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).
   Solution:

```
> library(faraway)
> data(hsb)
> library(nnet)
> mmod <- multinom(prog ~ gender + race + ses + schtyp + read + write + math +
+                    science + socst, hsb, trace = FALSE)
> summary(mmod)

Call:
multinom(formula = prog ~ gender + race + ses + schtyp + read +
    write + math + science + socst, data = hsb, trace = FALSE)

Coefficients:
         (Intercept)  gendermale raceasian racehispanic racewhite     seslow
general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156 1.09864111
vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881 0.04747323
         sesmiddle schtyppublic        read       write       math    science
general  0.7029621    0.5845405 -0.04418353 -0.03627381 -0.1092888 0.10193746
vocation 1.1815808    2.0553336 -0.03481202 -0.03166001 -0.1139877 0.05229938
              socst
general  -0.01976995
vocation -0.08040129

Std. Errors:
         (Intercept) gendermale raceasian racehispanic racewhite    seslow
general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6066763
vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772
         sesmiddle schtyppublic       read      write       math    science
general  0.5045938    0.5642925 0.03103707 0.03381324 0.03522441 0.03274038
vocation 0.5700833    0.8348229 0.03422409 0.03585729 0.03885131 0.03424763
             socst
general  0.02712589
vocation 0.02938212

Residual Deviance: 305.8705
AIC: 357.8705
```

   (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.
   Solution: I just used the AIC, as provided by `step`.

```
> mmod2 <- step(mmod, scope=~., direction="backward", trace = FALSE)
trying - gender
trying - race
trying - ses
trying - schtyp
trying - read
trying - write
trying - math
trying - science
trying - socst
trying - gender
trying - ses
trying - schtyp
trying - read
trying - write
trying - math
trying - science
trying - socst
trying - ses
trying - schtyp
trying - read
trying - write
trying - math
trying - science
trying - socst
trying - ses
trying - schtyp
trying - read
trying - math
trying - science
trying - socst
trying - ses
trying - schtyp
trying - math
trying - science
trying - socst
> summary(mmod2)
Call:
multinom(formula = prog ~ ses + schtyp + math + science + socst,
    data = hsb, trace = FALSE)

Coefficients:
         (Intercept)      seslow sesmiddle schtyppublic       math    science
general     2.587029  0.87607389 0.6978995    0.6468812 -0.1212242 0.08209791
vocation    6.687272 -0.01569301 1.2065000    1.9955504 -0.1369641 0.03941237
               socst
general  -0.04441228
vocation -0.09363417

Std. Errors:
         (Intercept)    seslow sesmiddle schtyppublic       math    science
general     1.686492 0.5758781 0.4930330     0.545598 0.03213345 0.02787694
vocation    1.945363 0.6690861 0.5571202     0.812881 0.03591701 0.02864929
              socst
general  0.02344856
vocation 0.02586717
```

```
Residual Deviance: 315.5511
AIC: 343.5511
```

Compared to students from a high socioeconomic class, students from a low socioeconomic class are more likely to choose a general high school program, while students from a middle socioeconomic class are more likely to choose a general program but even more likely to choose a vocational program. It is interesting that students from a low socioeconimic class do not show more of an interest in vocational programs.

Students from public schools are are more likely to choose a general program and much more likely to choose a vocational program, than students from private schools.

High scores in maths and social sciences indicate a higher chance of choosing an academic program, while (curiously) high scores in science indicate a lower chance of choosing an academic program.

If you wish to use a chisquared test instead of the AIC, then you will have to separately fit all the candidate models, and then compare them using `anova`. For example:

```
> mmodXgender <- multinom(prog ~ race + ses + schtyp + read + write + math +
+                     science + socst, hsb, trace = FALSE)
> anova(mmod, mmodXgender)
```

```
Likelihood ratio tests of Multinomial Models

Response: prog
                                                                    Model
1          race + ses + schtyp + read + write + math + science + socst
2 gender + race + ses + schtyp + read + write + math + science + socst
  Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1       376   306.2857
2       374   305.8705 1 vs 2     2 0.415142 0.8125556
```

Clearly considering all possible variables to drop will take some time.

