Exercises 8 solutions

1. (a) Log-likelihood function:

$$\ell = \sum_{i=1}^{N} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{N} n_i \log(1-\pi_i) + \text{ constant}$$

$$= \sum_{i=1}^{N} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{N} n_i \log(1+\mathrm{e}^{\beta_0+\beta_1 x_i}) + \text{ constant}$$

(b) MLE: let $\eta_i = \beta_0 + \beta_1 x_i$, then

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \frac{n_i}{1+\mathrm{e}^{-\eta_i}}, \qquad \frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} \frac{n_i x_i}{1+\mathrm{e}^{-\eta_i}},$$

Putting these derivatives = 0 gives the likelihood equations.

(c) For the elements of information matrix:

$$\mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \beta_0^2}\right) = \sum_{i=1}^{N} n_i \frac{\mathrm{e}^{-\eta_i}}{(1+\mathrm{e}^{-\eta_i})^2}, \qquad \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1}\right) = \sum_{i=1}^{N} n_i x_i \frac{\mathrm{e}^{-\eta_i}}{(1+\mathrm{e}^{-\eta_i})^2},$$

$$\mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \beta_1^2}\right) = \sum_{i=1}^{N} n_i x_i^2 \frac{\mathrm{e}^{-\eta_i}}{(1+\mathrm{e}^{-\eta_i})^2}.$$

(d) For the proposed model, the maximum of ℓ is

$$\hat{\ell}(\text{model}) = \sum_{i=1}^{N} y_i \log \hat{\pi}_i + \sum_{i=1}^{N} (n_i - y_i) \log(1-\hat{\pi}_i) + c$$

where c is the 'constant' term. Thus

$$\hat{\ell}(\text{model}) = \sum_{i=1}^{N} y_i \log(\hat{\mu}_i/n_i) + \sum_{i=1}^{N} (n_i - y_i) \log[1-(\hat{\mu}_i/n_i)] + c.$$

For the saturated model, the fitted values are the observed responses, so

$$\hat{\ell}(\text{sat}) = \sum_{i=1}^{N} y_i \log(y_i/n_i) + \sum_{i=1}^{N} (n_i - y_i) \log[1-(y_i/n_i)] + c.$$

Hence the scaled deviance is

$$D = 2\{\hat{\ell}(\text{sat}) - \hat{\ell}(\text{model})\} = 2\sum_{i=1}^{N} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right) \right\}.$$

and this is also the deviance as the scale parameter is 1.

2. (a) (i) Let $Y_i = 1$ if the ith child took the Leaving Certificate and 0 otherwise, and let $P(Y_i = 1) = \pi_i$. Then the fitted model asserts that

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -7.487 + 0.056 x_{1i} + 0.497 x_{2i} + 0.038 x_{3i}\ ,$$

where x_{1i} and x_{3i} are respectively the verbal reasoning score and father's occupational prestige for the ith child, and x_{2i} is a binary variable taking the value zero for boys and 1 for girls (this follows from the fact that `Gender` was coded as a factor and that by default, R codes factors using a corner-point parameterisation in which the first level — here, `Male` – is taken as the reference level[1]). Alternatively, the model could be written by defining Y_{ij} to be the response for the jth child in the ith gender group ($i = 1$ denoting males, $i = 2$ denoting females); then the model would be written as

$$\log\left(\frac{\pi_{ij}}{1-\pi_{i_{ij}}}\right) = -7.487 + 0.056 x_{1i} + \alpha_i + 0.038 x_{3ij} \ ,$$

in what is hopefully an obvious notation (but now $\alpha_1 = 0$ and $\alpha_2 = 0.497$).

(ii) The `GenderFemale` coefficient represents the average difference, on the log odds scale, between girls and boys — it is $\alpha_2 - \alpha_1$ in the second version of the model definition above. The associated p-value of 0.02 suggests that, even after adjustment for verbal reasoning ability and for father's occupation, boys and girls have different probabilities of taking the Leaving Certificate. The positive coefficient suggests that girls are more likely than boys to take the certificate.

(iii) For a boy (i.e. $x_2 = 0$) with $x_1 = 100$ and $x_3 = 40$, we have

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -7.487 + (0.056 \times 100) + (0.038 \times 40) = -0.367.$$

Thus the estimated probability of taking the Leaving Certificate is $1/(1 + e^{0.367}) = 0.409$.

