MA678 homework 05

Multinomial Regression

Xuan Zhu

October 21, 2018

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 7-point scale) using ideology and demographics with an ordered multinomial logit model.

1. Summarize the parameter estimates numerically and also graphically.

```
m1 <- polr(ordered(partyid7)~ideology+female+income+white,data=nes_data_comp,Hess=TRUE)
round(summary(m1)$coef,2)
```

```
##                                                         Value Std. Error
## ideology                                                 1.35       0.11
## female                                                  -0.24       0.17
## income2. 17 to 33 percentile                             0.42       0.34
## income3. 34 to 67 percentile                             0.37       0.32
## income4. 68 to 95 percentile                             0.38       0.33
## income5. 96 to 100 percentile                            1.41       0.45
## white                                                    0.76       0.21
## 1. strong democrat|2. weak democrat                     -0.86       0.35
## 2. weak democrat|3. independent-democrat                -0.01       0.34
## 3. independent-democrat|4. independent-independent       0.83       0.35
## 4. independent-independent|5. independent-republican     1.15       0.35
## 5. independent-republican|6. weak republican             2.13       0.36
## 6. weak republican|7. strong republican                  3.33       0.38
##                                                         t value
## ideology                                                  12.06
## female                                                    -1.36
## income2. 17 to 33 percentile                               1.25
## income3. 34 to 67 percentile                               1.14
## income4. 68 to 95 percentile                               1.16
## income5. 96 to 100 percentile                              3.14
## white                                                      3.66
## 1. strong democrat|2. weak democrat                       -2.48
## 2. weak democrat|3. independent-democrat                  -0.03
## 3. independent-democrat|4. independent-independent         2.39
## 4. independent-independent|5. independent-republican       3.30
## 5. independent-republican|6. weak republican               5.94
## 6. weak republican|7. strong republican                    8.79
```

Logit P($\hat{y}$>1.strong democrat)= 1.35ideo-0.24female+0.42$income_2$+0.37$income_3$+0.38$income_4$+1.41$income_5$+0.76white+0.86

Logit P($\hat{y}$>2.weak democrat)=1.35ideo-0.24female+0.42$income_2$+0.37$income_3$+0.38$income_4$+1.41$income_5$+0.76white+0.01

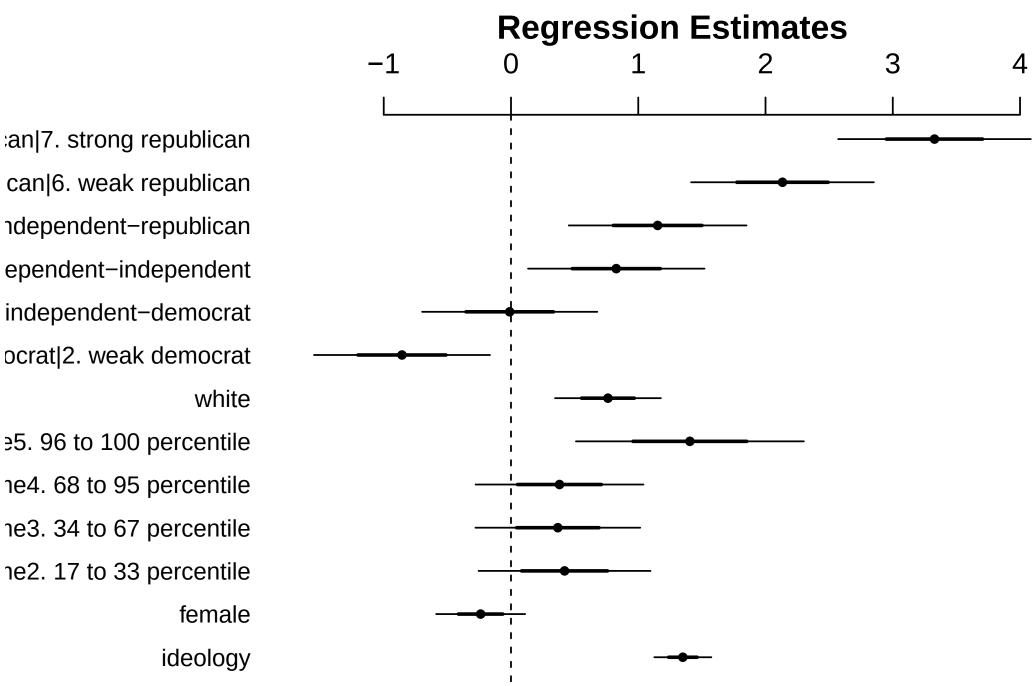
Logit P($\hat{y}$>3.independent-democrat)=1.35ideo-0.24female+0.42$income_2$+0.37$income_3$+0.38$income_4$+1.41$income_5$+0.76white-0.83

