Assignment Generalized Linear Models

A a single pdf-file containing solutions to the problems below is to be uploaded in Blackboard, due no later than the deadline for this assignment. Submitted report should be well-organized and include step-by-step explanations, relevant plots, tables, R code and conclusions.

I. Computational problems

1. Load the dataset `chemo` into `R`. The first four columns contain the number of epileptic seizures experienced during each of four consecutive two-week periods for $n = 58$ patients. The fifth column is an indicator of whether an active treatment was given to the patient (0=placebo, 1=active). Column six contains a baseline rate for each patient, established before the trial started. The last column contains the age of the patient in years. Add the first four columns together, and work with this sum as the response variable `seizures`.

   (i) Investigate the association between `seizures` (the total number of seizures during the eight-week period) and the predictors `treatment`, `baseline` and `age`, using a Poisson regression model with the canonical link function. For each patient report the estimate for the expected number of epileptic seizures. Provide a 95% confidence interval for each regression coefficient. Plot the observed values of the response against the fitted values, and comment on the quality of the fit.

   (ii) Comment on the significance of each predictor based on the P-values in the model summary and in the analysis of deviance table. For the analysis of deviance table, conduct a chi-square test for the significance of each predictor.

   (iii) Plot the deviance residuals against the fitted values, and plot the Pearson residuals against the fitted values and compare the two plots. Obtain the Pearson chi-squared statistic P and the deviance D. Comment on your findings.

   (iv) Suppose we use the link function $g(u) = \sqrt{u}$. Does this improve the fit?

2. The following data (taken from McCullagh and Nelder, *Generalized Linear Models*, 2nd ed., 1989, pp. 300-302, Taylor and Francis) provide clotting times (in seconds) for normal blood diluted to 9 different concentrations with prothrombin-free plasma. Clotting was induced by two different lots of thromboplastin.

```
clotting<-data.frame(conc=c(5,10,15,20,30,40,60,80,100,5,10,15,20,30,40,
60,80,100),lot=factor(c(rep(1,9),rep(2,9))),time=c(118,58,42,35,27,25,21,
19,18,69,35,26,21,18,16,13,12,12))
```

   (i) Use a Gamma model with the multiplicative inverse link (this is the default link function for the Gamma family in `R`) to investigate the association between the clotting time and the predictors `log(conc)` and `lot`. Plot clotting time against `log(conc)`, using a different plotting symbol for each `lot`. For each individual report the estimate for the expected clotting time. Provide a 90% confidence interval for each regression coefficient. Plot the observed values of the response against the fitted values, and comment on the quality of the fit.

(ii) Comment of the significance of each predictor based on the P-values in the model summary and in the analysis of deviance table. For the analysis of deviance table, conduct an F test for the significance of each predictor.

(iii) Plot the deviance residuals against the Pearson residuals. Obtain the Pearson chi-squared statistic P and the deviance D. Comment on your findings.

3. The file `birthweight.txt` (available on Blackboard) contains the following measurements for each of 189 infants: birth weight (grams), mother's age (years), mother's weight (pounds), mother's race (1=white, 2=black, 3=other), smoking status during pregnancy (0=No, 1=Yes), history of premature labor (number of occurrences), history of hypertension (0=No, 1=Yes), presence of uterine irritability (0=No, 1=Yes), and number of visits to a physician during the first trimester of the pregnancy. Convert the columns `Race`, `Smoker`, `Hypertension` and `UterineIrrit` into the factor format. Define `Low Birth Weight` as a birth weight below 2500 grams. Based on this criterion, create a factor `Low` which indicates whether or not each subject had a low birth weight. Add this as a new column to the data frame.

(i) Fit a full binomial model using logistic regression of infant low birth weight status against mother's age, weight, race, smoking status, premature labor history, hypertension history, uterine irritability status and number of first-trimester physician visits. Based on the analysis of deviance table for the full model with a chi-squared test, select only those predictors whose P-values are below 0.1, and fit a reduced model involving only these predictors. How does the AIC compare between the full and reduced models?

(ii) Consider the reduced model obtained in (i). For each infant report the estimate for the probability of low birth weight. Discuss the quality of the fit based on the deviance and Pearson chi-squared statistics.

II. Theoretical problems

1. Prove formulas (3.10) on page 46 from the lecture notes.

2. Show that the Poisson probability mass function $f(y;\lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^y}{y!}$, with $\lambda > 0$ and $y = 0, 1, \ldots$, can be put in form (3.9) (see page 46 in the lecture notes) with canonical parameter $\theta = \log\lambda$. Identify the functions $b(\theta)$, $c(y,\phi)$ and the parameters ϕ and A. Give the canonical link function and use (3.10) to derive that $\mathbb{E}Y = \mathrm{Var}(Y) = \lambda$, where $Y \sim \mathrm{Poisson}(\lambda)$. Assuming that one uses Poisson regression with the canonical link, what is the form of the ith diagonal element of the weight matrix $\mathbb{W}$, in terms of λ_i?

3. Data $Y_1, \ldots, Y_n$ are assumed to follow a binary logistic model in which Y_j takes value 1 with probability $\pi_j = \exp(x_j^T\beta)/((1+\exp(x_j^T\beta))$ and value 0 otherwise, for $j = 1, \ldots, n$. Show that the deviance for a model with fitted probabilities $\hat{\pi}_j = \pi_j(\hat{\beta})$ can be written as $D = -2\big[Y^T X\hat{\beta} + \sum_{j=1}^n \log(1-\hat{\pi}_j)\big]$ and the likelihood equation is $X^T Y = X^T\hat{\pi}$.