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 backward elimination to reduce the model to one where all predictors are statistically significant. Give an interpretation of the resulting model.

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

https://www.coursehero.com/file/21838991/mast30027-wk6soltn/

This study resource was shared via CourseHero.com

```
> 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
```

This study resource was shared via CourseHero.com

https://www.coursehero.com/file/21838991/mast30027-wk6soltn/

```
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)
                                                                      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           NA       NA        NA
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.

(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)
> (obs <- prop.table(xtabs(Freq ~ status + year, pneumo), 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
```

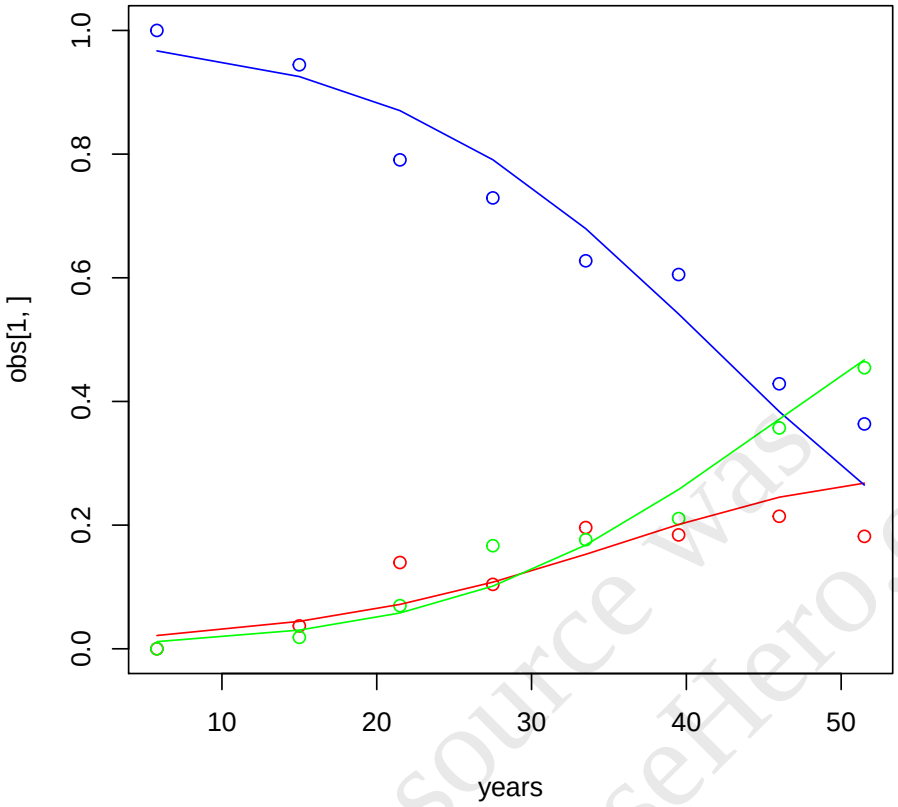

https://www.coursehero.com/file/21838991/mast30027-wk6soltn/

This study resource was shared via CourseHero.com

```
> years <- c(5.8, 15, 21.5, 27.5, 33.5, 39.5, 46, 51.5)
> par(mfrow=c(1,1))
> plot(years, obs[1,], col="red", ylim=c(0,1))
> points(years, obs[2,], col="blue")
> points(years, obs[3,], col="green")
> pneumo2 <- data.frame(status = rep(pneumo$status, pneumo$Freq),
+                       year = rep(pneumo$year, pneumo$Freq))
> mmod <- multinom(status ~ year, data = pneumo2, trace = FALSE)
> summary(mmod)

Call:
multinom(formula = status ~ year, data = pneumo2, trace = FALSE)

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

Std. Errors:
       (Intercept)       year
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")
```

This study resource was shared via CourseHero.com

https://www.coursehero.com/file/21838991/mast30027-wk6soltn/



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

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

```
      mild     normal     severe
0.09148821 0.82778696 0.08072483
```

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

Solution:

