MAST30027: Modern Applied Statistics

Week 6 Lab

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, $\mathbf{p}$).

1. 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. Treating the pneumonoconiosis status as response variable, 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.

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

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

5. The following three-way table refers to results of a case-control study about effects of cigarette smoking and coffee drinking on myocardial infarction (MI) or heart attack for a sample of men under 55 years of age.

Cups Coffee per Day	Cigarettes per Day							
	0		1-24		25-34		≥ 35	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
0	66	123	30	52	15	12	36	13
1-2	141	179	59	45	53	22	69	25
3-4	113	106	63	65	55	16	119	30
≥ 5	129	80	102	58	118	44	373	85

   Eight log-linear models with Poisson error have been fitted, with the results of residual deviances and residual degrees of freedom being given in the following table.

Model	Residual deviance	Residual df
`coffee + cigar + MI`	607.25	24
`coffee + cigar*MI`	394.43	21
`cigar + coffee*MI`	484.70	21
`MI + coffee*cigar`	271.40	15
`coffee*cigar + coffee*MI`	148.81	12
`coffee*cigar + cigar*MI`	58.55	12
`coffee*MI + cigar*MI`	271.88	18
`coffee*cigar + coffee*MI + cigar*MI`	11.17	9

(a) Give an interpretation of each of the following models.

i. `coffee + cigar + MI`

ii. `MI + coffee*cigar`

iii. `coffee*cigar + coffee*MI`

(b) i. Test the hypothesis that there is no association between `coffee` and `MI` when `cigar` level is given.

ii. Test the hypothesis that the association between `MI` and `cigar` is the same for all `coffee` levels. That is, test that there is no three-way interaction. (We could equivalently ask to test if the association between `MI` and `coffee` is the same for all `cigar` levels, or if the association between `cigar` and `coffee` is the same for all levels of `MI`.)

(c) Suppose we want to collapse the three-way table over `coffee` to investigate the relation between `cigar` and `MI`.

i. Give the value of the residual deviance that will be obtained if the 'no association' model is fitted to the collapsed table. Is the model adequate?

ii. Is it reasonable to collapse the table over `coffee`? Explain.