STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 9

1. The data file `Leuk.txt` contains the survival time in weeks and the initial white blood cell count for leukemia patients. It also contains an indicator of whether the patient was positive (AG = 1) or negative (AG = 2) for a particular blood factor called AG.

   (a) To model survival times in weeks, fit a gamma GLM with a logarithmic link and the logarithm of the white blood cell count as the predictor. Plot the deviance residuals versus the linear predictor values. Do you see any pattern to be concerned about? Replot the deviance residuals using different symbols for the residuals from the positive and the negative patients. Do you see any pattern to be concerned about now?

   (b) Include the blood factor as a categorical predictor. Is it significant? Again, plot the deviance residuals using separate symbols for residuals from the positive and negative patients. Do you see any patterns of concern?

   (c) Include an interaction term between the blood factor and the logarithm of the white blood cell count in the model. What does the inclusion of this interaction term allow in the model structure? Test whether the interaction is significant.

2. The data file `Transaction.txt` contains data regarding the audit of a large Australian corporation. The audit process incurred different transaction costs for examining each of two different transaction types. The available data consist of the total time (in minutes) spent on examining both types of transactions in 261 different branch offices of the enterprise as well as the total number of transactions of the two different types occuring at each branch during the 1985-86 financial year. The aim is to develop a model for the cost associated with dealing with each transaction type. Since cost is related to time, we can treat time as the response and try to model its relationship to the number of different transactions. Moreover, since all the time is spent on one or the other of the two transaction types, the response should be additively related to the explanatory variables and ought to go through the origin.

   (a) Fit an ordinary least-squares regression to this data. Do you think that an intercept of zero is reasonable based on this analysis? Plot the residuals versus fitted values, as well as a normal Q-Q plot of the residuals. Do you think that a normal linear model is reasonable for these data based on these plots?

   (b) Generally speaking random times are asymmetrically distributed, and indeed often exponentially distributed. Thus, the distribution of the total of a number of random

times often falls into the gamma distribution family. Fit a gamma generalised linear model to the transaction data, maintaining the additive structure of the relationship between the mean response and the predictors (i.e., retain the identity link function). Plot the deviance residuals versus the linear predictor values. Comment on the plot. Do you think that a model without intercept is justifiable for this model?

(c) Suppose that we believe the model without intercept is appropriate and that a new branch is surveyed. If it is found that the 100,000 transactions of each type were recorded for this branch during the 1985-86 financial year, estimate the amount of time that will be necessary to complete the audit process for this branch. In addition, estimate and find a 95% confidence interval for the proportion of this total audit time spent to examine transactions of the first type.

Note: the proportion of time spent on the first transaction will be estimated by a non-linear function of the parameters $h(\beta)$. To find the confidence interval, use the following approximation for the variance:

$$Var(h(\hat{\beta})) \approx \frac{\partial h(\beta)}{\partial \beta}^T Var(\hat{\beta}) \frac{\partial h(\beta)}{\partial \beta} = \phi \frac{\partial h(\beta)}{\partial \beta}^T (X^T W X)^{-1} \frac{\partial h(\beta)}{\partial \beta}$$

This is the delta-approximation to the variance of a function of a random quantity. Note that $Var(\hat{\beta}) = \phi(X^T W X)^{-1}$. From R, we obtain the estimate $\hat{\phi}$ with the command `m1$dispersion` , and we obtain an estimate of the matrix $(X^T W X)^{-1}$ with the command `m1$cov.unscaled` (for the `glm` object named `m1`).