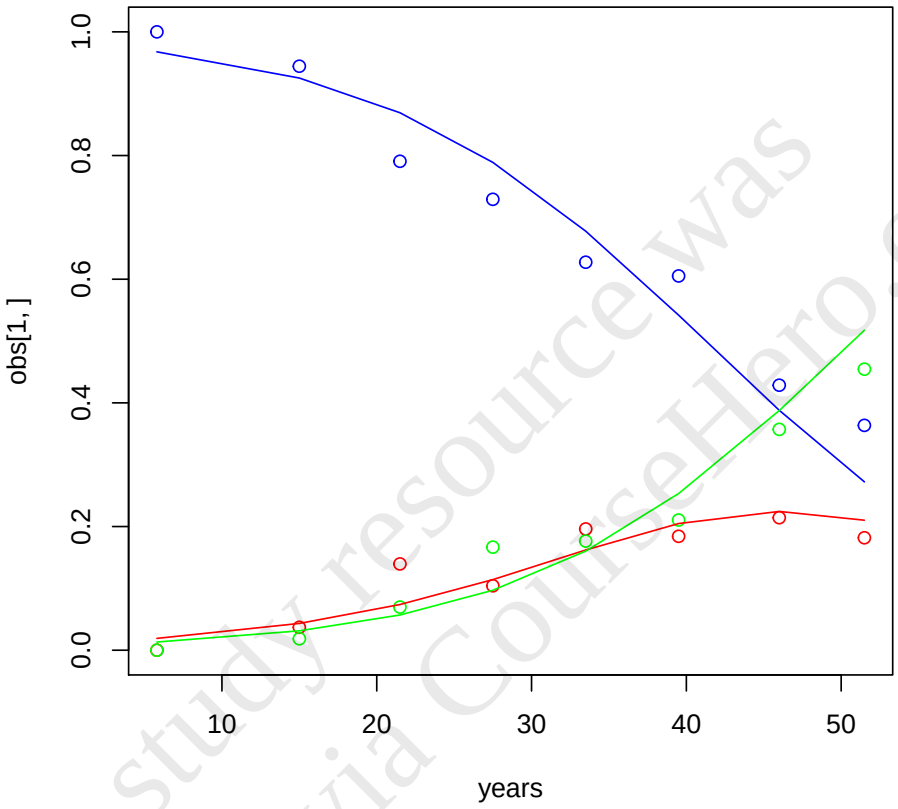
First we convert `status` into an ordered factor (take care to get the order correct), then use the `polr` function. The fit looks good, and the AIC for this model is slightly smaller than that for the multinomial logistic regression model, so we prefer it.

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

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

Coefficients:
      Value Std. Error t value
year 0.0959    0.01194   8.034

Intercepts:
            Value   Std. Error t value
normal|mild  3.9558  0.4097     9.6558
mild|severe  4.8690  0.4411    11.0383
```


https://www.coursehero.com/file/21838991/mast30027-wk6soltn/

```
Residual Deviance: 416.9188
AIC: 422.9188
> plot(years, obs[1,], col="red", ylim=c(0,1))
> points(years, obs[2,], col="blue")
> points(years, obs[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
0.82610096 0.09601474 0.07788430
```

3. Suppose that $\mathbf{X} = (X_1, \ldots, X_k) \sim$ multinomial(n, π) where $\pi = (\pi_1, \ldots, \pi_k)$. 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$(n,\ \pi)$ 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

$$\text{Cov}\,(X_i, X_j) \quad = \quad \tfrac{1}{2}(\text{Var}\,(X_i + X_j) - \text{Var}\, X_i - \text{Var}\, X_j)$$

This study resource was Shared via CourseHero.com

https://www.coursehero.com/file/21838991/mast30027-wk6soltn/

$$
\begin{aligned}
&= \tfrac{1}{2}(n(\pi_i+\pi_j)(1-\pi_i-\pi_j) - n\pi_i(1-\pi_i) - n\pi_j(1-\pi_j)) \\
&= -n\pi_i\pi_j
\end{aligned}
$$

4. Suppose that $(X, Y, Z) \sim \text{multinomial}(n, (p_1, p_2, p_3))$. 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. **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 $\gamma_{ij} = \mathbb{P}(Y_i \le j)$ is

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

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

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 $\{Y \le j|\mathbf{x}_A\}$ and $\{Y \le j|\mathbf{x}_B\}$ do not depend on j. For this reason, this model is often called the proportional odds model.

This independence of the odds ratio on j can be used to check the suitability of the model. For $j = 1$ and $j = 2$, calculate the difference between the *observed* log odds at income level 1.5 and levels 4, 6, 8, 9.5, ... (the values in the vector `inca`). Do they look roughly the same?

Solution: The odds ratio is

$$
\begin{aligned}
\frac{\frac{\mathbb{P}(Y\le j|\mathbf{x}_A)}{1-\mathbb{P}(Y\le j|\mathbf{x}_A)}}{\frac{\mathbb{P}(Y\le j|\mathbf{x}_B)}{1-\mathbb{P}(Y\le j|\mathbf{x}_B)}} &= \frac{\exp(\text{logit}(\mathbb{P}(Y \le j|\mathbf{x}_A)}{\exp(\text{logit}(\mathbb{P}(Y \le j|\mathbf{x}_B)} \\
&= \frac{\exp(\theta_j - \mathbf{x}_A^T\beta)}{\exp(\theta_j - \mathbf{x}_B^T\beta)} \\
&= \exp(-(\mathbf{x}_A - \mathbf{x}_B)^T\beta)
\end{aligned}
$$

which does not depend on j, as required.

Note that the difference between the log odds is just $-(\mathbf{x}_A - \mathbf{x}_B)^T\beta$.

We check this for the `nes96` data. First we calculate `log_o1`, the observed log odds of being Democrat, and `log_o2`, the observed log odds of being Democrat or Independent. We do this for each income level.