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

Solution:

```
> hsb[hsb$id==99,]
```

```
    id gender  race  ses schtyp    prog read write math science socst
102 99 female white high public general   47    59   56      66    61
```

```
> predict(mmod2, newdata = hsb[hsb$id==99,], type="probs")
```

```
  academic    general   vocation
0.64426309 0.27665609 0.07908082
```

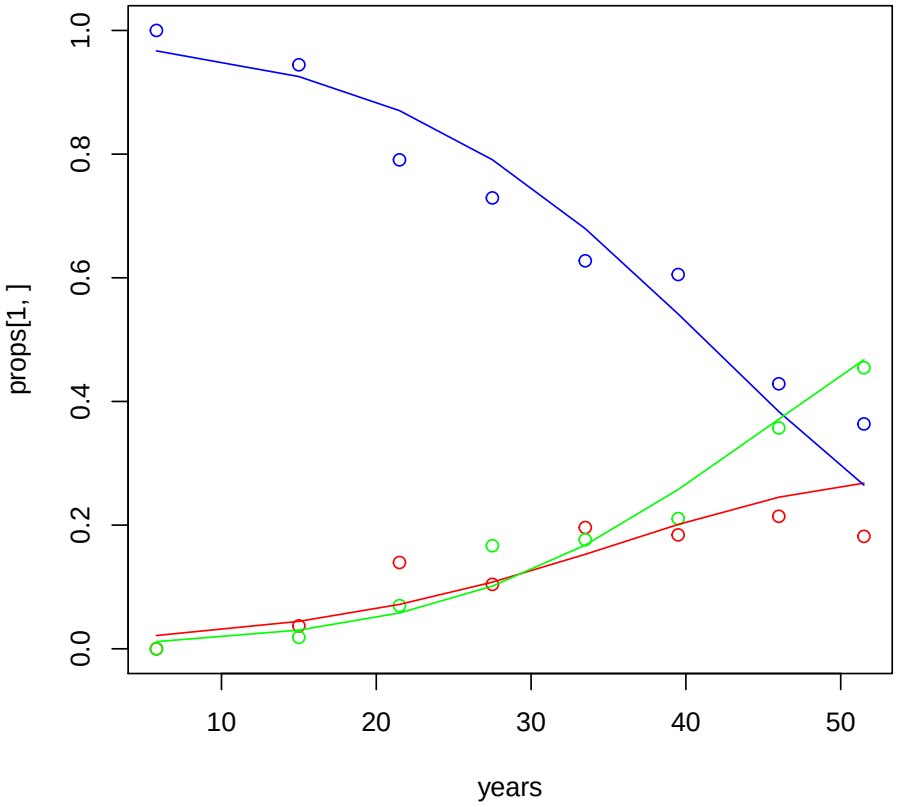
2. The `pneumo` data from the `faraway` package gives the number of coal miners classified by radiological examination into one of three categories of pneumonoconiosis and by the number of years spent working at the coal face divided into eight categories. Treating the pneumonoconiosis status as response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

