Exercises 8

1. Suppose that $Y_1, \ldots, Y_N$ are N independent responses such that $Y_i \sim \text{Bin}(n_i, \pi_i)$ and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \text{ for } i = 1, \ldots, N.$$

   (a) Obtain the log-likelihood function ℓ as a function of β_0 and β_1 (given the data).

   (b) Obtain the likelihood equations (by differentiating ℓ with respect to β_0 and β_1).

   (c) Obtain the information matrix.

   (d) Obtain algebraically the form of the deviance given in Section 3.3.2.

   Hint: You may find it useful to note that if $\log[\pi/(1-\pi)] = z$ then $\pi = e^z/(1+e^z) = 1/(1+e^{-z})$.

2. In the Irish education system, the Leaving Certificate is the final examination for high school students and is usually taken between the ages of 16 and 19 years (see `http://en.wikipedia.org/wiki/Leaving_Certificate`). In a study to investigate educational opportunities in Ireland[1], data were gathered on 441 school leavers. Variables recorded included: gender (male or female, denoted `Gender` below); the Drumcondra Verbal Reasoning Test score (a measure of verbal reasoning ability, measured on a continuous scale and denoted by `DVRT` below); whether or not the Leaving Certificate was taken (0=no, 1=yes, denoted by `CertTaken` below); a measure of the father's occupational prestige (continuous-valued with higher values indicating more prestigious occupations, denoted by `Prestige` below); and the type of school (secondary or vocational, denoted by `SchoolType` below).

   (a) The following (edited) R output is from an analysis of these data in which logistic regression was used to explore the social and academic factors associated with taking the Leaving Certificate (note that `Gender` was defined to the system as a factor, whereas `DVRT` and `Prestige` are both continuous covariates):

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.486916   0.923494  -8.107 5.18e-16 ***
DVRT          0.055578   0.008306   6.692 2.21e-11 ***
GenderFemale  0.496952   0.217793   2.282   0.0225 *
Prestige      0.037782   0.007593   4.976 6.50e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

      (i) Write down the mathematical representation of the model summarised by the output above, taking care to define any notation that you use. Give the estimated values of any model parameters.

      (ii) How do you interpret the coefficient labelled `GenderFemale`, and the associated p-value?

      (iii) According to the model above, what is the probability of taking a Leaving Certificate for a boy with a DVRT score of 100 and whose father has a prestige score of 40? (these are both close to the average values in the data set).

[1] Reported in Raftery, A.E. and M. Hout (1985). Does Irish education approach the meritocratic ideal? A logistic analysis. *Economic and Social Review* **16**, pp115-140.

(b) The model above was extended by adding the `SchoolType` variable (again, coded as a factor). The result was as follows:

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.571088   1.028429  -4.445 8.80e-06 ***
DVRT                  0.040374   0.009259   4.361 1.30e-05 ***
GenderFemale         -0.004924   0.252142  -0.020  0.98442
Prestige              0.025694   0.008456   3.039  0.00238 **
SchoolTypeVocational -3.149843   0.421097  -7.480 7.43e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

What do you conclude about the effect of `Gender` now? What has changed, and why?

