APPLIED STATISTICS, MATH10096

Workshop 4

1. Let y_i be independent random variables each with mean μ_i and variance $V(\mu_i)\phi$, where $\mu_i = g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$, g is a smooth monotonic function, $\boldsymbol{\beta}$ a vector of parameters, and $\mathbf{X}$ a design matrix of known entries. Let $z_i = g'(\mu_i)(y_i - \mu_i) + \mathbf{X}_i\boldsymbol{\beta}$. Find the expected value and covariance matrix of the vector $\mathbf{z} = (z_1, \ldots, z_n)^\mathsf{T}$.

2. This question is about getting a feel for the IRLS method used to fit GLMs.

   (a) Find an expression for the value of β minimizing

   $$\sum_{i=1}^{n} w_i(z_i - \beta x_i)^2$$

   in terms of w_i, z_i, and x_i.

   (b) Consider fitting the following GLM

   $$\mathbb{E}(y_i) = \beta x_i, \quad y_i \sim \text{Poi},$$

   to the data:

x	1	2	3	4	5
y	0	1	3	7	4

   Apply two steps of the IRLS scheme, by hand, in order to obtain an MLE of β. You will need to modify the default starting values for the IRLS, by changing the initial value used for $\hat{\mu}_1$ (why?): I suggest starting with $\hat{\mu}_1 = 1$.

   (c) Obtain an estimate of the standard error of $\hat{\beta}$.

3. If you add a completely irrelevant predictor variable, with one associated parameter, to a GLM, how much decrease do you expect in the scaled deviance of the GLM?

4. Consider independent Inverse Gaussian response variables, y_i, for $i = 1, \ldots, n$, modelled using a GLM that predicts a mean $\hat{\mu}_i$ for each y_i. Further note that a saturated model which uses one parameter per observation will produce predicted values $\tilde{\mu}_i = y_i$. Show that the deviance is given by

   $$D = \sum_{i=1}^{n} \left\{ \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2} \right\}.$$

5. This question covers the Belgian AIDS epidemic data discussed in section 10.1 of the lecture notes. The data can be entered into R as follows.

```
y <- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
t <- 1:13
```

   Recall that y refers to the number of new AIDS cases per year and t denotes the year since 1980 (and until 1993).

(a) Plot AIDS cases against year since 1980.

(b) Use the `glm` function to fit the model for the AIDS data given in the notes. You will need to use the `poisson` family and `"log"` link function.

(c) From the scaled deviance, does it appear that the model is adequate?

(d) Examine the residuals for the fitted model. Is the model adequate?

(e) Propose an alternative model for the data and fit it.

(f) Check the alternative model, and see if all its terms are 'significant', using the `summary` command.

(g) Use AIC to compare your revised model and the original model.

(h) Which model do you think best describes the data?

(i) What can you conclude about whether or not the epidemic is continuing to increase exponentially by the end of the data series?

6. This question is about analyzing the O-rings data discussed in Question 5 of workshop sheet 3. The data is available on Learn in the file `orings.rda` and can be read in `R` as follows.

```
load("orings.rda")
```

Recall the context: in January 1986 the space shuttle Challenger exploded shortly after take-off. Subsequent investigation eventually focused on the possibility of o-rings in the fuel tanks having failed as a result of the unusually low launch temperature (31 degrees F), hence allowing a fuel leak. You can read more in Wikipedia:

https://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster

It turned out that data on o-ring failure in previous launches was available for various launch temperatures. In particular, for each temperature considered, we have the number of o-rings, out of 6, that failed/were damaged. As discussed in the solutions of workshop 3, a logistic regression model might be appropriate. Viewing each of the 6 o-rings at each launch temperature as independent binomial trials, with a probability of failure that depends on temperature, gives:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 \texttt{temp}_i, \quad y_i \overset{\text{indep.}}{\sim} \text{binom}(p_i, 6), \tag{1}$$

where p_i denotes the probability of o-ring failure/damage at temperature temp_i and y_i denotes the number of damaged o-rings, out of 6, at temperature temp_i.

(a) Reproduce the plot from workshop 3 sheet.

(b) Fit the model suggested in (1) in `R`. You will need to use the `binomial` family and `"logit"` link function.

(c) The standard residual plots for this model are of little use. Why is that?

(d) To get a feel for the plausibility of the model, overlay a curve showing the expected number of failures as a function of temperature on your original data plot. Plot also approximate 95% confidence limits for the mean number of failures at each temperature. Note that an inverse function of the logit link can be created in `R` as follows:

```
ilogit <- binomial()$linkinv
```

(e) Obtain a point estimate and approximate 95% CI for the expected number of o-ring failures at the Challenger launch temperature of 31 degrees F. What does this suggest? How reliable are the results likely to be?

(f) In some respects the first observation in the `orings` dataset looks out of line with the others, and is an extremely influential point in the fit. Use the `cooks.distance` function applied to your fitted model object to confirm this. What do you think should be done about this?

7. In this question we will be analyzing the data in the 'contingency' table from Question 4 in workshop 3 exercises sheet. Contingency tables occur when you have observations of several factor variables for a number of study subjects, and what is of interest is how many subjects/observations occur at each combination of the factor variables. The questions of interest are usually about whether the factors are linked in some way. Question 4 on workshop sheet 3 concerns a simple example of a contingency table, useful for investigating whether there is an association between gender and belief in life after death. We can address this question by using analysis of deviance to compare the fit of two competing models of these data: one in which belief is modelled as independent of gender, and a second in which there is an interaction between belief and gender. The data are as follows:

	Believer	Non-Believer
Female	435	147
Male	375	134

The data can be entered in `R` as follows (below 1 stands for believer and 0 for non-believer).

```
al <- data.frame(y = c(435, 147, 375, 134),
                 gender = as.factor(c("F","F","M","M")),
                 belief = as.factor(c(1, 0, 1, 0)))
```

As we have seen, under the null model of independence between belief and gender and if y_i is an observation of the counts in one of the cells of the table, then we could model the expected number of counts as

$$\mu_i \equiv \mathbb{E}(Y_i) = n\gamma_k\alpha_j \text{ if } y_i \text{ is data for gender } k, \text{ and faith } j.$$

where n is the total number of people sampled, γ_1 and γ_2 are the proportion female and male in the population sampled, and α_1 and α_2 are the proportion of believers and non-believers in the population sampled. As seen in the previous workshop, we can yield a linear predictor by taking logs

$$\log(\mu_i) = \log(n) + \log(\gamma_k) + \log(\alpha_j) \text{ if } y_i \text{ is gender } k \text{ and faith } j,$$

and defining $\tau = \log(n)$, $\beta_k = \log(\gamma_k)$ and $\delta_j = \log(\alpha_j)$, we get

$$\log(\mu_i) = \tau + \beta_k + \delta_j \text{ if } y_i \text{ is gender } k \text{ and faith } j.$$

Note that the appropriate distribution for the data is really multinomial, but the resulting likelihood would look just like the likelihood that we would get by assuming that the counts in each cell of the table are observations of Poisson random variables, so the Poisson distribution can be used in practice.

(a) Fit the model assuming that belief and gender are independent. Since there are only 4 residuals with 1 degree of freedom between them, there is no point looking at residual plots for this one. From the scaled deviance, does it appear that the model is adequate?

(b) Do these data provide any evidence for an association between gender and belief in an afterlife.