Solution: First we have a look at the data. Then the data needs to be reformatted before we can use the `multinom` function to fit a model. The fit looks quite good.

```
> data(pneumo)
> counts <- xtabs(Freq ~ status + year, pneumo)
> (props <- prop.table(counts, 2))
```

```
        year
status          5.8         15       21.5       27.5       33.5       39.5
  mild   0.00000000 0.03703704 0.13953488 0.10416667 0.19607843 0.18421053
  normal 1.00000000 0.94444444 0.79069767 0.72916667 0.62745098 0.60526316
  severe 0.00000000 0.01851852 0.06976744 0.16666667 0.17647059 0.21052632
        year
```

```
status           46       51.5
  mild   0.21428571 0.18181818
  normal 0.42857143 0.36363636
  severe 0.35714286 0.45454545

> years <- c(5.8, 15, 21.5, 27.5, 33.5, 39.5, 46, 51.5)
> par(mfrow=c(1,1))
> plot(years, props[1,], col="red", ylim=c(0,1))
> points(years, props[2,], col="blue")
> points(years, props[3,], col="green")
> mmod <- multinom(t(counts) ~ years, trace=FALSE)
> summary(mmod)

Call:
multinom(formula = t(counts) ~ years, trace = FALSE)

Coefficients:
       (Intercept)       years
normal   4.2916723 -0.08356506
severe  -0.7681706  0.02572027

Std. Errors:
       (Intercept)      years
normal   0.5214110 0.01528044
severe   0.7377192 0.01976662

Residual Deviance: 417.4496
AIC: 425.4496

> fitted <- predict(mmod, newdata=list(year=years), type="probs")
> lines(years, fitted[,1], col="red")
> lines(years, fitted[,2], col="blue")
> lines(years, fitted[,3], col="green")
```



For a miner with 25 year down pit we have the following fitted probabilities

```
> predict(mmod, newdata=list(years=25), type="probs")

      mild     normal     severe
0.09148821 0.82778696 0.08072483
```

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

   Solution: If $\mathbf{X} \sim$ multinomial$(1, \pi)$ then for $i \neq j$ we have $\mathbb{E}X_iX_j = 0$ and thus $\text{Cov}\,(X_i, X_j) = 0 - \mathbb{E}X_i\mathbb{E}X_j = -\pi_i\pi_j$. If $\mathbf{X} \sim$ multinomial(n, π) then it can be written as the sum of n independent multinomial$(1, \pi)$, whence we can multiply the covariances by n to get the result.

   Alternatively, if we add X_i and X_j it is just as if we combined these two cases into a single case with probability $\pi_i + \pi_j$. Thus

$$\begin{aligned} \text{Cov}\,(X_i, X_j) &= \tfrac{1}{2}(\text{Var}\,(X_i + X_j) - \text{Var}\, X_i - \text{Var}\, X_j) \\ &= \tfrac{1}{2}(n(\pi_i + \pi_j)(1 - \pi_i - \pi_j) - n\pi_i(1-\pi_i) - n\pi_j(1-\pi_j)) \\ &= -n\pi_i\pi_j \end{aligned}$$

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

Solution:

$$\begin{aligned}
\mathbb{P}(Y=y|X=x) &= \mathbb{P}(Y=y, Z=n-x-y|X=x) \\
&= \mathbb{P}(X=x, Y=y, Z=n-x-y)/\mathbb{P}(X=x) \\
&= \frac{n!/(x!y!(n-x-y)!)p_1^x p_2^y p_3^{n-x-y}}{n!/(x!(n-x)!)p_1^x(1-p_1)^{n-x}} \\
&= \frac{(n-x)!}{y!(n-x-y)!}\left(\frac{p_2}{1-p_1}\right)^y\left(\frac{p_3}{1-p_1}\right)^{n-x-y}
\end{aligned}$$

But $p_3/(1-p_1) = 1-p_2/(1-p_1)$, so this is of the right form.

We get immediately that $\mathbb{E}(Y|X=x) = (n-x)p_2/(1-p_1)$. That is, given $X=x$, we divvy up the remaining $n-x$ trials between Y and Z proportionately to p_2 and p_3.

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`

Solution:

i. The three factors `coffee`, `cigar` and `MI` are mutually independent;

ii. `MI` is independent of `coffee` and `cigar` together;

iii. Given `coffee`, `cigar` and MI are independent (unrelated).

