Exercises: Generalized Linear Models (part 2)

Exercise 1: Exam 25/01/2024

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

c) Write the approximate distribution of the maximum likelihood estimator $\hat{\boldsymbol{B}}$ of $\boldsymbol{\beta} = (\beta_1, \beta_2)$, and the 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.

Exercise 2

The *Pima* dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements. In particular, the $n = 724$ patients in this dataset are females at least 21 years old of Pima heritage.

The datasets has one response variable (`diabetes`: 1 = positive; 0 = negative), and it is known that, of these women, 249 have diabetes, while 475 do not.
Moreover, we have the following medical predictor variables:

- `pregnant`: Presence of present/past pregnancies: 0 = no pregnancies; 1 = at least one pregnancy.
- `glucose` : Plasma glucose concentration, numeric.
- `pressure`: Diastolic blood pressure (mm Hg), numeric.
- `BMI` : Body mass index, numeric.
- `age` : Age (years), numeric.

Fitting a logistic regression on R returns the following output ("model A"):

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-8.9267	0.8537	-10.46	0.0000
pregnant1	0.2465	0.2931	0.84	0.4004
glucose	0.0349	0.0035	9.92	0.0000
pressure	-0.0078	0.0084	-0.93	0.3515
BMI	0.0941	0.0154	6.09	0.0000
age	0.0328	0.0086	3.81	0.0001

Null deviance: 931.94 on 723 degrees of freedom
Residual deviance: 694.45 on 718 degrees of freedom

Answer the following:

a) Write the corresponding model.

b) Provide the interpretation of the `age` and `pregnant` coefficients.

c) Is it reasonable to remove the `pregnant` variable from the regression? Why?

A new model ("model B") is then fitted removing the `pregnant` and `pressure` variables. This model returns the following output:

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-9.0085	0.7261	-12.41	0.0000
glucose	0.0346	0.0035	9.90	0.0000
BMI	0.0884	0.0147	6.01	0.0000
age	0.0317	0.0079	3.99	0.0001

Null deviance: 931.94 on 723 degrees of freedom
Residual deviance: 696.15 on 720 degrees of freedom

d) Perform a test to compare model A and model B using a 5% significance level. Which one do you prefer?

e) Define the null model. Obtain the estimate of the regression coefficients in this model.

	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
$\chi^2_{1;p}$	2.7055	3.8415	5.0239	6.6349	7.8794	9.1406	10.8276
$\chi^2_{2;p}$	4.6052	5.9915	7.3778	9.2103	10.5966	11.9829	13.8155
$\chi^2_{3;p}$	6.2514	7.8147	9.3484	11.3449	12.8382	14.3203	16.2662
$\chi^2_{4;p}$	7.7794	9.4877	11.1433	13.2767	14.8603	16.4239	18.4668
$\chi^2_{5;p}$	9.2364	11.0705	12.8325	15.0863	16.7496	18.3856	20.515

Table 2: Some quantiles of the χ^2 distribution: $p = \mathbb{P}(X \leq \chi^2_{df;p})$ with $X \sim \chi^2_{df}$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom df.