Logit P($\hat{y}$>4. independent-independent)=1.35ideo-0.24female+0.42$income_2$+0.37$income_3$+0.38$income_4$+1.41$income_5$+0.76wh
1.15

Logit P($\hat{y}$>5. independent-republican)=1.35ideo-0.24female+0.42$income_2$+0.37$income_3$+0.38$income_4$+1.41$income_5$+0.76whi
2.13

Logit P($\hat{y}$>6. weak republican)=1.35ideo-0.24female+0.42$income_2$+0.37$income_3$+0.38$income_4$+1.41$income_5$+0.76white-
3.33

```
coefplot(m1)
```



2. Explain the results from the fitted model.

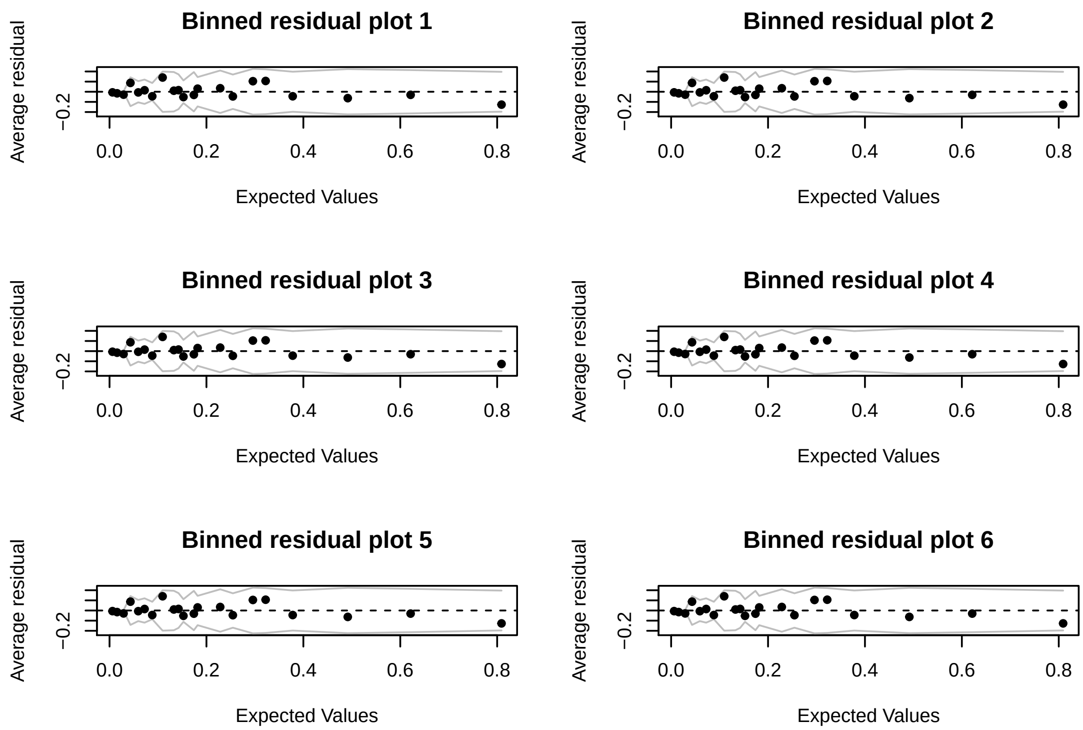
What are the probabilities for male, income 0-16 percentile, 0 ideo & non-white?

Solve:

$(\pi_2+\pi_3+\pi_4+\pi_5+\pi_6+\pi_7)/\pi_1) = exp(0.86)$ $(\pi_3+\pi_4+\pi_5+\pi_6+\pi_7)/(\pi_1+\pi_2) = exp(0.01)$ $(\pi_4+\pi_5+\pi_6+\pi_7)/(\pi_1+\pi_2+\pi_3) = exp(-0.83)$ $(\pi_5+\pi_6+\pi_7)/(\pi_1+\pi_2+\pi_3+\pi_4) = exp(-1.15)$ $(\pi_6+\pi_7)/(\pi_1+\pi_2+\pi_3+\pi_4+\pi_5) = exp(-2.13)$ $(\pi_7)/(\pi_1+\pi_2+\pi_3+\pi_4+\pi_5+\pi_6)) = exp(-3.33)$

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

```
nes <- cbind(partyid7=nes_data_comp$partyid7, female=nes_data_comp$female, income=nes_data_comp$income,
nes <- data.frame(na.omit(nes))
resid <- model.matrix(~factor(partyid7)-1, data=nes)-fitted(m1)
par(mfrow = c(3, 2))
p1 <- binnedplot(fitted(m1)[,1], resid[,1], cex.main=1.3, main="Binned residual plot 1")
p2 <- binnedplot(fitted(m1)[,1], resid[,1], cex.main=1.3, main="Binned residual plot 2")
p3 <- binnedplot(fitted(m1)[,1], resid[,1], cex.main=1.3, main="Binned residual plot 3")
p4 <- binnedplot(fitted(m1)[,1], resid[,1], cex.main=1.3, main="Binned residual plot 4")
p5 <- binnedplot(fitted(m1)[,1], resid[,1], cex.main=1.3, main="Binned residual plot 5")
p6 <- binnedplot(fitted(m1)[,1], resid[,1], cex.main=1.3, main="Binned residual plot 6")
```



The graph looks good.

