First name:

Last name:

Student ID number:

Statistical Modelling
Exam 25/01/2024

Exercise 1

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? Why?

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

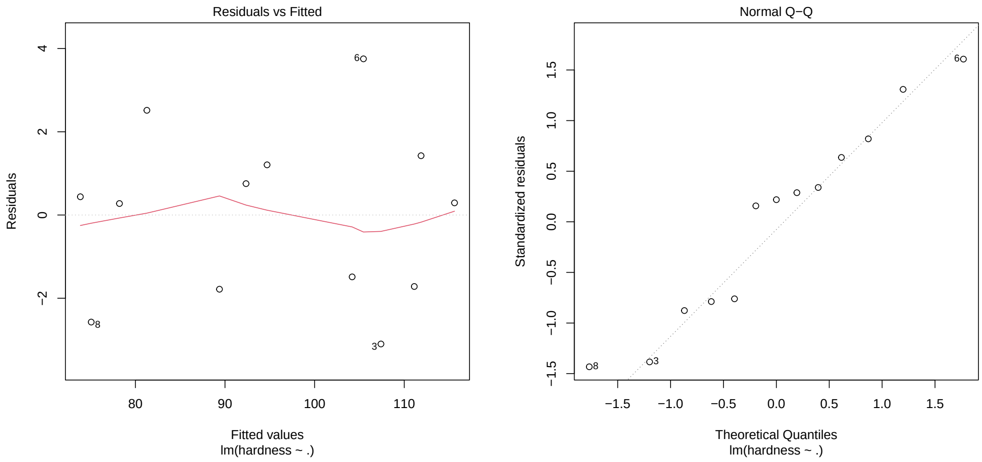

Figure 1:

Exercise 2

Let $(y_1, \ldots, y_5)$ and $(y_6, \ldots, y_{10})$ be two independent samples from a Poisson distribution of mean $\exp\{\beta_1\}$ and from a Poisson distribution of mean $\exp\{\beta_1 + \beta_2\}$, respectively.

a) Formulate an appropriate Poisson regression model for the expected value of Y_i, $i = 1, \ldots, 10$.

b) Write the log-likelihood function of $\underline{\beta} = (\beta_1, \beta_2)$ and the score function. Find the maximum likelihood estimate of (β_1, β_2). Finally, obtain the observed information matrix.

c) Determine an approximate distribution of the maximum likelihood estimator $\underline{\hat{\beta}}$ of $\underline{\beta} = (\beta_1, \beta_2)$, and an approximate distribution of the maximum likelihood estimator $\hat{\beta}_1$ of β_1.

d) Provide the interpretation of the coefficient β_2.

e) Define the concept of "saturated model" and obtain the expression of maximum of the log-likelihood for this 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 4 df	$t_{4,p}$	1.5332	2.1318	2.7764	3.7469	4.6041	5.5976	7.1732
t with 5 df	$t_{5,p}$	1.4759	2.0150	2.5706	3.3649	4.0321	4.7733	5.8934
t with 6 df	$t_{6,p}$	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.2076
t with 7 df	$t_{7,p}$	1.4149	1.8946	2.3646	2.9980	3.4995	4.0293	4.7853
t with 8 df	$t_{8,p}$	1.3968	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008
t with 9 df	$t_{9,p}$	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968
t with 10 df	$t_{10,p}$	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437
t with 11 df	$t_{11,p}$	1.3634	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247
t with 12 df	$t_{12,p}$	1.3562	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296
t with 13 df	$t_{13,p}$	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520

Table 1: Some quantiles of Gaussian and Student's t distribution: $p = \mathbb{P}(X \leq q_p)$. Columns correspond to probabilities p. Rows correspond to different distributions, in particular, for the t, each row corresponds to different degrees of freedom (df).

	0.90	0.95	0.975	0.99	0.995	0.9975	0.999
$f_{1,4;p}$	4.5448	7.7086	12.2179	21.1977	31.3328	45.6740	74.1373
$f_{1,5;p}$	4.0604	6.6079	10.0070	16.2582	22.7848	31.4067	47.1808
$f_{1,8;p}$	3.4579	5.3177	7.5709	11.2586	14.6882	18.7797	25.4148
$f_{1,13;p}$	3.1362	4.6672	6.4143	9.0738	11.3735	13.9468	17.8154
$f_{2,4;p}$	4.3246	6.9443	10.6491	18.0000	26.2843	38.0000	61.2456
$f_{2,5;p}$	3.7797	5.7861	8.4336	13.2739	18.3138	24.9640	37.1223
$f_{2,8;p}$	3.1131	4.4590	6.0595	8.6491	11.0424	13.8885	18.4937
$f_{2,13;p}$	2.7632	3.8056	4.9653	6.7010	8.1865	9.8392	12.3127
$f_{4,4;p}$	4.1072	6.3882	9.6045	15.9770	23.1545	33.3027	53.4358
$f_{4,5;p}$	3.5202	5.1922	7.3879	11.3919	15.5561	21.0478	31.0850
$f_{4,8;p}$	2.8064	3.8379	5.0526	7.0061	8.8051	10.9407	14.3916
$f_{4,13;p}$	2.4337	3.1791	3.9959	5.2053	6.2335	7.3728	9.0727
$f_{5,4;p}$	4.0506	6.2561	9.3645	15.5219	22.4564	32.2609	51.7116
$f_{5,5;p}$	3.4530	5.0503	7.1464	10.9670	14.9396	20.1783	29.7524
$f_{5,8;p}$	2.7264	3.6875	4.8173	6.6318	8.3018	10.2834	13.4847
$f_{5,13;p}$	2.3467	3.0254	3.7667	4.8616	5.7910	6.8200	8.3541
$f_{8,4;p}$	3.9549	6.0410	8.9796	14.7989	21.3520	30.6167	48.9962
$f_{8,5;p}$	3.3393	4.8183	6.7572	10.2893	13.9610	18.8022	27.6495
$f_{8,8;p}$	2.5893	3.4381	4.4333	6.0289	7.4959	9.2358	12.0455
$f_{8,13;p}$	2.1953	2.7669	3.3880	4.3021	5.0761	5.9318	7.2061
$f_{13,4;p}$	3.8859	5.8911	8.7150	14.3065	20.6027	29.5042	47.1627
$f_{13,5;p}$	3.2567	4.6552	6.4876	9.8248	13.2934	17.8667	26.2240
$f_{13,8;p}$	2.4876	3.2590	4.1622	5.6089	6.9384	8.5146	11.0596
$f_{13,13;p}$	2.0802	2.5769	3.1150	3.9052	4.5733	5.3113	6.4094

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).