

UNIVERSITY OF WOLLONGONG
School of Mathematics and Applied Statistics
STAT902. Advanced Data Analysis

ASSIGNMENT 5

Due: 5:00pm Monday 19th April, 2010.

Local students: please put under lecturer's door.

Remote students: please e-mail or post by this date.

1. Consider the Poisson GLM:

$$y_i \stackrel{\text{ind.}}{\sim} \text{Poisson}(\exp((\mathbf{X}\boldsymbol{\beta})_i)), \quad 1 \leq i \leq n$$

(also known as Poisson regression) with the canonical link $\eta = \ln(\mu)$. Derive the expression for the deviance:

$$D = 2\{\mathbf{y}^T \ln(\mathbf{y}/\hat{\boldsymbol{\mu}}) - \mathbf{1}^T(\mathbf{y} - \hat{\boldsymbol{\mu}})\} = 2 \sum_{i=1}^n \{y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}.$$

2. This question involves doing some diagnostic checks and model adjustments for the ragweed pollen analysis corresponding to Question 3 of Assignment 4.
- (a) Make sure that the data set file `ragweed1994.dat` from Assignment 4 is available.
 - (b) Save the files `galapagos.txt` and `galapagos.Rs` from the Computer Code and Data page on the course web-site.
 - (c) Start an R session and change directories to the one containing `galapagos.txt` and `galapagos.Rs`. Type `source("galapagos.Rs")` to run the script. This should produce a set of diagnostic plots for both the ordinary Poisson regression model and the improved quasi-likelihood Poisson model – corresponding to the slides titled *Additional Aspects of Generalised Linear Models* presented in class.
 - (d) Copy the file `galapagos.Rs` to a new file `ragweed1994.Rs`.
 - (e) Modify `ragweed1994.Rs` so that it does similar diagnoses and model improvements for the ragweed pollen count data analysis. As it turns out, the quasi-likelihood extension of the Poisson model still does not fit the data very well and advanced variance modelling is required. Do not concern yourself with these issues in this assignment question. The aim is to learn some of the mechanics of GLM diagnostics. You should include the diagnostic plots in your assignment submission. (Since rain is binary there is no need to have a partial residual diagnostic for this particular predictor.)

3. Random variables X and Y have joint density function

$$[x, y] = \frac{2}{3}(x + 2y), \quad 0 < x < 1, 0 < y < 1.$$

What is the best prediction of X if Y is observed to be $1/2$?

4. Random variables X, Y and Z are related and distributed as follows:

$$Y = X + Z \quad \text{where} \quad \begin{bmatrix} X \\ Z \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 7 \end{bmatrix} \right)$$

Only the random variable Y is observed.

- (a) What is the *best prediction* of X given $Y = y$? Your answer should derive the required result from first principles, not just using results on partitioned normal random vectors.
- (b) If Y is observed to be 2.34 then what is the predicted value of X ?

Hint: The joint distribution of X and Y is also bivariate normal.