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).
   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 sesmiddle
general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6066763 0.5045938
vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772 0.5700833
           schtyppublic    read      write       math    science      socst
general     0.5642925 0.03103707 0.03381324 0.03522441 0.03274038 0.02712589
vocation    0.8348229 0.03422409 0.03585729 0.03885131 0.03424763 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: Use backward selection with AIC

```
> mmod2 <- step(mmod, scope=~.,  trace = FALSE, direction = "backward")

> 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      306.28565
2        374      305.87051   1 vs 2   2   0.41514199   0.81255555
```

(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
0.644263088 0.276656088 0.079080824
```

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.

   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
mild   0.000000000 0.037037037 0.139534884 0.104166667 0.196078431
normal 1.000000000 0.944444444 0.790697674 0.729166667 0.627450980
severe 0.000000000 0.018518519 0.069767442 0.166666667 0.176470588
year
status     39.5          46        51.5
mild   0.184210526 0.214285714 0.181818182
normal 0.605263158 0.428571429 0.363636364
severe 0.210526316 0.357142857 0.454545455
> 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  4.29167227 -0.083565062
severe -0.76817058  0.025720269

Std. Errors:
        (Intercept)     years
normal  0.52141098 0.015280443
severe  0.73771918 0.019766616

Residual Deviance: 417.44956
AIC: 425.44956
```

   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
0.091488209 0.827786959 0.080724832
```

   In the model above we had eight multinomial observations, with the number of trials equal to 98, 54, 43, 48, 51, 38, 28, 11. Each of these multinomials can be regarded as the sum of a number of independent multinomials each based on a single trial (just as a binomial is a sum of independent Bernoulli random variables). If we treat the data this way and fit a multinomial logistic regression, we get the same model.

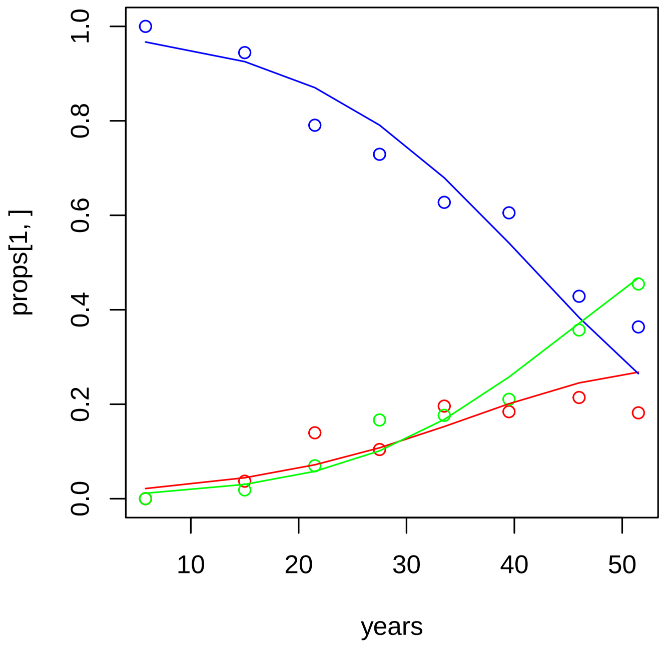

Figure 1: Fitted probability of the three outcomes vs length of service

```
> pneumo2 <- data.frame(status = rep(pneumo$status, pneumo$Freq),
year = rep(pneumo$year, pneumo$Freq))
> mmod2 <- multinom(status ~ year, data = pneumo2, trace = FALSE)
> summary(mmod2)
Call:
multinom(formula = status ~ year, data = pneumo2, trace = FALSE)

Coefficients:
        (Intercept)        year
normal   4.29167227 -0.083565062
severe -0.76817058  0.025720269

Std. Errors:
        (Intercept)        year
normal   0.52141098 0.015280443
severe   0.73771918 0.019766616

Residual Deviance: 417.44956
AIC: 425.44956
```

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

Solution: First we convert `status` into an ordered factor (take care to get the order correct), then use the `polr` function.

```
> # b
> pneumo2$status <- ordered(pneumo2$status, levels=c("normal", "mild", "severe"))
> library(MASS)
> omod <- polr(status ~ year, pneumo2)
> summary(omod)
```

```
Re-fitting to get Hessian

Call:
polr(formula = status ~ year, data = pneumo2)

Coefficients:
      Value    Std. Error t value
year 0.095904   0.011938  8.0336

Intercepts:
              Value  Std. Error t value
normal|mild  3.95584  0.40969    9.65576
mild|severe  4.86905  0.44110   11.03833

Residual Deviance: 416.91883
AIC: 422.91883
```

Note: An alternative solution that will give the same answer using the original data is

```
pneumo$status <- ordered(pneumo$status, levels  = c('normal','mild','severe'))
omod2 <- polr(status~year, data = pneumo, weights = Freq)
```

The fit looks good (Figure 2), and the AIC for this model is slightly smaller than that for the multinomial logistic regression model, so we prefer it.

```
> plot(years, props[1,], col="red", ylim=c(0,1))
> points(years, props[2,], col="blue")
> points(years, props[3,], col="green")
> fitted <- predict(omod, newdata=list(year=years), type="probs")
> lines(years, fitted[,1], col="blue")
> lines(years, fitted[,2], col="red")
> lines(years, fitted[,3], col="green")
```

For a miner with 25 years exposure we have the following fitted probabilities

```
> predict(omod, newdata=list(year=25), type="probs")
    normal        mild      severe
0.826100955 0.096014743 0.077884302
```

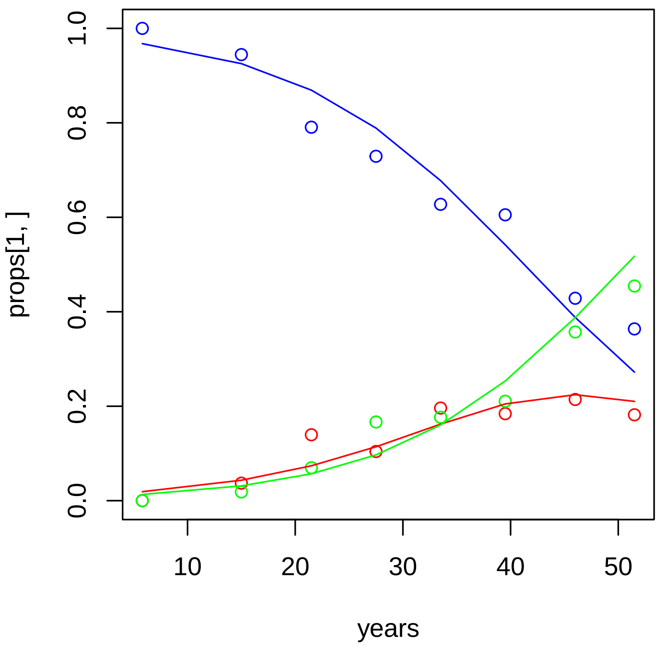

Figure 2: Fitted probability of the three outcomes vs length of service from the ordinal model