{\varepsilon}^*$.

c) Write the expression of the maximum likelihood estimator $\underline{\hat{\beta}}^*$ and its exact distribution.

d) Let $\underline{e} = \underline{y}^* - X^* \underline{\hat{\beta}^*}$ be the vector of the residuals. State which of the following identities are satisfied and motivate the answer:

$$\sum_{i=1}^n e_i = 0 \qquad \sum_{i=1}^n e_i x_{i,2} = 0$$

$$\sum_{i=1}^n e_i \log(x_{i,2}) = 0 \qquad \sum_{i=1}^n e_i \log(x_{i,2}^2) = 0$$

Exercise 2

A corporation sells computer parts and performs maintenance and repair service. The data below have been collected from 18 recent calls to perform maintenance service; for each call, x_1 is the number of repairs and y is the total number of minutes spent by the service person. Moreover, it is also available the information about the type of computer: in particular, the first 12 observations refer to business computers, while the last 6 refer to personal computers. Specifically,

business computers

i	1	2	3	4	5	6	7	8	9	10	11	12
number of repairs	7	6	5	5	4	7	7	4	2	8	5	5
total minutes	97	86	78	75	62	101	105	53	33	118	65	71

personal computers

i	13	14	15	16	17	18
number of repairs	2	1	3	1	4	3
total minutes	25	10	39	17	49	28

with

$$\sum_{i=1}^{18} y_i = 1112 \qquad \sum_{i=1}^{18} x_{i,1} = 79$$

$$s_y^2 = 1040.889 \qquad s_{x_1}^2 = 4.4869$$

$$R^2 = 0.9808$$

a) Formulate an appropriate Gaussian linear model ("model A") to study how the total minutes of intervention depend on the number of repairs and the type of computer. Write the model formulation and assumptions.

b) State the decomposition of the sum of squares and specify each term for the fitted model.

c) Test the hypothesis about the overall significance of the model and interpret its result.

d) Specify a new model ("model B") which assumes that there is an interaction effect between the number of repairs and the type of computer. Write the model formulation.

e) Can you choose between the two models (model A and model B) by simply comparing their coefficients of determination R^2? Explain.

f) In the model that includes the interaction term (model B) it is obtained a residual (error) sum of squares equal to $SSE_B = 285.657$. Perform a test to compare this model with model A using a significance level 0.05. Which model do you prefer?

Exercise 3

Consider an experiment to study the resistance to the tension of a machine component. The dataset studies how many breaks occurred during 54 replications of the experiment for two types of material (A and B) and different levels of tension (L = low; M = medium; H = high).
To study such relationship we fit a Poisson regression model. The output of the model is the following:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.6920	0.0454	81.30	0.0000
material B	-0.2060	0.0516	-3.99	0.0001
tension M	-0.3213	0.0603	-5.33	0.0000
tension H	-0.5185	0.0640	-8.11	0.0000

Null deviance: 297.37 on 53 degrees of freedom
Residual deviance: 210.39 on 50 degrees of freedom

a) Write the model formulation and assumptions.

b) Derive and explain the interpretation of the coefficient associated with the variable "material B".

c) A second model ("model B") assumes that the type of material and the level of tension do not have an impact on the number of breaks. Specify the model and perform a test to compare the model fitted in point (a) with model B.

		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
$f_{1,18;p}$	3.0070	4.4139	5.9781	8.2854	10.2181	12.3208	15.3793
$f_{2,18;p}$	2.6239	3.5546	4.5597	6.0129	7.2148	8.5130	10.3899
$f_{3,18;p}$	2.4160	3.1599	3.9539	5.0919	6.0278	7.0351	8.4875
$f_{1,17;p}$	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

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).