MAST90104: A First Course in Statistical Learning

Week 10 Lab and Workshop

1 Practical questions

1. The `cornnit` dataset in the `faraway` package contains data on the effect of nitrogen on the yield of corn. Fit a gamma regression to this data, using the `glm` command. You will need to pay attention to the choice of link function (inverse, identity or log), and consider transforming the predictor variable (your first step should be to plot the data).

   (a) Extract the Pearson residuals from the fitted model using the `residuals` function, then use them to estimate the dispersion parameter. Check that your answer agrees with the summary output from your model.

   (b) Suppose your fitted model is `gmod`, then the command `anova(gmod, test="F")` will compare your model against the null model, using an F test. Using the deviances and dispersion estimates reported by `summary(gmod)`, check that the F statistic reported by the `anova` function is correct.

   (c) Now do some diagnostic plots. Can you identify a potential outlier?

   (d) Fit a linear model to the `cornnit` data.
   Which do you prefer, the linear model or the gamma model, and why?

   (e) Re-do the gamma and linear model diagnostic plots with the standard R diagnostic plots and comment.

2. In the `multinom` function from the `nnet` package, the response should be a factor with K levels or a matrix with K columns, which will be interpreted as counts for each of K classes. The first case is a short hand for responses of the form multinomial(1, $\mathbf{p}$). The `hsb` data from the `faraway` package was collected as a subset of the "High School and Beyond" study, conducted by the National Education Longitudinal Studies program of the U.K. National Center for Education Statistics. The variables are gender; race; socioeconomic status; school type; chosen high school program type; scores on reading, writing, math, science, and social studies. We want to determine which factors are related to the choice of the type of program—academic, vocational, or general—that the students pursue in high school. The response is multinomial with three levels.

   (a) Fit a trinomial response model with the other relevant variables as predictors (untransformed).

   (b) Use either backward elimination with χ^2 tests (using the `anova` command), or the AIC (using `step`), to produce a parsimonious model. Give an interpretation of the resulting model.

   (c) For the student with id 99, compute the predicted probabilities of the three possible choices.

3. The `pneumo` data from the `faraway` package gives the number of coal miners classified by radiological examination into one of three categories of pneumonoconiosis and by the number of years spent working at the coal face divided into eight categories.

   (a) Treating the pneumonoconiosis status as response variable as nominal, build a model for predicting the frequency of the three outcomes in terms of length of service and use it to predict the outcome for a miner with 25 years of service.

   (b) Repeat the analysis with the pneumonoconiosis status being treated as ordinal.

2 Workshop questions

1. Suppose $Y_i, i = 1, \cdots, n$ are from a generalised linear model so they are independent from an exponential family:

$$f(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

with the parameter ϕ constant and supposed known but θ_i varies. Recall that

$$\begin{aligned} \mu &= \mathbb{E}Y = b'(\theta) \\ V(\mu) &= \operatorname{Var} Y = b''(\theta)a(\phi) \\ v &= b'' \circ (b')^{-1} \end{aligned}$$

and that there is a link function, g, so that $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ are the parameters of interest, $\mu_i = \mathbb{E}Y_i$ and $\mathbf{x}_i$ is a vector of explanatory variables (this is the ith row of the predictor matrix X). In answering the questions below, you will establish that the Newton-Raphson method with Fisher scoring is the same as the iteratively weighted least squares algorithm introduced in lectures.

(a) Write down the log likelihood as a function of $\boldsymbol{\beta}$ and show that its derivative, $U(\beta_j)$, with respect to β_j may be written as:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}.$$

(b) Hence show that

$$Cov(U(\beta_j)U(\beta_k)) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{V(\mu_i)(g'(\mu_i))^2}.$$

(c) Find the Fisher information and show that it is $X^TW(\boldsymbol{\beta})X$ where $W(\boldsymbol{\beta})$ is a diagonal matrix whose ith diagonal entry is

$$\frac{1}{V(\mu_i)(g'(\mu_i))^2}.$$

2. Suppose that students answer questions on a test and that a specific student has an aptitude T. A particular question might have difficulty d_i and the student will get the answer correct only if $T > d_i$. Consider d_i fixed and $T \sim N(\mu, \sigma^2)$, then the probability that a randomly selected student will get the answer wrong is $p_i = \mathbb{P}(T < d_i)$.

   Show how you might model this situation using a probit regression model.

3. **Proportional odds in ordinal regression.** Suppose that Y_i takes values in the ordered set $\{1, \ldots, J\}$. Using a logit link, our model for $\gamma_{ij} = \mathbb{P}(Y_i \leq j)$ is

$$\gamma_{ij} = \operatorname{logit}^{-1}(\theta_j - \mathbf{x}_i^T\boldsymbol{\beta}).$$

   Thinking of γ_{ij} as a function of $\mathbf{x}_i$, we can rewrite it as $\gamma_j(\mathbf{x}_i) = \mathbb{P}(Y \leq j|\mathbf{x}_i)$.

   Recall the odds for an event A are given by $\mathbb{P}(A)/(1 - \mathbb{P}(A))$. By relative odds we mean the ratio of two odds. Show that the relative odds for $\{Y \leq j|\mathbf{x}_A\}$ and $\{Y \leq j|\mathbf{x}_B\}$ do not depend on j.