Exercises: Generalized Linear Model

Valentina Zangirolami - valentina.zangirolami@unimib.it

December 21, 2023

(Referring to the theoretical parts: 20, 21, 22, 23, 24, 25, 26)

1 Exercise 1

To meet competition or cope with economic slowdowns, corporations sometimes undertake a “reduction in force” (RIF), in which substantial numbers of employees are terminated. Federal and various state laws require that employees be treated equally regardless of their age. In particular, employees over the age of 40 years are in a “protected” class, and many allegations of discrimination focus on comparing employees over 40 with their younger coworkers. Here are the data for a recent RIF:

Terminated	**Over40: No**	**Over40: Yes**
Yes	17	71
No	564	835

Exercise 1.1
(a) Choose an appropriate response variable and then a regression model. Justify your answer.
(b) Further, find the estimated regression model.

Exercise 1.2

Software gives the estimated slope $\hat{\beta}_2 = 1.0371$ and its standard error $SE(\hat{\beta}_2) = 0.2755$. Transform the results to the odds scale. Summarize the results (using confidence intervals) and write a short conclusion.

Exercise 1.3

If additional explanatory variables were available, for example, a performance evaluation, how would you use this information to study the RIF?

2 Exercise 2

The acquisition literature suggests that takeovers occur either due to conflicts between managers and shareholders or to create a new entity that exceeds the sum of its previously separate components. Other research has offered managerial hubris as a third option, but it has not been studied empirically. Recently, some researchers revisited acquisitions over a 10-year period in the Australian financial system. A measure of CEO overconfidence was based on the CEO's level of media exposure, and a measure of dominance was based on the CEO's remuneration relative to the firm's total assets. They then used logistic regression to see whether CEO overconfidence and dominance were positively related to the probability of at least one acquisition in a year. To help isolate the effects of CEO hubris, the model included explanatory variables of firm characteristics and other potentially important factors in the decision to acquire. The following table summarizes the results for the two key explanatory variables:

Covariates	**Estimates**	**SE**
Overconfidence	0.0878	0.0402
Dominance	1.5067	0.0057

Exercise 2.1
Write the estimated regression model and interpret the coefficients estimates.

Exercise 2.2

Perform the significance tests and determine whether the variables are significant at the 0.05 level.

Exercise 2.3

Estimate the odds ratio for each variable and construct a 95% confidence interval.

3 Exercise 3

Let consider a dataset on the number of research articles published by 915 graduate students in biochemistry PhD programs. The variables for this dataframe are

- **art**: count of articles produced during last 3 years of PhD
- **fem**: factor indicating gender of student, with levels Men and Women
- **mar**: factor indicating marital status of student, with levels Single and Married
- **kid5**: number of children aged 5 or younger
- **phd**: prestige of PhD department
- **ment**: count of articles produced by PhD mentor during last 3 years

Exercise 3.1
Choose an appropriate response variable and then a regression model. Justify your answer.

Exercise 3.2

Complete the following table.

Coefficients	**Estimates**	**SE**	**Z-obs**	**P-value**
β_1	0.304617	0.102981	2.958	0.0031
β_2	-0.224594	0.054613		
β_3	0.155243			0.0114
β_4		0.040127	-4.607	
β_5		0.026397		0.6271
β_6	0.025543	0.002006	12.733	$<$ 2e-16

Then, interpret the coefficient β_6.

Exercise 3.3

Knowing that the null deviance is equal to 1817.4 while the residual deviance is equal to 1634.4. Perform a statistical test about the overall significance. Specify the hypothesis, the test statistic and the p-value.

Exercise 3.4

Knowing the value of the following quantity

$$\sum_{i=1}^{n} y_i \log(\hat{\mu}_i) - \hat{\mu}_i = -642.0261$$

Find the log-likelihood of the saturated model. Further, interpret the results in terms of goodness of fit.

4 Exercise 4

A researcher is interested in how variables, such as **GRE** (Graduate Record Exam scores), **GPA** (grade point average) and prestige of the undergraduate institution (**Rank**), effect admission into graduate school. The response variable (**Admit**), admit/don't admit, is a binary variable. **Rank** takes on the values 1 through 4. The total number of observations is 400.

Exercise 4.1

Write the equation of the regression model using probit. Which other kind of model can we use?

Exercise 4.2

Knowing the estimates of regression coefficients using a probit model, such that

Coefficients	**Estimates**
β_1	-2.38684
β_2	0.00138
β_3	0.47773
β_4	-0.41540
β_5	-0.81214
β_6	-0.93590

Interpret them.

Exercise 4.3

Now, let consider another link function (should have already been specified in ex. 4.1). Write the regression model.

Exercise 4.4

We are interested in performing a test for comparing models. Let consider the previous model as full model, and the restricted model as a model which just includes **Rank**. Knowing the value of the observed test statistic, $w^{obs} = 16.449$, specify the hypothesis, the theoretical test statistic and p-value. Discuss about the result.

Exercise 4.5

Knowing the deviance of the previous restricted model $D(restricted) = 458.52$, and the deviance of the null model $D(null) = 499.98$, perform a test about overall significance by specifying the hypothesis, the test statistic and the p-value.

Exercise 4.6

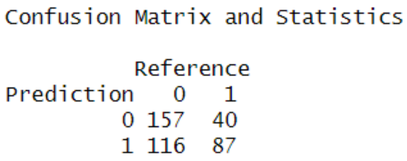
Referring to the full model, we know

- $\hat{\beta}_3 = 0.804038$
- $z^{obs} = 2.423$

Interpret the value of β_3. Perform the test of significance and discuss about the result.

Exercise 4.7

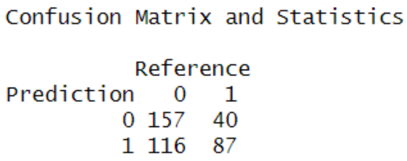
Now, we would like to evaluate our models in terms of mis-classification. Let consider two logistic regression models with probit and logit link functions, involving all covariates. The confusion matrix for each model corresponds to



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 157  40
         1 116  87
```

Figure 1: Probit



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 157  40
         1 116  87
```

Figure 2: Logit

Compute the mis-classification error and interpret the results.

5 Exercise 5

A researcher reports an experiment on the toxicity to the tobacco budworm *Heliothis virescens* of doses of the pyrethroid *trans*-cypermethrin to which the moths were beginning to show resistance. Batches of 20 moths of each sex were exposed for three days to the pyrethroid and the number in each batch that were dead or knocked down was recorded. The results were

	Dose					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Exercise 5.1
Write an appropriate regression model using $\log_2(Dose)$ and *Sex*. Justify your answer.

Exercise 5.2

Considering the following table

Covariates	**Estimates**	**SE**
Intercept	-3.4732	0.4685
Dose	1.0642	0.1311
Sex	1.1007	0.3558

Interpret the estimated regression coefficients and obtain a confidence interval at 95% confidence level.