# MA678 homework 06

Multinomial Regression

Anna Cook

October 22, 2020

## Multinomial logit:

Using the individual-level survey data from the 2000 National Election Study (data in folder NES), predict party identification (which is on a five-point scale) using ideology and demographics with an ordered multinomial logit model.

```
## Call:
## polr(formula = factor(partyid7) ~ age + gender + race + ideo,
##     data = nes, Hess = TRUE)
##
## Coefficients:
##         age      gender        race        ideo
## -0.01193745 -0.17256010 -0.25333872  0.42067639
##
## Intercepts:
##        1|2        2|3        3|4        4|5        5|6        6|7
## -1.3901593 -0.3133816  0.2740848  0.6796139  1.2965932  2.3213684
##
## Residual Deviance: 44523.29
## AIC: 44543.29
## (22681 observations deleted due to missingness)
```

1. Summarize the parameter estimates numerically and also graphically.

```
summary(fit)
```

```
## Call:
## polr(formula = factor(partyid7) ~ age + gender + race + ideo,
##     data = nes, Hess = TRUE)
##
## Coefficients:
##           Value Std. Error t value
## age    -0.01194   0.000982 -12.156
## gender -0.17256   0.032103  -5.375
## race   -0.25334   0.015284 -16.575
## ideo    0.42068   0.009925  42.387
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -1.3902   0.0795   -17.4885
## 2|3  -0.3134   0.0780    -4.0193
## 3|4   0.2741   0.0780     3.5154
## 4|5   0.6796   0.0783     8.6817
```

```
## 5|6   1.2966   0.0791    16.4009
## 6|7   2.3214   0.0814    28.5017
##
## Residual Deviance: 44523.29
## AIC: 44543.29
## (22681 observations deleted due to missingness)
```

2. Explain the results from the fitted model.

```
#a person is more likely to be higher in partyid if they are younger, have a lower indicator value for
```

3. Use a binned residual plot to assess the fit of the model.

## (Optional) Choice models:

Using the individual-level survey data from the election example described in Section 10.9 (data available in the folder NES),

1. fit a logistic regression model for the choice of supporting Democrats or Republicans. Then interpret the output from this regression in terms of a utility/choice model.
2. Repeat the previous exercise but now with three options: Democrat, no opinion, Republican. That is, fit an ordered logit model and then express it as a utility/choice mode

## Contingency table and ordered logit model

In a prospective study of a new living attenuated recombinant vaccine for influenza, patients were randomly allocated to two groups, one of which was given the new vaccine and the other a saline placebo. The responses were titre levels of hemagglutinin inhibiting antibody found in the blood six weeks after vaccination; they were categorized as "small", "medium" or "large".

| treatment | small | moderate | large | Total |
|---|---|---|---|---|
| placebo | 25 | 8 | 5 | 38 |
| vaccine | 6 | 18 | 11 | 35 |

The cell frequencies in the rows of table are constrained to add to the number of subjects in each treatment group (35 and 38 respectively). We want to know if the pattern of responses is the same for each treatment group.

1. Using a chisqure test and an appropriate log-linear model, test the hypothesis that the distribution of responses is the same for the placebo and vaccine groups.

```
chisq.test(data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 17.648, df = 2, p-value = 0.0001472
```

```
stan_glm(Freq ~ treatment + levels, data = df, refresh = 0)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      Freq ~ treatment + levels
##  observations: 6
##  predictors:   4
## ------
```

```
##                 Median MAD_SD
## (Intercept)      15.5    8.0
## treatmentvaccine -1.0    8.2
## levelsmoderate   -2.0   10.0
## levelslarge      -7.0    9.8
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 10.4    3.9
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

2. For the model corresponding to the hypothesis of homogeniety of response distributions, calculate the fitted values, the Pearson and deviance residuals, and the goodness of fit statistics $X^2$ and $D$. Which of the cells of the table contribute most to $X^2$ and $D$? Explain and interpret these results.

```
chisq <- chisq.test(data)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 17.648, df = 2, p-value = 0.0001472
```

3. Re-analyze these data using ordered logit model (use `polr`) to estiamte the cut-points of a latent continuous response varaible and to estimate a location shift between the two treatment groups. Sketch a rough diagram to illustrate the model which forms the conceptual base for this analysis.

# High School and Beyond

The hsb data was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the 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.