(b) i. Test the hypothesis that there is no association between `coffee` and `MI` when `cigar` level is given.
Solution: This is to test the adequacy of `coffee*cigar+cigar*MI`. The residual deviance is 58.55 on 12 df. $\chi^2_{12,0.95} = 21.026$ so `coffee*cigar+cigar*MI` is not adequate, and we will reject the hypothesis.

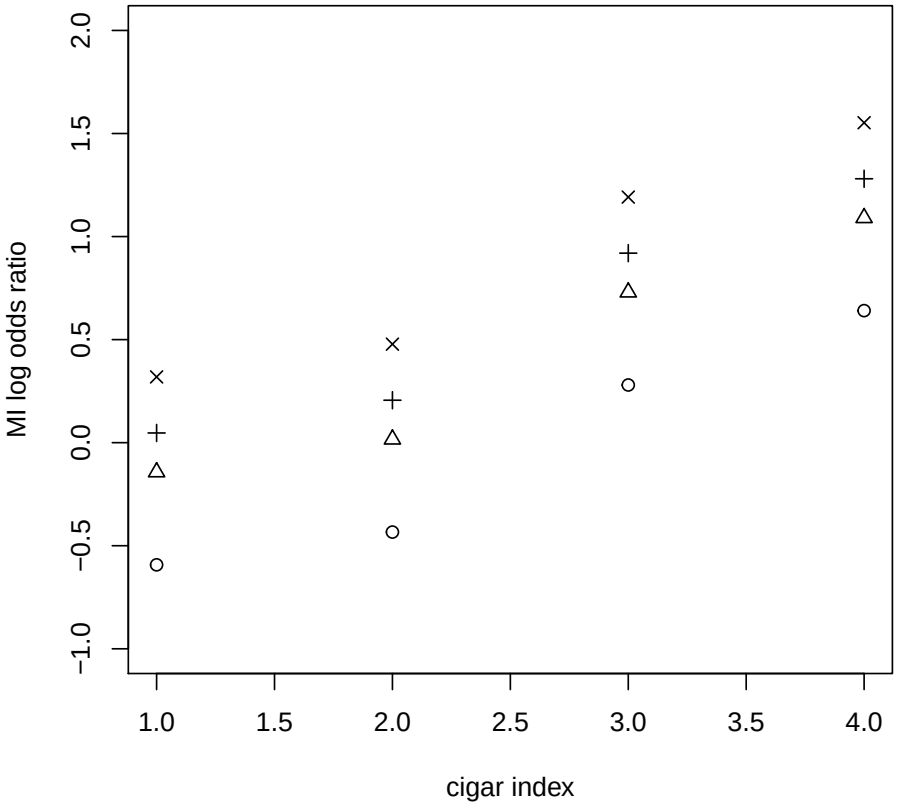
We can think of this test as LR test between the fitted model and the full/saturated model, which has deviance zero and degrees of freedom zero. In this case it also makes sense to compare the fitted model to the model `coffee*cigar + coffee*MI + cigar*MI`, which is asking the question "is there in addition a significant pairwise interaction between coffee and MI". In theory, if there was no pairwise interaction between coffee and MI, but there was a three-way interaction, then a test of `coffee*cigar + coffee*MI + cigar*MI` against `coffee*cigar + cigar*MI` would not be significant, whereas a test of `coffee*cigar + cigar*MI` against the full model (our model adequacy test) would be significant. In practice however, it is unlikely that we will have a significant three-way interaction without all the corresponding pair-wise interactions being significant.