```
> data(nes96)
> sPID <- nes96$PID
> levels(sPID) <- c("Democrat","Democrat","Independent","Independent",
+                   "Independent","Republican","Republican")
> inca <- c(1.5,4,6,8,9.5,10.5,11.5,12.5,13.5,14.5,16,18.5,21,23.5,
+           27.5,32.5,37.5,42.5,47.5,55,67.5,82.5,97.5,115)
> nincome <- inca[unclass(nes96$income)]
> obs <- prop.table(table(nincome, sPID), 1)
```

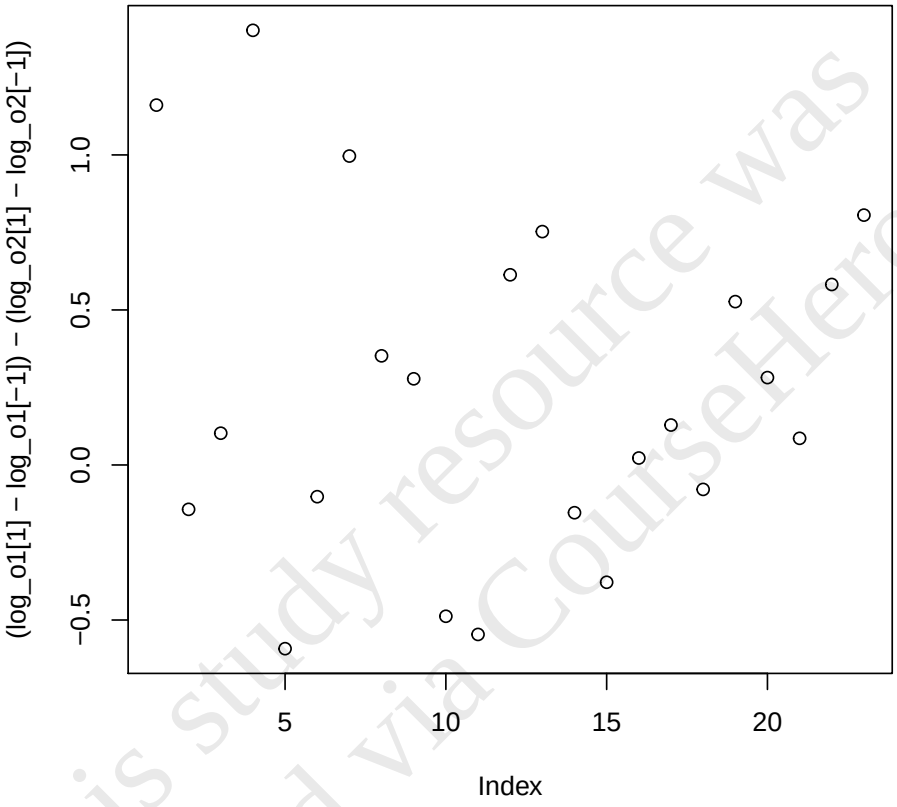

https://www.coursehero.com/file/21838991/mast30027-wk6soltn/

This study resource was shared via CourseHero.com

```
> p1 <- obs[,1]
> p2 <- obs[,1] + obs[,2]
> log_o1 <- log(p1/(1-p1))
> log_o2 <- log(p2/(1-p2))
```

For the log odds of being Democrat, the difference between income level 1.5 and the rest is just given by `log_o1[1] - log_o1[-1]`. This should look the same as `log_o2[1] - log_o2[-1]`, which we check by taking differences. These differences are not particularly close to zero, but at least they don't display any trend.

```
> plot((log_o1[1] - log_o1[-1]) - (log_o2[1] - log_o2[-1]))
```



This study resource was shared via CourseHero.com

https://www.coursehero.com/file/21838991/mast30027-wk6soltn/