MA678 Homework 6

Reese Mullen

11/5/2024

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.

```
## Warning in data.table::fread("/Users/reesemullen/Desktop/Statistical
## Practice/MA 678/678-HW/nes.txt"): Detected 70 column names but the data has 71
## columns (i.e. invalid file). Added 1 extra default column name for the first
## column which is guessed to be row names or an index. Use setnames() afterwards
## if this guess is not correct, or fix the file write command that created the
## file to create a valid file.
```

1. Summarize the parameter estimates numerically and also graphically.

```
party7 = nes_data_comp$partyid7
nes_data_comp <- nes_data_comp[!is.na(levels(party7)[party7]),]
nesmod1 <- polr(factor(partyid7) ~ ideo + age + gender + race + south, Hess = TRUE, data = nes_data_comp
summary(nesmod1)
```

```
## Call:
## polr(formula = factor(partyid7) ~ ideo + age + gender + race +
##     south, data = nes_data_comp, Hess = TRUE)
##
## Coefficients:
##                         Value Std. Error t value
## ideomoderate          0.95244   0.330457  2.8822
## ideoconservative      1.94258   0.181627 10.6954
## age                  -0.01343   0.004948 -2.7140
## genderfemale         -0.39117   0.155717 -2.5121
## raceblack            -1.79778   0.277314 -6.4828
## raceasian             0.12366   0.544593  0.2271
## racenative american -0.14122   0.368345 -0.3834
## racehispanic         -0.63028   0.297507 -2.1185
## south1                0.21075   0.175175  1.2031
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -1.4013  0.3061    -4.5779
## 2|3 -0.5501  0.2981    -1.8457
```

```
## 3|4  0.2422  0.2977     0.8136
## 4|5  0.6280  0.3002     2.0922
## 5|6  1.4280  0.3051     4.6811
## 6|7  2.4084  0.3163     7.6152
##
## Residual Deviance: 1884.263
## AIC: 1914.263
```

```
round(summary(nesmod1)$coef,2)
```

```
##                     Value Std. Error t value
## ideomoderate         0.95       0.33    2.88
## ideoconservative     1.94       0.18   10.70
## age                 -0.01       0.00   -2.71
## genderfemale        -0.39       0.16   -2.51
## raceblack           -1.80       0.28   -6.48
## raceasian            0.12       0.54    0.23
## racenative american -0.14       0.37   -0.38
## racehispanic        -0.63       0.30   -2.12
## south1               0.21       0.18    1.20
## 1|2                 -1.40       0.31   -4.58
## 2|3                 -0.55       0.30   -1.85
## 3|4                  0.24       0.30    0.81
## 4|5                  0.63       0.30    2.09
## 5|6                  1.43       0.31    4.68
## 6|7                  2.41       0.32    7.62
```