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.
Solution: On the face of it, this seems to be asking if MI and cigar are independent of coffee (that is, is the model `coffee + MI*cigar` adequate), but it is actually trying to ask if there is no three-way interaction between MI, cigar and coffee. That is, we should test the adequacy of `coffee*cigar + coffee*MI + cigar*MI`. The deviance is 11.17 on 9 df and $\chi^2_{9,0.95} = 16.919$, so `coffee*cigar + coffee*MI + cigar*MI` is adequate, and we cannot reject the hypothesis.
To see exactly what is meant by the question, we can have a look at the fitted values of the model with all two-way interactions.

```
> cigar <- gl(4, 8, 32, labels=c("0", "1-24", "25-34", ">=35"))
> MI <- gl(2, 4, 32, labels=c("case", "control"))
> coffee <- gl(4, 1, 32, label=c("0", "1-2", "3-4", ">=5"))
> y <- c(66,141,113,129, 123,179,106,80, 30,59, 63,102, 52,45,65,58,
+        15, 53, 55,118,  12, 22, 16,44, 36,69,119,373, 13,25,30,85)
> ftable(xtabs(y ~ cigar + MI + coffee))
              coffee   0 1-2 3-4 >=5
cigar MI
0     case            66 141 113 129
      control        123 179 106  80
1-24  case            30  59  63 102
      control         52  45  65  58
25-34 case            15  53  55 118
      control         12  22  16  44
>=35  case            36  69 119 373
      control         13  25  30  85
> model <- glm(y ~ (cigar + MI + coffee)^2, family="poisson")
```

To see how MI and cigar interact, we can look at the log odds ratio for MI = case versus MI = control (which is just the difference in η values), for different values of cigar. We do this for each value of coffee and plot them:

```
> fitted <- predict(model)
> plot(   fitted[MI=="case" & coffee=="0"] - fitted[MI=="control" & coffee=="0"],
+         ylim=c(-1,2), pch=1, xlab="cigar index", ylab="MI log odds ratio")
> points( fitted[MI=="case" & coffee=="1-2"] - fitted[MI=="control" & coffee=="1-2"],
+         pch=2)
> points( fitted[MI=="case" & coffee=="3-4"] - fitted[MI=="control" & coffee=="3-4"],
+         pch=3)
> points( fitted[MI=="case" & coffee==">=5"] - fitted[MI=="control" & coffee==">=5"],
+         pch=4)
```



We see that for all levels of cigar the effect of coffee is the same. Similarly, if we take a log odds ratio for different levels of cigar, we will see that for all levels of MI the effect of coffee is the same.

(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?
Solution: Collapsing the table over coffee is equivalent to omitting the coffee factor from the log poisson model. We wish to compare the additive model `MI + cigar` to the full model `MI*cigar`.

```
> mod1 <- glm(y ~ MI + cigar, family=poisson)
> c(deviance(mod1), mod1$df.residual)
[1] 941.8928  27.0000
> mod2 <- glm(y ~ MI*cigar, family=poisson)
> c(deviance(mod2), mod2$df.residual)
[1] 729.0725  24.0000
> c(deviance(mod1) - deviance(mod2), mod1$df.residual - mod2$df.residual)
[1] 212.8202   3.0000
```

An observed value of 212.82 with 3 d.f. is clearly significant ($\chi^2_{0.95,3} = 7.815 < 212.82$), so we reject the hypothesis that MI and cigar are independent (in the absence of coffee). Alternatively we can combine the counts for different levels of coffee. In this case the full model is saturated (so has zero deviance and zero d.f.)

```
> cigar <- gl(4, 2, 8, labels=c("0", "1-24", "25-34", ">=35"))
> MI <- gl(2, 1, 8, labels=c("case", "control"))
> z <- colSums(matrix(y, nrow=4))
```

```
> mod3 <- glm(z ~ MI + cigar, family=poisson)
> c(deviance(mod3), mod3$df.residual)
[1] 212.8202   3.0000
```

In fact, this result can be obtained without fitting any new models, by taking the difference in the deviances for the models `coffee + cigar + MI` and `coffee + cigar*MI`. This is analogous to the additivity of the residual sums of squares coming from the main effects in a three-factor linear model.

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

Solution:

Not reasonable to collapse. While the 3-factor interaction is not significant, both of the 2-factor interactions involving `coffee` (`coffee:cigar` and `coffee:MI`) are significant.