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 ac ademic, 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).

```
library(VGAM)
try<-vglm(prog~gender+race+ses+schtyp+read+math+write+science+socst,family=multinomial,data=hsb)
anova(try) #model selection
```

```
## Analysis of Deviance Table (Type II tests)
##
## Model: 'multinomial', 'VGAMcategorical'
##
## Links: 'multilogit'
##
## Response: prog
##
##
##         Df Deviance  Pr(>Chi)
```

```
## gender   2   0.4151 0.8125559
## race     6   5.6816 0.4597805
## ses      4  12.1658 0.0161598 *
## schtyp   2   8.3799 0.0151471 *
## read     2   2.2600 0.3230355
## math     2  13.9886 0.0009171 ***
## write    2   1.3819 0.5010997
## science  2  10.6158 0.0049523 **
## socst    2   8.4230 0.0148244 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hsbm <-vglm(prog~ses+schtyp+math+science+socst,family=multinomial,data=hsb)
```

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

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
id99 <- hsb %>% filter(id=="99")%>%select(ses,schtyp,math,science,socst)
predict(newdata=id99,hsbm,type="response")
```

```
##    academic   general   vocation
## 1 0.6442696 0.2766495 0.07908086
```

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)
```

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

```r
happym <- polr(ordered(happy)~money+sex+love+work,data=happy,Hess=TRUE)
round(summary(happym)$coef,2)
```

```
##        Value Std. Error t value
## money  0.02       0.01    2.11
## sex   -0.47       0.79   -0.60
## love   3.61       0.80    4.50
## work   0.89       0.41    2.17
## 2|3    5.47       1.99    2.75
## 3|4    6.47       1.92    3.36
## 4|5    9.16       2.17    4.22
## 5|6   10.97       2.32    4.73
## 6|7   11.51       2.37    4.85
## 7|8   13.54       2.67    5.08
## 8|9   17.29       3.15    5.50
## 9|10  19.01       3.33    5.71
```

2. Interpret the parameters of your chosen model.

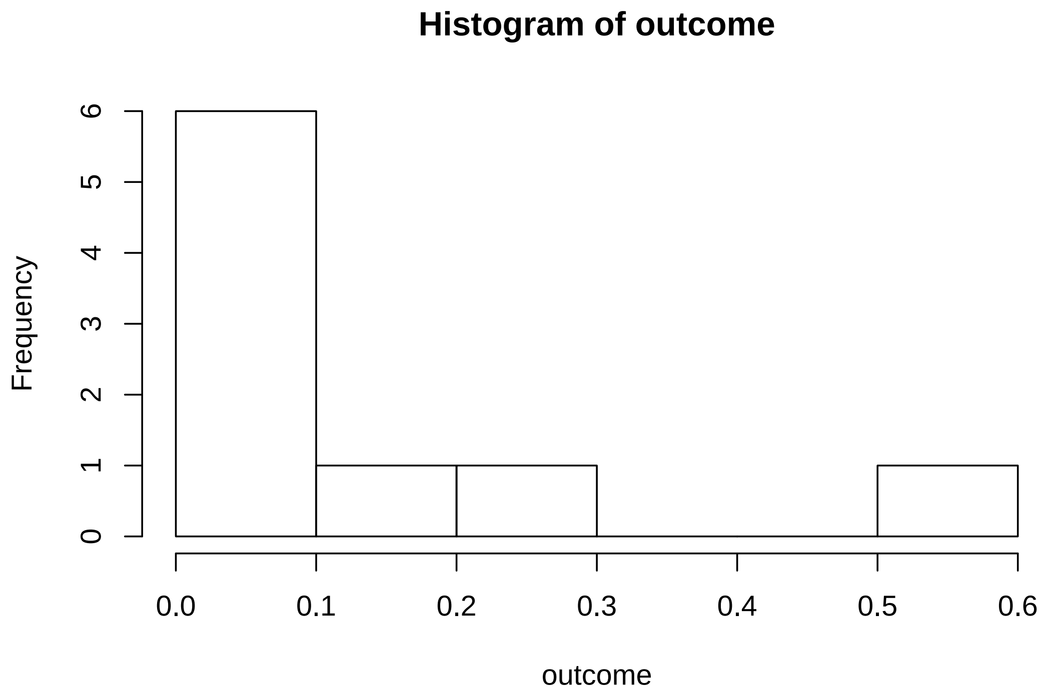
For example, the first logit function is

Logit P($\hat{y}$>level2)= 0.02money-0.47sex+3.61love+0.89work-5.47,and we can write 7 other functions in a similar way.