```
data(hsb)
?hsb
```

1. Fit a trinomial response model with the other relevant variables as predictors (untransformed).

```
fit <- multinom(factor(prog)~gender + race + ses + read + write + math + science + socst + schtyp, data
```

```
## # weights:  42 (26 variable)
## initial  value 219.722458
## iter  10 value 171.814970
## iter  20 value 153.793692
## iter  30 value 152.935260
## final  value 152.935256
## converged
```

```
fit
```

```
## Call:
```

```
## multinom(formula = factor(prog) ~ gender + race + ses + read +
##     write + math + science + socst + schtyp, data = hsb, hess = TRUE)
##
## 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        read       write       math    science       socst
## general  0.7029621 -0.04418353 -0.03627381 -0.1092888 0.10193746 -0.01976995
## vocation 1.1815808 -0.03481202 -0.03166001 -0.1139877 0.05229938 -0.08040129
##          schtyppublic
## general     0.5845405
## vocation    2.0553336
##
## Residual Deviance: 305.8705
## AIC: 357.8705
```

2. For the student with id 99, compute the predicted probabilities of the three possible choices.

```
subset <- subset(hsb, hsb$id==99)
fitted(fit)[102,]
```

```
##  academic   general  vocation
## 0.5076752 0.3753090 0.1170158
```

## Happiness

Data were collected from 39 students in a University of Chicago MBA class and may be found in the dataset happy.

```
library(faraway)
data(happy)
?happy
```

1. Build a model for the level of happiness as a function of the other variables.

```
fit <- polr(factor(happy)~money + sex + love + work, data = happy, Hess = TRUE)
fit
```

```
## Call:
## polr(formula = factor(happy) ~ money + sex + love + work, data = happy,
##     Hess = TRUE)
##
## Coefficients:
##      money        sex       love       work
##  0.0224593 -0.4734369  3.6076452  0.8875135
##
## Intercepts:
##       2|3       3|4       4|5       5|6       6|7       7|8       8|9      9|10
##  5.470845  6.468394  9.159127 10.972524 11.511333 13.543305 17.290890 19.011197
##
## Residual Deviance: 94.86029
## AIC: 118.8603
```

2. Interpret the parameters of your chosen model.

```r
# a person is most likely to be happier if they have more money, more love, a good job, and not satisfa
```

3. Predict the happiness distribution for subject whose parents earn $30,000 a year, who is lonely, not sexually active and has no job.

```r
newdata = data.frame(money=30, sex=0,love=1,work=1)
predict(fit, newdata=newdata, type = "probs")
```

```
##            2            3            4            5            6            7
## 5.749087e-01 2.108348e-01 1.960962e-01 1.515266e-02 1.250656e-03 1.526336e-03
##            8            9           10
## 2.252137e-04 4.465166e-06 9.736048e-07
```

## newspaper survey on Vietnam War

A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset `uncviet`. Treat the opinion as the response and the sex and year as predictors. Build a proportional odds model, giving an interpretation to the estimates.

```r
data(uncviet)
?uncviet
fit <- model <- polr(policy ~ sex + year, weights=y, data=uncviet)
fit
```

```
## Call:
## polr(formula = policy ~ sex + year, data = uncviet, weights = y)
##
## Coefficients:
##    sexMale   yearGrad yearJunior yearSenior   yearSoph
## -0.6470352  1.1769887  0.3964211  0.5443945  0.1315047
##
## Intercepts:
##         A|B         B|C         C|D
## -1.10979578 -0.01304875  2.44169665
##
## Residual Deviance: 7757.056
## AIC: 7773.056
```

```r
# students are more likely to want the US to withdraw from Vietnam if they are older and if they are fe
```

## pneumonoconiosis of coal miners

The pneumo data 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.

```r
library(faraway)
data(pneumo,package="faraway")
?pneumo
```

```
## Help on topic 'pneumo' was found in the following packages:
##
##   Package               Library
##   VGAM                  /Library/Frameworks/R.framework/Versions/4.0/Resources/library
```

```
##   faraway               /Library/Frameworks/R.framework/Versions/4.0/Resources/library
##
##
## Using the first match ...
```

1. 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.