2. Explain the results from the fitted model.

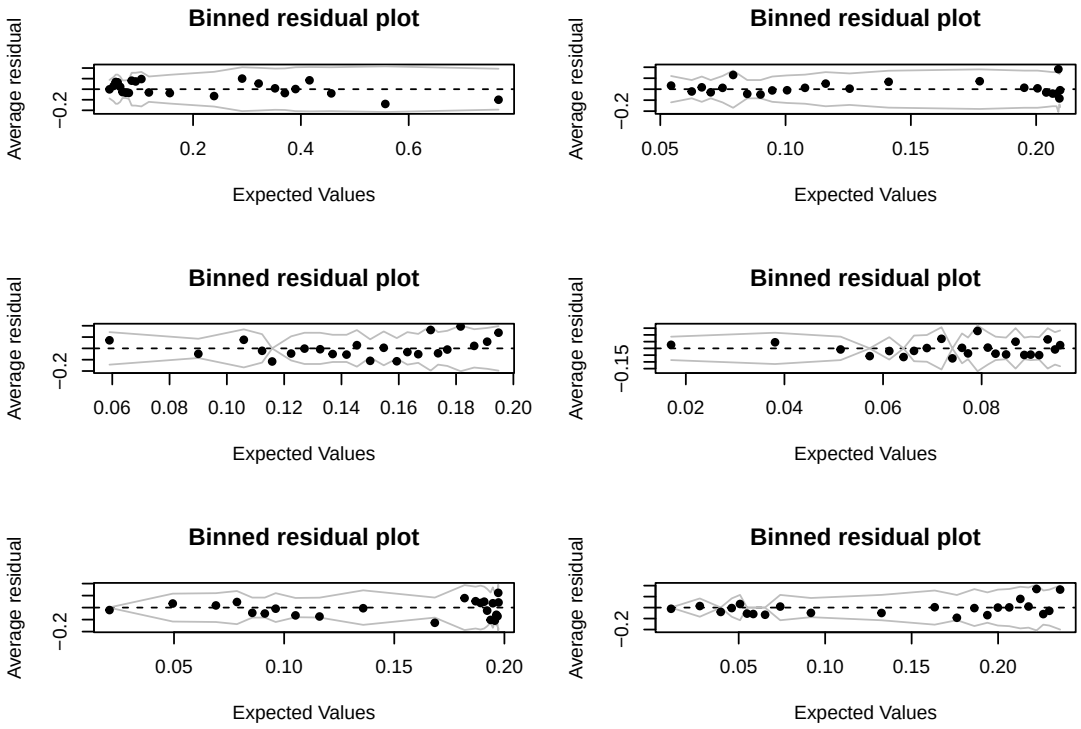
```
confint(nesmod1)
```

```
## Waiting for profiling to be done...
```

```
##                         2.5 %       97.5 %
## ideomoderate         0.3020516  1.601688963
## ideoconservative     1.5900104  2.302367108
## age                 -0.0231608 -0.003753405
## genderfemale        -0.6969767 -0.086312430
## raceblack           -2.3508599 -1.261785345
## raceasian           -0.9485309  1.210930077
## racenative american -0.8697439  0.580818771
## racehispanic        -1.2178816 -0.048884839
## south1              -0.1329627  0.554188748
```

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

```
nes <- cbind(partyid7 = nes_data_comp$partyid7, ideo = nes_data_comp$ideo, race = nes_data_comp$race, ag
nes <- data.frame(na.omit(nes))
resid <- model.matrix(~ factor(partyid7) - 1, data = nes) - fitted(nesmod1)
par(mfrow = c(3, 2))
for (i in 1:6) {
  binnedplot(fitted(nesmod1)[, i], resid[, i], cex.main = 1.3, main = "Binned residual plot")
}
```



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 hemaglutinin 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 chi-square test and an appropriate log-linear model, test the hypothesis that the distribution of responses is the same for the placebo and vaccine groups.

```
treat_chi <- chisq.test(treatment_table)
print(treat_chi)
```

```
##
##  Pearson's Chi-squared test
##
## data:  treatment_table
## X-squared = 17.648, df = 3, p-value = 0.0005199
```

2. For the model corresponding to the hypothesis of homogeneity 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.

```
treatment <- as.data.frame(treatment_table)
names(treatment) <- c("Treatment", "Response", "Count")

treatmod1<-glm(Count~Treatment+Response, family = poisson, data = treatment)
fit_values<-fitted(treatmod1)

pear_resid<-residuals(treatmod1, type = "pearson")
dev_resid<-residuals(treatmod1, type = "deviance")

X2<-sum(pear_resid^2)
D <-sum(dev_resid^2)


cat("Pearson chi-square (X^2):", X2, "\n")
```

```
## Pearson chi-square (X^2): 17.64783
```

```
cat("Deviance (D):", D, "\n")
```

```
## Deviance (D): 18.64253
```

