Exercises: Generalized Linear Models

Exercise 1: Exam 22/02/2024

The `titanic` dataset is a collection of data about 714 passengers, and the goal is to predict whether each passenger survived (`Survival`: 1 if the passenger survived, 0 if they did not) based on some personal characteristics. In particular, here we consider the ticket class (`Class`: 1 = first, 0 = second or third), the gender (`Gender`: man = 1, woman = 0), and the age (`Age`, in years). Fitting a logistic regression model in R produces the following summary:

	Estimate	Std. Error	z value	Pr(>|z|)
`(Intercept)`	1.5003	0.2462	6.09	0.0000
`Class`	2.0103	0.2479	8.11	0.0000
`Gender`	-2.5473	0.2017	-12.62	0.0000
`Age`	-0.0299	0.0074	??	??
Null deviance:	964.52			
Residual deviance:	675.14			

a) Write the corresponding model formulation.

b) Write the likelihood and log-likelihood function.

c) Perform the statistical test to complete the missing values in the table. What variables are statistically significant?

d) Estimate the odds of survival for a 30-year-old woman traveling in first class (denote this individual as "A"). How would the odds change for an otherwise identical woman aged 31? (denote this individual as "B")

e) Interpret the coefficient associated with the `Class` variable. In particular, discuss its effect on the odds and on the survival probability.

f) Define the "null deviance" and "residual deviance" in the output.

g) Perform a test about the overall significance.

Exercise 2 - Exam practice

Consider a sample of $n = 30$ independent observations and fit the logistic regression model $Y_i \sim \text{Bernoulli}(\pi_i)$ where $\text{logit}(\pi_i) = \beta_1 + \beta_2 x_i$, with x_i a dummy variable taking value 1 for the first 10 observations and 0 otherwise. Fitting this model returns the following output

	Estimate	Std. Error	z value	Pr(>|z|)
`(Intercept)`	1.7346	0.6262	2.77	0.0056
`x`	-2.5819	0.9319	-2.77	0.0056

Null deviance: 38.191 on 29 degrees of freedom
Residual deviance: 29.126 on 28 degrees of freedom

Answer the following:

a) Write the likelihood, log-likelihood and score functions for (β_1, β_2).

b) Compute the estimated probabilities $\hat{\pi}$ for $x = 0$ and $x = 1$. Obtain the odds for $x = 0$ and $x = 1$, and interpret the result. Compute an estimate of the odds ratio.

c) Test the hypothesis $H_0 : \beta_2 = -1$ vs $H_1 : \beta_2 \neq -1$.

Exercise 3: Exam 03/09/2024

Consider an experiment designed to study the resistance of a machine component to tensile stress. The dataset records the number of breaks observed in 54 replications of the experiment, considering two types of material (A and B) and three levels of tension (L = low, M = medium, H = high).
To investigate this relationship, we fit a Poisson regression model. The output of the model is the following:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.6920	0.0454	81.30	0.0000
material B	-0.2060	0.0516	-3.99	0.0001
tension M	-0.3213	0.0603	-5.33	0.0000
tension H	-0.5185	0.0640	-8.11	0.0000

Null deviance: 297.37 on 53 degrees of freedom
Residual deviance: 210.39 on 50 degrees of freedom

a) Write the model formulation.

b) Derive and explain the interpretation of the coefficient associated with the variable "material B".

c) Through an adequate statistical test, evaluate whether the tension covariate can be removed, knowing that the residual deviance of the corresponding model is equal to 281.33.

d) A second model ("model B") assumes that both the type of material and the level of tension do not have an impact on the number of breaks. Specify the model and perform a test to compare the model fitted in point (a) with model B.

Exercise 4: Exam 25/01/2024

Let $(y_1, \ldots, y_5)$ and $(y_6, \ldots, y_{10})$ be two independent samples. The random variables in the first sample follow a Poisson distribution with mean $\exp\{\beta_1\}$, while those in the second sample follow a Poisson distribution with mean $\exp\{\beta_1 + \beta_2\}$.

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.

Exercise 5

The *Pima* dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict whether a patient has diabetes based on a set of diagnostic measurements. In particular, the $n = 724$ patients in this dataset are females at least 21 years old of Pima heritage. Among the 724 patients, 249 have diabetes, while 475 do not.

The dataset contains one response variable (`diabetes`: 1 = positive; 0 = negative). Moreover, we have the following medical predictor variables:

- `pregnant`: Whether the patient has ever been pregnant: 0 = no; 1 = yes.
- `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 model formulation.

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.