Exercises: Multiple Gaussian Linear Model

Exercise 1: Exam 03/09/2024

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 $\boldsymbol{Y}^* = X^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*$, explicitly stating $\boldsymbol{Y}^*$ (and its distribution), X^*, $\boldsymbol{\beta}^*$, and $\boldsymbol{\varepsilon}^*$.

c) Write the expression of 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 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: Exam 25/01/2024

The data contained in the `cement` dataset represent the hardness (`hardness` variable) of 13 types of cement with different chemical compositions. Specifically, each type is obtained with varying proportions of aluminium (`aluminium` variable), silicate (`silicate` variable), calcium aluminoferrite (`aluminium_ferrite`), and silicate bic (`silicate_bic`). The interest is explaining how the hardness of cement depends on the proportions of chemicals.

A regression model was fitted for this purpose and produced the following result:

	Estimate	Std. Error	t statistic	Pr(>|t|)
`(Intercept)`	124.4809	26.7557	4.653	0.0016
`aluminium`	0.9739	??	3.435	0.0089
`silicate`	-0.1405	0.2891	-0.486	0.6400
`aluminium_ferrite`	-0.4974	0.2751	??	??
`silicate_bic`	??	0.3214	-2.481	0.0381

Error sum of squares	49.378
Total sum of squares	2715.763
R^2 coefficient	??

a) Write the model formulation and assumptions.

b) Complete the missing values in the table. For "Pr(> |t|)" of `aluminium_ferrite` provide an approximate value. What variables have a statistically significant effect?

c) Test the statistical hypothesis corresponding to the statement "the covariates do not have an effect on the hardness of cement".

d) On a reduced model ("model B") that includes only the variables `aluminium` and `silicate_bic` the error sum of squares is equal to $SSE_B = 74.762$. Perform an F test to compare this model with the complete model ("model A") that includes all the covariates. Interpret the result: which model would you prefer?

e) Obtain the coefficient R^2 of model B. Instead of performing the test in point (d), could you have simply compared the coefficient R^2 of the two models to choose between them? Why?

f) Figure 1 shows two plots regarding the complete model (model A). Explain what they represent and interpret them.

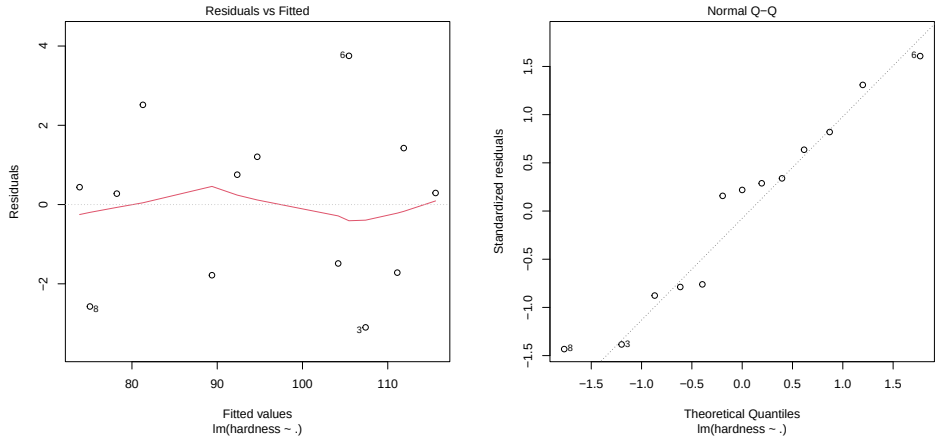

Figure 1:

	p						
	0.90	0.95	0.975	0.99	0.995	0.9975	0.999
z_p	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902

Table 1: Some quantiles of the Gaussian distribution: $p = \mathbb{P}(Z \leq z_p)$. Columns correspond to probabilities p.

	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$t_{1;p}$	3.0777	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088
$t_{3;p}$	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145
$t_{8;p}$	1.3968	1.8595	2.306	2.8965	3.3554	3.8325	4.5008
$t_{9;p}$	1.383	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968
$t_{10;p}$	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437
$t_{12;p}$	1.3562	1.7823	2.1788	2.681	3.0545	3.4284	3.9296
$t_{13;p}$	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	3.852
$t_{14;p}$	1.345	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874
$t_{15;p}$	1.3406	1.7531	2.1314	2.6025	2.9467	3.286	3.7328
$t_{16;p}$	1.3368	1.7459	2.1199	2.5835	2.9208	3.252	3.6862

Table 2: Some quantiles of the t distribution: $p = \mathbb{P}(T \leq t_{\alpha;p})$ with $T \sim t_\alpha$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom α.

	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$f_{1,8;p}$	3.4579	5.3177	7.5709	11.2586	14.6882	18.7797	25.4148
$f_{2,8;p}$	3.1131	4.459	6.0595	8.6491	11.0424	13.8885	18.4937
$f_{3,8;p}$	2.9238	4.0662	5.416	7.591	9.5965	11.9786	15.8295
$f_{4,8;p}$	2.8064	3.8379	5.0526	7.0061	8.8051	10.9407	14.3916
$f_{1,9;p}$	3.3603	5.1174	7.2093	10.5614	13.6136	17.1876	22.8571
$f_{2,9;p}$	3.0065	4.2565	5.7147	8.0215	10.1067	12.5392	16.3871
$f_{3,9;p}$	2.8129	3.8625	5.0781	6.9919	8.7171	10.7265	13.9018
$f_{4,9;p}$	2.6927	3.6331	4.7181	6.4221	7.9559	9.7411	12.5603
$f_{1,10;p}$	3.285	4.9646	6.9367	10.0443	12.8265	16.0363	21.0396
$f_{2,10;p}$	2.9245	4.1028	5.4564	7.5594	9.427	11.5723	14.9054
$f_{3,10;p}$	2.7277	3.7083	4.8256	6.5523	8.0807	9.8334	12.5527
$f_{4,10;p}$	2.6053	3.478	4.4683	5.9943	7.3428	8.8876	11.2828
$f_{1,11;p}$	3.2252	4.8443	6.7241	9.646	12.2263	15.1674	19.6868
$f_{2,11;p}$	2.8595	3.9823	5.2559	7.2057	8.9122	10.848	13.8116
$f_{3,11;p}$	2.6602	3.5874	4.63	6.2167	7.6004	9.1668	11.5611
$f_{4,11;p}$	2.5362	3.3567	4.2751	5.6683	6.8809	8.2521	10.3461
$f_{1,12;p}$	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
$f_{4,12;p}$	2.4801	3.2592	4.1212	5.412	6.5211	7.7618	9.6327

Table 3: Some quantiles of the F distribution: $p = \mathbb{P}(F \leq f_{df_1,df_2;p})$ with $F \sim F_{df_1,df_2}$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom (df_1, df_2).