Statistical Modelling
Exam 27/06/2024

Exercise 1

Assume that $y_1, \dots, y_{200}$ are realizations of independent Gaussian random variables with variance equal to 1 and mean $\beta_1 + \beta_2 \exp\{z_i\}$ for $i = 1, \dots, 120$, and mean $\beta_1 + \beta_3 \exp\{z_i^2\}$ for $i = 121, \dots, 200$, where the z_i are known constants and $(\beta_1, \beta_2, \beta_3)$ are unknown real parameters.

a) Are the assumptions of a Gaussian linear model satisfied in the above formulation? Justify the answer.

b) State the parameter space and sample space.

c) Consider the model in matrix form: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Explicitly define $\boldsymbol{Y}$, X, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$, and specify their dimensions. Write the distribution of $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$.

d) Obtain the expression of the matrix $X^T X$ and the vector $X^T \boldsymbol{y}$, and explain how these elements can be used to obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

e) Write the distribution of the maximum likelihood estimator $\hat{\boldsymbol{B}}$.

f) Let $\boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ be the vector of the residuals. Indicate which of the following identities hold, and justify your answer:

$$\sum_{i=1}^{200} e_i = 0 \qquad \sum_{i=1}^{200} e_i z_i = 0 \qquad \sum_{i=1}^{200} e_i z_i^2 = 0$$

$$\sum_{i=1}^{200} e_i \exp\{z_i\} = 0 \qquad \sum_{i=1}^{200} e_i \exp\{z_i^2\} = 0 \qquad \sum_{i=1}^{120} e_i \exp\{z_i\} = 0$$

(Hint: pay attention to the indices in each summation.)

Exercise 2

The `chdage` dataset contains measurements on 100 patients for two variables: `age` (in years) and a binary variable (`CHD`) that takes the value 1 if the individual has coronary heart disease and 0 otherwise.

a) To investigate the relationship between age and the probability of coronary heart disease, a researcher fits a logistic regression (Model A), yielding the following output:

	Estimate	Std. Error	z value	Pr(>|z|)
`Intercept`	-5.3095	1.1337	-4.68	0.0000
`age`	0.1109	0.0241	4.610	0.0000

Null deviance:	136.66
Residual deviance:	107.35

a1) Write the corresponding statistical model.

a2) Interpret the coefficient associated with the age variable.

a3) State the null and alternative hypotheses and perform a test to compare the fitted model with a model that includes only the intercept. Specifically, write the test statistic, its distribution, observed value, and the resulting p-value.

b) The researcher then considers whether age might have a quadratic effect and includes the corresponding covariate in the model. The fitted model (Model B) yields the following output:

	Estimate	Std. Error	z value	Pr(>|z|)
`Intercept`	?	4.2901	-0.99	0.3229
`age`	?	0.1947	0.315	0.7527
$\texttt{age}^2$	0.0005	0.0021	?	?

Null deviance:	136.66
Residual deviance:	107.29

b1) Write the corresponding statistical model.

b2) Complete the missing values in the table.

b3) Perform a statistical test to compare Model B and the null model.

b4) Perform a statistical test to determine which of Model A and Model B is preferable.

c) To further investigate the relationship between age and the presence of heart disease, the `age` variable is transformed into a dummy variable. Specifically, the new variable `age<50` takes the value 1 if `age` is less than 50 and 0 otherwise. The output of this model (Model C) is as follows:

	Estimate	Std. Error	z value	Pr(>|z|)
`Intercept`	1.0609	0.3867	2.74	0.0061
`age<50`	-2.0989	0.4788	-4.38	0.0000

Null deviance:	136.66
Residual deviance:	114.61

c1) Write the corresponding statistical model.

c2) Interpret the coefficient of `age<50`.