3. Re-analyze these data using ordered logit model (use `polr`) to estimate the cut-points of a latent continuous response variable 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.

```
treatment2<- data.table(
  Treatment = factor(rep(c("placebo", "vaccine"), each = 3)),
  Response = factor(rep(c("small", "moderate", "large"), times = 2), ordered = TRUE),
  Count = c(25, 8, 5, 6, 18, 11)
)

treatment2_exp <- treatment2[rep(1:.N, Count)]
treatmod2 <- polr(Response ~ Treatment, data = treatment2, Hess = TRUE)
summary(treatmod2)
```

```
## Call:
## polr(formula = Response ~ Treatment, data = treatment2, Hess = TRUE)
##
## Coefficients:
##                               Value Std. Error            t value
## Treatmentvaccine 0.0000000000001804        1.5 0.0000000000001203
##
## Intercepts:
##                 Value   Std. Error t value
## large|moderate -0.6931  1.1456    -0.6050
## moderate|small  0.6931  1.1456     0.6050
##
## Residual Deviance: 13.18335
## AIC: 19.18335
```

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)
```

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

```
hsbmod1 <- multinom(prog ~race + read + write + math + science,data = hsb,trace = FALSE,HESS = TRUE)
summary(hsbmod1)
```

```
## Call:
## multinom(formula = prog ~ race + read + write + math + science,
##     data = hsb, trace = FALSE, HESS = TRUE)
##
## Coefficients:
##          (Intercept)  raceasian racehispanic   racewhite        read
## general     4.924957 1.11489221  -0.60687283 -0.01313942 -0.05388450
## vocation    8.777829 0.08636574   0.07298783  0.42373684 -0.05594167
##                write       math    science
## general  -0.03946933 -0.1071044 0.09229507
## vocation -0.06281609 -0.1253231 0.05262485
##
## Std. Errors:
##          (Intercept) raceasian racehispanic racewhite       read      write
## general     1.528744 0.9950814    0.8707214 0.6995466 0.02853999 0.02864533
## vocation    1.629837 1.3388885    0.7864713 0.6836971 0.03052243 0.02855810
##                math    science
## general  0.03391490 0.03053422
## vocation 0.03616922 0.03106921
##
## Residual Deviance: 332.6696
## AIC: 364.6696
```

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

```
predict(hsbmod1,newdata=hsb[hsb$id==99,],type="probs")
```

```
##  academic   general  vocation
## 0.3756043 0.4338602 0.1905356
```

Happiness

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

```r
data(happy)
```

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

```r
happymod1 <- polr(factor(happy) ~ money + sex + love + work,data = happy)
summary(happymod1)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = factor(happy) ~ money + sex + love + work, data = happy)
##
## Coefficients:
##          Value Std. Error t value
## money  0.02246    0.01066  2.1064
## sex   -0.47344    0.79498 -0.5955
## love   3.60764    0.80114  4.5031
## work   0.88751    0.40826  2.1739
##
## Intercepts:
##      Value   Std. Error t value
## 2|3   5.4708  1.9891     2.7504
## 3|4   6.4684  1.9223     3.3650
## 4|5   9.1591  2.1698     4.2212
## 5|6  10.9725  2.3213     4.7268
## 6|7  11.5113  2.3720     4.8530
## 7|8  13.5433  2.6673     5.0776
## 8|9  17.2909  3.1454     5.4971
## 9|10 19.0112  3.3270     5.7142
##
## Residual Deviance: 94.86029
## AIC: 118.8603
```

2. Interpret the parameters of your chosen model.

```r
confint(happymod1)
```

```
## Waiting for profiling to be done...

##
## Re-fitting to get Hessian

##              2.5 %     97.5 %
## money  0.002276811 0.04490097
## sex   -2.068912556 1.07918378
## love   2.168908580 5.37172934
## work   0.123787533 1.74622976
```

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

```
predict(happymod1,newdata = data.frame(love = 1,sex = 0,work = 1,money = 30),type = "probs")
```

```
##               2               3               4               5               6
## 0.5749087495352 0.2108349557988 0.1960959529833 0.0151526917408 0.0012506577394
##               7               8               9              10
## 0.0015263392585 0.0002252141575 0.0000044651793 0.0000009736073
```

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.