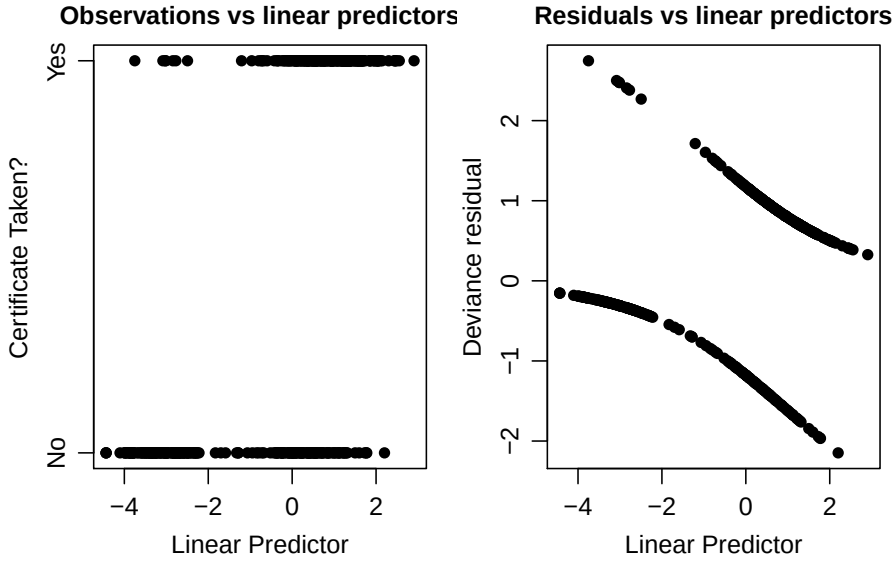
(c) The following analysis of deviance table was used to compare three models in R (you aren't told explicitly what the models are, but the table gives you all the information you need to figure it out):

```
Analysis of Deviance Table

Model 1: CertTaken ~ DVRT
Model 2: CertTaken ~ DVRT + Gender + Prestige
Model 3: CertTaken ~ DVRT + Gender + Prestige + SchoolType
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       439     545.24
2       437     515.01  2   30.227 2.731e-07 ***
3       436     417.82  1   97.187 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

What are the hypotheses being tested in this table, and what do the results tell you?

(d) The following figure shows plots of both the response variables and the deviance residuals against the linear predictors, for the model in part (b) above.



What does the "residuals versus linear predictors" plot tell you about the (lack of) model fit? (**Hint:** think *carefully* about what is being plotted!)

3. An alternative way to check the residual structure in a GLM with binary responses (as in the previous question) is to group the observations based on the values of the linear predictor or, equivalently, of the fitted values $\{\hat{\pi}_i\}$. Specifically, consider a subgroup $\{Y_i : a \leq \hat{\pi}_i < b\}$ for which the fitted values lie in the range $[a, b)$. Let $T_{(a,b)} = \sum_{a \leq \hat{\pi}_i < b} Y_i$ be the total number of 'successes' in this subgroup. Write down expressions for the expected value and variance of $T_{(a,b)}$ under the fitted model. Hence find values $m_{(a,b)}$ and $s_{(a,b)}$ such that, if the fitted model is correct and there is a large number of observations with $a \leq \hat{\pi}_i < b$, the statistic $Z_{(a,b)} = \left[T_{(a,b)} - m_{(a,b)}\right] / s_{(a,b)}$ has approximately a standard normal distribution.

   Suppose now that the interval $[0, 1]$ is partitioned into g subintervals as $0 = a_0 < a_1 < \cdots < a_{g-1} < a_g = 1$, so that the ith subinterval runs from a_{i-1} to a_i. Suppose also that the statistics $Z_{(a_0,a_1)}, Z_{(a_1,a_2)}, \cdots, Z_{(a_{g-1},a_g)}$ have been calculated. Can you suggest a way to combine all of these Zs into a single test statistic that could be used to measure the goodness of fit of the model? Ignoring the effect of estimation error (i.e. that the fitted values are the $\{\hat{\pi}_i\}$ rather than the $\{\pi_i\}$ themselves), what would be the distribution of this test statistic under the null hypothesis that the data were generated from the model?

4. (a) Using the definition of the exponential family at equation (3.3) of the lecture notes, write down an expression for the log-likelihood of a GLM with independent responses $y_1, \ldots, y_N$, canonical parameters $\theta_1, \ldots, \theta_N$ and dispersion parameter ϕ (leave your expression in terms of θs and ϕ).

   (b) Denoting by $\tilde{\theta}_1, \ldots, \tilde{\theta}_N$ the values of the canonical parameter for the saturated model, use the result from part (a) to write down an expression for the scaled deviance, D^*. Deduce that the unscaled deviance, $D = \phi D^*$, does not depend on the dispersion parameter.