STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 7

1. In the following situations, identify whether the described random variable plausibly has a Bernoulli, Binomial, Poisson, Normal or some other type of probability distribution. Briefly justify your answer.

   (a) Let Y be the number of iPhone owners out of a random sample of 100 that download an app today.

   (b) Let Y be the number of babies born on a single day in Canberra.

   (c) Let Y be the initial weight (in kilograms) of a randomly selected male enrolling in the Biggest Loser diet.

   (d) Let Y be the initial weight (in kilograms) of a randomly selected person on the Biggest Loser diet.

   (e) Let Y be 1 if a randomly selected person who saw the movie The Wolverine liked it, and 0 if not.

   (f) Let Y be 1 if you flip a coin and it lands tails, and 2 if it lands heads.

2. This question concerns the choice of link functions.

   (a) Suppose a probability distribution has an unknown mean μ that is restricted to be greater than 1. Why is $g(\mu) = \log(\mu)$ NOT a sensible link function? What would be a more reasonable link function?

   (b) Suppose that a probability distribution has an unknown mean μ that is restricted to be between -1 and 1. What would be a reasonable link function for μ?

3. The data file `Heart.txt` is located on Wattle and contains data for 99 individuals ordered in increasing age groups. For each age group (`age`), the total number of individuals at that age (`ssize`) and the number of subjects at that age who have symptoms of heart disease (`disease`) are given.

   weighted least square

   (a) Fit an unweighted least-squares empirical logit model to this data. In other words, fit the model

$$\log\left(\frac{\texttt{disease/ssize}}{1-\texttt{disease/ssize}}\right) = \beta_0 + \beta_1\texttt{age} + \epsilon,$$

   where ϵ is assumed to be normal with mean zero and constant variance. Note that, since many of the observed proportions are either zero or one, you will need to modify them to, say, 0.05 or 0.95, respectively, before taking the logit transformation. Also, investigate how important the choices of these arbitrary values are by re-fitting the regression after having set the zero and one values to 0.005 and 0.995.

(b) The Delta method estimate of variance for $g(Y)$ is given by:

$$\mathbb{V}g(Y) \doteq \left[g'(\mathbb{E}Y)\right]^2 \mathbb{V}Y.$$

Use this fact, and the knowledge that Y is binomially distributed, to find the variance of the empirical logit transformed proportions. In addition, use these weights to fit a weighted least-squares empirical logit to the data. In other words, fit the models in (a), but use weighted least-squares with the appropriate weights. [NOTE: You will have to use the iterative scheme outlined in class.]

(c) Fit a logistic regression to the data. Compare your parameter estimates to those from (a) and (b).

(d) Using the results of the model in (c), by what factor has your odds of having symptoms of heart disease increased once you are 10 years older? Does your answer to this question depend on what age you currently are? [HINT: Recall that your odds of having symptoms at a particular age is just $\pi(\texttt{age})/[1-\pi(\texttt{age})]$.]

4. The data for the anaesthetic depth example shown in lectures is stored in the file `AnsthcSum.txt` on Wattle. In Tutorial 5, we saw that the estimate for the 50%-response concentration, $x_{0.5}$, was given by

$$x_{0.5} = \frac{g(0.5) - \widehat{\beta}_0}{\widehat{\beta}_1},$$

where g was either the logistic, probit or complementary log-log link function.

(a) Calculate a 95% confidence interval for this quantity for each of these three link functions. [NOTE: Recall that for binomial data we assume that $\phi = 1$, since we have included the factors $1/n_i$ into the weights.]

(b) Now, calculate 95% confidence intervals for the parameters β_0 and β_1.

(c) Using the confidence intervals from (b), we might suggest an alternative confidence interval for the 50%-response concentration be constructed from the largest and smallest values that the 50%-response concentration can take for parameter values within the intervals of (b). Use the logistic link model to investigate this alternative method. Do you think that this alternative method will generally provide reasonable intervals?