```
data(uncviet)

uncvietmod1<-vglm(policy~sex+year, family = cumulative(parallel = TRUE, link = "logitlink"), data = unc
```

```
## Warning in eval(slot(family, "initialize")): response should be ordinal---see
## ordered()
```

```
summary(uncvietmod1)
```

```
## Call:
## vglm(formula = policy ~ sex + year, family = cumulative(parallel = TRUE,
##     link = "logitlink"), data = uncviet)
##
## Coefficients:
##                              Estimate             Std. Error z value Pr(>|z|)
## (Intercept):1 -1.098612288668111144  0.730296743340221315  -1.504    0.132
## (Intercept):2 -0.000000000000017200  0.707106781186547572   0.000    1.000
## (Intercept):3  1.098612288668108058  0.730296743340221543   1.504    0.132
## sexMale        0.000000000000002512  0.565685424949238457   0.000    1.000
## yearGrad       0.000000000000001146  0.894427190999916854   0.000    1.000
## yearJunior     0.000000000000008896  0.894427190999915855   0.000    1.000
## yearSenior     0.000000000000008569  0.894427190999915966   0.000    1.000
## yearSoph       0.000000000000005958  0.894427190999916410   0.000    1.000
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3])
##
## Residual deviance: 110.9035 on 112 degrees of freedom
##
## Log-likelihood: -55.4518 on 112 degrees of freedom
##
## Number of Fisher scoring iterations: 1
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##    sexMale   yearGrad yearJunior yearSenior   yearSoph
##          1          1          1          1          1
```

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.

```
data(pneumo, package = "faraway")
```

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.

```
pneumomod1 <- multinom(status ~ year, data = pneumo, weights = Freq)
```

```
## # weights:  9 (4 variable)
## initial  value 407.585159
## iter  10 value 208.724810
## final  value 208.724782
## converged
```

```
summary(pneumomod1)
```

```
## Call:
## multinom(formula = status ~ year, data = pneumo, weights = Freq)
##
## 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
```

```
predict(pneumomod1, data.frame (year = 25), type = "probs")
```

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

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

```
pneumomod2 <- polr(factor(status) ~ year, data = pneumo, weights = Freq, Hess = TRUE)
summary(pneumomod2)
```

```
## Call:
## polr(formula = factor(status) ~ year, data = pneumo, weights = Freq,
```

```
##     Hess = TRUE)
##
## Coefficients:
##        Value Std. Error t value
## year 0.01566   0.009057    1.73
##
## Intercepts:
##               Value   Std. Error t value
## mild|normal   -1.8449  0.2492    -7.4039
## normal|severe  2.3676  0.2709     8.7411
##
## Residual Deviance: 502.1551
## AIC: 508.1551
```

```
predict(pneumomod2, data.frame (year = 25), type = "probs")
```

```
##       mild     normal     severe
## 0.09652357 0.78172799 0.12174844
```

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.

```
pneumo$disease <- ifelse(pneumo$status == "normal", 0, 1)
pneumo1 <- as.data.frame(cbind(Freq = pneumo$Freq, normal = ifelse(pneumo$status == "normal",1,0), mild

pneumomod3 <- multinom(cbind(normal,mild,severe) ~ year, weights = Freq, data = pneumo1)
```

```
## # weights:  9 (4 variable)
## initial  value 407.585159
## iter  10 value 208.809599
## final  value 208.724782
## converged
```

```
summary(pneumomod3)
```

```
## Call:
## multinom(formula = cbind(normal, mild, severe) ~ year, data = pneumo1,
##     weights = Freq)
##
## Coefficients:
##        (Intercept)       year
## mild     -4.291680 0.08356529
## severe   -5.059849 0.10928549
##
## Std. Errors:
##        (Intercept)       year
## mild     0.5214120 0.01528046
## severe   0.5964319 0.01646978
##
## Residual Deviance: 417.4496
## AIC: 425.4496
```

```r
predict(pneumomod3, data.frame(year = 25), type = "probs")
```

```
##     normal       mild     severe
## 0.82778727 0.09148803 0.08072470
```

4. Compare the three analyses.

```r
#The results from these three analyses are similar, and mild is between 0.08-0.10, normal is betwen 0.7
```