First name:

Last name:

Student ID number:

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 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? Motivate the answer.

b) State the parameter space and sample space.

c) Express the model in matrix form: $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, explicitly stating how $\underline{Y}$, X, $\underline{\beta}$, and $\underline{\varepsilon}$ are defined and their dimensions. Write the distribution of $\underline{Y}$ and $\underline{\varepsilon}$.

d) Obtain the expression of the matrix $X^T X$ and the vector $X^T \underline{y}$; state how these elements should be used to obtain the maximum likelihood estimate $\underline{\hat{\beta}}$.

e) Write the distribution of the maximum likelihood estimator $\underline{\hat{\beta}}(\underline{Y})$.

f) Let $\underline{e} = \underline{y} - X\underline{\hat{\beta}}$ be the vector of the residuals. State which of the following identities are satisfied and motivate the 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: read the indices in the sum!)

Exercise 2

The data contained in the `chdage` dataset represent the measurements on 100 patients of two variables: the age expressed in years (`AGE`) and a binary variable (`CHD`) which assumes value 1 if the individual has a coronary heart disease and 0 otherwise.

a) To investigate whether there is a relationship between the probability of having a coronary heard disease and the age of the individuals, a researcher fitted a generalized linear model (using the canonical link function) that produced 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 statistical model corresponding to such output (assumptions and model specification).

a2) Write the interpretation of the coefficient associated with the age variable.

a3) Write the system of hypotheses and perform a test to compare the fitted model with a model that includes only the intercept. Comment the result.

b) The researcher then wonders whether the age might have a quadratic effect and adds the corresponding covariate to the model. The fitted model produced 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 statistical model corresponding to such output.

b2) Complete the missing values in the table.

b3) Write the system of hypotheses and perform a test to compare the fitted model with a model that includes only the intercept. Comment the result.

b4) Write the system of hypotheses and perform a test to evaluate which model is preferable between model (a) and (b). Comment the result.

c) To further investigate the relationship between the age and the presence of heart disease, the `age` variable was then transformed into a dummy variable. Specifically, the new variable `age<50` takes value 1 if `age` is smaller than 50 and 0 otherwise. With this new variable, the following output is produced when fitting the model:

	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 statistical model corresponding to such output.

c2) Write the interpretation of the slope coefficient.

		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
χ^2 with 1 df	$\chi^2_{1,p}$	2.7055	3.8415	5.0239	6.6349	7.8794	9.1406	10.8276
χ^2 with 2 df	$\chi^2_{2,p}$	4.6052	5.9915	7.3778	9.2103	10.5966	11.9829	13.8155
χ^2 with 3 df	$\chi^2_{3,p}$	6.2514	7.8147	9.3484	11.3449	12.8382	14.3203	16.2662
χ^2 with 4 df	$\chi^2_{4,p}$	7.7794	9.4877	11.1433	13.2767	14.8603	16.4239	18.4668
χ^2 with 5 df	$\chi^2_{5,p}$	9.2364	11.0705	12.8325	15.0863	16.7496	18.3856	20.5150

Table 1: Some quantiles of Gaussian, and χ^2 distribution: $p = \mathbb{P}(X \leq q_p)$. Columns correspond to probabilities p. Rows correspond to different distributions, in particular, for the χ^2, each row corresponds to different degrees of freedom (df).