(b) In the new model, the effect of `Gender` is negligible: the coefficient is close to zero and the p-value is almost 1. The conclusion from this is very different from that suggested in part (a): now it seems that gender is irrelevant as a determinant of whether or not a student takes the Leaving Certificate. Presumably, this is because the primary determinant is the type of school attended: the output from the new model suggests that students at vocational schools are far less likely to take their Leaving Certificate than those at secondary schools (the associated coefficient is negative and highly significant). We might speculate that vocational schools tend to be more male-dominated than secondary schools (obviously, we could check this in a complete analysis of the data): thus the apparent association with gender is in fact due to the type of school attended

(c) The hypothesis being tested in moving from row i to $i + 1$ of the table is that the data were generated from the model summarised in row i. Thus, in moving from row 1 to row 2, the null hypothesis is that the probability of taking the Leaving Certificate depends solely on "academic ability" (as measured by the DVRT). The tiny p-value here indicates overwhelming rejection of this hypothesis. Similarly, in moving from row 2 to row 3, the null hypothesis is that the probability of taking the Leaving Certificate depends solely on "academic

[1] If you didn't know this already, you could figure it out from the fact that the coefficient in the output is labelled `GenderFemale`.

ability", gender and father's occupation (which presumably is a surrogate for social status). Once again, the tiny p-value indicates overwhelming rejection of this hypothesis. Strictly speaking, we can't conclude anything more from this: however, if we think about what is actually happening then it is not unreasonable to conclude that the type of school attended is an important determinant of whether or not a student takes their Leaving Certificate.

(d) The "residuals versus linear predictors" plot shows two well-defined curves. Normally, such clear structure in a residual plot would indicate a problem with the model. However, in this case the observations are either 0 or 1 (this is clear from the 'observations versus linear predictors' plot): thus, if the model predicts a probability μ_i for the ith case in the data set, the residual will be a function of either $-\mu_i$ or $1-\mu_i$. Any plot of residuals against linear predictors (which are a 1-1 function of the fitted values) will therefore show points on one of two curves. The shape of the curves is determined by the linear predictors, not by the responses — thus the apparent structure in this plot tells us essentially nothing about the (lack of) fit of the model.

3. Under the fitted model, $\mathrm{E}\left(T_{(a,b)}\right) = \sum_{a\leq\hat{\pi}_i<b}\hat{\pi}_i$ and $\mathrm{Var}\left(T_{(a,b)}\right) = \sum_{a\leq\hat{\pi}_i<b}\mathrm{Var}\left(Y_i\right) = \sum_{a\leq\hat{\pi}_i<b}\hat{\pi}_i\left(1-\hat{\pi}_i\right)$. If there is a large number of observations with $a\leq\hat{\pi}_i<b$ then, as a sum of a large number of observations, $T_{(a,b)}$ will be approximately normally distributed by the central limit theorem; thus

$$Z_{(a,b)} = \frac{T_{(a,b)} - \sum_{a\leq\hat{\pi}_i<b}\hat{\pi}_i}{\sqrt{\sum_{a\leq\hat{\pi}_i<b}\hat{\pi}_i\left(1-\hat{\pi}_i\right)}}$$

has approximately a standard normal distribution.

Since the observations are assumed to be independent, the statistic $\sum_{j=1}^{g}Z_j^2$ should have approximately a χ_g^2 distribution if the data really were generated from the model.

[This idea forms the basis of the Hosmer-Lemeshow test discussed in §3.3.2 of the lecture notes — although the Hosmer-Lemeshow test statistic is a slightly simplified version of that derived above, and the degrees of freedom of null distribution are adjusted to account for the estimation of the $\{\pi_i\}$].

4. (a) The log-likelihood is just

$$\ell = \sum_{i=1}^{N}\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right] = \sum_{i=1}^{N}\left[w_i\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right]\ .$$

(b) The scaled deviance is defined as $D^* = 2\left[\ell(\mathrm{sat}) - \ell(\mathrm{model})\right]$

$$= \frac{2}{\phi}\sum_{i=1}^{N}w_i\left[\left(y_i\tilde{\theta}_i - b(\tilde{\theta}_i)\right) - \left(y_i\theta_i - b(\theta_i)\right)\right]\ .$$

The unscaled deviance is thus

$$D = \phi D^* = 2\sum_{i=1}^{N}w_i\left[\left(y_i\tilde{\theta}_i - b(\tilde{\theta}_i)\right) - \left(y_i\theta_i - b(\theta_i)\right)\right]\ ,$$

which is independent of ϕ as claimed.