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 alumino-ferrite (`aluminium_ferrite`), and silicate bic (`silicate_bic`). The interest is in 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 (call it "Model A") :

	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.

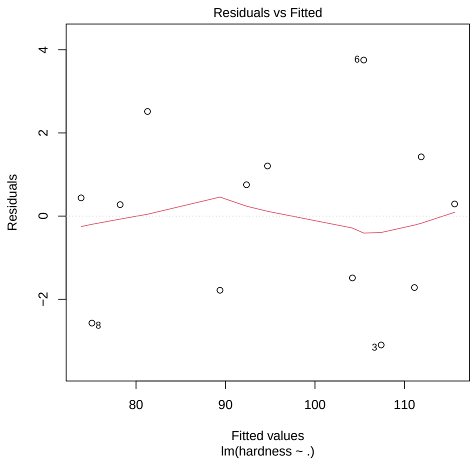
b) Complete the missing values in the table. For "Pr(> |t|)" of `aluminium_ferrite` provide an approximate value. What variables have a statistically significant effect?

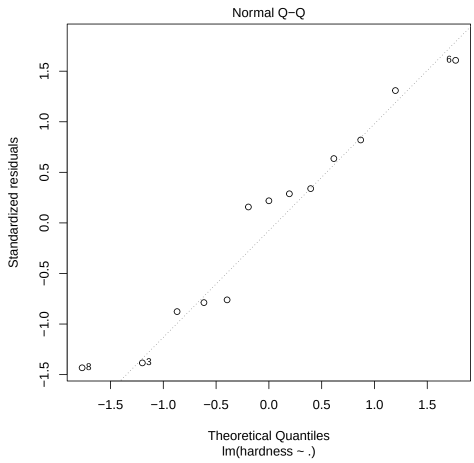
c) Test the statistical hypothesis corresponding to the statement "the covariates do not have an effect on the hardness of cement".

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 Model A. 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 Model A and Model B? Why?

f) The figure below shows two plots regarding the complete model (Model A). Explain what they represent and interpret them.





Exercise 2

Let $(y_1, \dots, y_5)$ and $(y_6, \dots, y_{10})$ 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, $i = 1, \dots, 10$.

b) Write the log-likelihood function of $\boldsymbol{\beta} = (\beta_1, \beta_2)$ 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 $\hat{\boldsymbol{B}}$ of $\boldsymbol{\beta}$, and an approximate distribution of the maximum likelihood estimator $\hat{B}_1$ 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.