MAST90104: A First Course in Statistical Learning

Week 12 Lab and Workshop

1 Practical questions

1. In the `multinom` function from the `nnet` package, the response should be a factor with J levels or a matrix with J columns, which will be interpreted as counts for each of J 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.

2. 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 that $\mathbf{X} = (X_1, \ldots, X_k) \sim$ multinomial(n, π) where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$. Since $X_i \sim \text{bin}(n, \pi_i)$, we have $\mathbb{E}(X_i) = n\pi_i$ and $\text{Var}\,(X_i) = n\pi_i(1 - \pi_i)$. Show that for $i \neq j$, $\text{Cov}\,(X_i, X_j) = -n\pi_i\pi_j$.

   Hint: just as for the binomial, we can write a multinomial$(n,\ \boldsymbol{\pi})$ as the sum of n independent multinomial$(1,\ \pi)$ random variables.

   Alternative hint: $\text{Var}\,(X + Y) = \text{Var}\,(X) + \text{Var}\,(Y) + 2\text{Cov}\,(X, Y)$.

2. Suppose that $(X, Y, Z) \sim$ multinomial$(n, (p_1, p_2, p_3))$. Show that

$$Y|\{X = x\} \sim \text{binomial}(n - x, p_2/(1 - p_1)).$$

   Hence obtain $\mathbb{E}(Y|X = x)$.

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

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

   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.