Statistical Modelling
Exam 03/09/2024

Exercise 1

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

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

Answer the following questions:

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

b) Consider Model 4 after an appropriate transformation, and denote the transformed quantities by Y_i^*, $x_{i,2}^*$, (β_1^*, β_2^*), and ε_i^*.
Express the model in matrix form $\boldsymbol{Y}^* = X^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*$, explicitly stating $\boldsymbol{Y}^*$ X^*, $\boldsymbol{\beta}^*$, and $\boldsymbol{\varepsilon}^*$.
Write the distribution of $\boldsymbol{Y}^*$ and $\boldsymbol{\varepsilon}^*$.

c) Write the expression for the maximum likelihood estimator $\hat{\boldsymbol{B}}^*$ and its exact distribution.

d) Let $\boldsymbol{e} = \boldsymbol{y}^* - X^* \hat{\boldsymbol{\beta}}^*$ be the vector of the residuals. State which of the following identities are satisfied and justify your 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 services. The data below were collected from 18 recent maintenance service calls; for each call, x_1 denotes the number of repairs and y the total number of minutes spent by the technician. Moreover, information about the type of computer is also available: in particular, the first 12 observations refer to business computers, while the last 6 refer to personal computers. The data are as follows:

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

Additionally, the following summary statistics are provided:

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

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

$$R^2 = 0.9808$$

a) Formulate an appropriate Gaussian linear model ("Model A") to study how the total service time depends 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) Perform a test of overall significance and interpret its result.

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

e) Can you choose between Model A and Model B by simply comparing their coefficients of determination R^2? Justify your answer.

f) For Model B, the residual sum of squares is $SSE_B = 285.657$. Perform a test to compare this model with Model A using a significance level of 0.05. Which model do you prefer?

Exercise 3

Consider an experiment to study the resistance of a machine component to tension. The dataset records the number of breaks observed in 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 this relationship, we fit a Poisson regression model. The output of the model is as follows:

	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 its assumptions.

b) Interpret 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 have no effect on the number of breaks. Specify the model and perform a test to compare the model fitted in point (a) with model B.