```r
fit <- multinom(status ~ year, data = pneumo, Hess = TRUE)
```

```
## # weights:  9 (4 variable)
## initial  value 26.366695
## final  value 26.366695
## converged
```

```r
fit
```

```
## Call:
## multinom(formula = status ~ year, data = pneumo, Hess = TRUE)
##
## Coefficients:
##         (Intercept)         year
## normal 2.109424e-15 2.486900e-14
## severe 2.664535e-15 3.552714e-14
##
## Residual Deviance: 52.73339
## AIC: 60.73339
```

```r
newdata=data.frame(year=25)
predict(fit, newdata=newdata, type="probs")
```

```
##      mild    normal    severe
## 0.3333333 0.3333333 0.3333333
```

2. Repeat the analysis with the pneumonoconiosis status being treated as ordinal.

```r
fit <- polr(status ~ year, data = pneumo, Hess = TRUE)
fit
```

```
## Call:
## polr(formula = status ~ year, data = pneumo, Hess = TRUE)
##
## Coefficients:
##         year
## 4.340705e-11
##
## Intercepts:
##   mild|normal normal|severe
##    -0.6931472     0.6931472
##
## Residual Deviance: 52.73339
## AIC: 58.73339
```

```r
newdata=data.frame(year=25)
predict(fit, newdata=newdata, type="probs")
```

```
##      mild    normal    severe
## 0.3333333 0.3333333 0.3333333
```

3.Now treat the response variable as hierarchical with top level indicating whether the miner has the disease and the second level indicating, given they have the disease, whether they have a moderate or severe case.

4. Compare the three analyses.

```
# the first two are the same
```

## (optional) Multinomial choice models:

Pardoe and Simonton (2006) fit a discrete choice model to predict winners of the Academy Awards. Their data are in the folder AcademyAwards.

| name | description |
|---|---|
| No | unique nominee identifier |
| Year | movie release year (not ceremony year) |
| Comp | identifier for year/category |
| Name | short nominee name |
| PP | best picture indicator |
| DD | best director indicator |
| MM | lead actor indicator |
| FF | lead actress indicator |
| Ch | 1 if win, 2 if lose |
| Movie | short movie name |
| Nom | total oscar nominations |
| Pic | picture nom |
| Dir | director nom |
| Aml | actor male lead nom |
| Afl | actor female lead nom |
| Ams | actor male supporting nom |
| Afs | actor female supporting nom |
| Scr | screenplay nom |
| Cin | cinematography nom |
| Art | art direction nom |
| Cos | costume nom |
| Sco | score nom |
| Son | song nom |
| Edi | editing nom |
| Sou | sound mixing nom |
| For | foreign nom |
| Anf | animated feature nom |
| Eff | sound editing/visual effects nom |
| Mak | makeup nom |
| Dan | dance nom |
| AD | assistant director nom |
| PrNl | previous lead actor nominations |
| PrWl | previous lead actor wins |
| PrNs | previous supporting actor nominations |
| PrWs | previous supporting actor wins |
| PrN | total previous actor/director nominations |
| PrW | total previous actor/director wins |
| Gdr | golden globe drama win |
| Gmc | golden globe musical/comedy win |
| Gd | golden globe director win |
| Gm1 | golden globe male lead actor drama win |

| name | description |
|---|---|
| Gm2 | golden globe male lead actor musical/comedy win |
| Gf1 | golden globe female lead actor drama win |
| Gf2 | golden globe female lead actor musical/comedy win |
| PGA | producer's guild of america win |
| DGA | director's guild of america win |
| SAM | screen actor's guild male win |
| SAF | screen actor's guild female win |
| PN | PP*Nom |
| PD | PP*Dir |
| DN | DD*Nom |
| DP | DD*Pic |
| DPrN | DD*PrN |
| DPrW | DD*PrW |
| MN | MM*Nom |
| MP | MM*Pic |
| MPrN | MM*PrNl |
| MPrW | MM*PrWl |
| FN | FF*Nom |
| FP | FF*Pic |
| FPrN | FF*PrNl |
| FPrW | FF*PrWl |

1. Fit your own model to these data.
2. Display the fitted model on a plot that also shows the data.
3. Make a plot displaying the uncertainty in inferences from the fitted model.