When interpreting the parameters, we can plug in the values of predictors as (money=a,sex=0/1,. . . .) to calculate the probabilities.

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
library(dplyr)
preda <- expand.grid(money=30,sex=0,love=1,work=1)
outcome<-predict(happym,newdata=preda,type="probs")
hist(outcome)
```



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)
surveym <- polr(ordered(policy)~sex+year,weights = y,data=uncviet,Hess=TRUE)
round(summary(surveym)$coef,2)
```

```
##             Value Std. Error t value
## sexMale     -0.65       0.08   -7.61
## yearGrad     1.18       0.10   11.51
## yearJunior   0.40       0.11    3.61
## yearSenior   0.54       0.11    4.84
## yearSoph     0.13       0.11    1.15
## A|B         -1.11       0.11  -10.02
## B|C         -0.01       0.11   -0.12
## C|D          2.44       0.12   20.45
```

Logit P($\hat{y}$>A)= -0.65sexMale+1.18yearGrad+0.4yearJunior+0.54yearSenior+0.13yearSoph+1.11

Logit P($\hat{y}$>B)=-0.65sexMale+1.18yearGrad+0.4yearJunior+0.54yearSenior+0.13yearSoph+0.01

Logit P($\hat{y}$>C)=-0.65sexMale+1.18yearGrad+0.4yearJunior+0.54yearSenior+0.13yearSoph-2.44

What if the student is a female freshman?

Solve:

$(\pi_B) + (\pi_C) + (\pi_D)/(\pi_A)$ = exp(1.11) $(\pi_C) + (\pi_D)/(\pi_A) + (\pi_B)$ = exp(0.01) $(\pi_D)/(\pi_A) + (\pi_B) + (\pi_C)$ = exp(-2.44)

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.

```
library(faraway)
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.

```
library(nnet)
nom <-multinom(status~year,family=multinomial,weights = Freq,data = pneumo)
```

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

```
nom
```

```
## Call:
## multinom(formula = status ~ year, data = pneumo, weights = Freq,
##     family = multinomial)
##
## Coefficients:
##        (Intercept)        year
## normal   4.2916723 -0.08356506
## severe  -0.7681706  0.02572027
##
## Residual Deviance: 417.4496
## AIC: 425.4496
```

```
predb <-expand.grid(year=25)
predict(newdata=predb,nom,type="probs")
```

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

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

```
ord <-polr(ordered(status)~year,weights = Freq,data=pneumo,Hess=TRUE)
ord
```

```
## Call:
## polr(formula = ordered(status) ~ year, data = pneumo, weights = Freq,
##     Hess = TRUE)
##
## Coefficients:
##       year
## 0.01566431
```

```
##
## Intercepts:
##   mild|normal normal|severe
##     -1.844855      2.367584
##
## Residual Deviance: 502.1551
## AIC: 508.1551
```

```
predict(newdata=predb,ord,type="probs")
```

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

3.Now treat the response variable as hier archical 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.

```
disease <- c(1:24)
logit1model <-cbind(pneumo,disease)
for (i in 1:24){
  if (logit1model[i,2]=="mild"){
    logit1model[i,4] <- 0
  } else {
    logit1model[i,4] <- 1
  }
}
logitm1 <- glm(disease~year,data=logit1model,family=binomial(link="logit"),weights = Freq)
predc <- expand.grid(year=25)
p1 <-predict(newdata=predc,logitm1,type="response")
logit2model <-cbind(logit1model,fitted(logitm1))
subset <- logit2model %>% filter(disease=="1")
severe <- c(1:16)
subset <- cbind(subset,severe)
for (i in 1:16){
  if (subset[i,2]=="normal"){
    subset[i,6] <- 0
  } else {
    subset[i,6] <- 1
  }}
logitm2 <- glm(severe~year,data=subset,family=binomial(link="logit"),weights = Freq)
p2 <- predict(newdata=predc,logitm2,type="response")
matrix(c(1-p1,p1*(1-p2),p1*p2),1,3)
```

```
##            [,1]      [,2]       [,3]
## [1,] 0.08532836 0.8316481 0.08302356
```

4. Compare the three analyses.

I think the second analysis is more accurate since the levels of mild, normal and severe can be seen as ordered. However, A1 & A3 are similar to each other.

(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 academy.awards.

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
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

name	description
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.