MA678 homework 05

Multinomial Regression

Jiahao Xu

October 22, 2018

Multinomial logit:

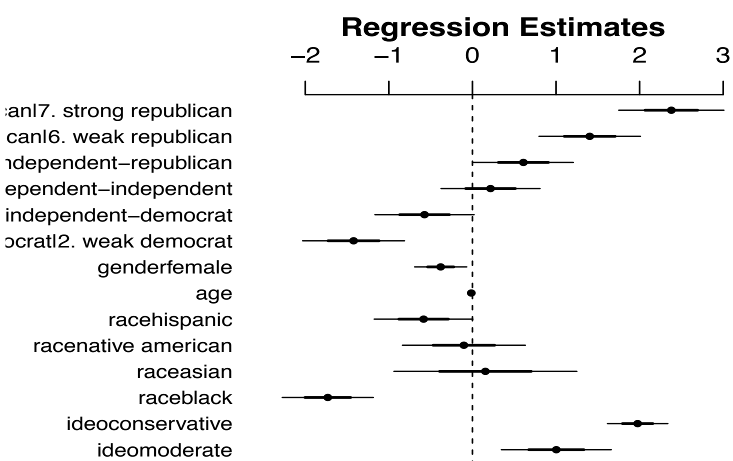
Using the individual-level survey data from the 2000 National Election Study (data in folder nes), predict party identification (which is on a 7-point scale) using ideology and demographics with an ordered multinomial logit model.

1. Summarize the parameter estimates numerically and also graphically.

```
require(VGAM)
mod1<-polr(partyid7~ideo+race+age+gender, data=nes_data_comp)
summary(mod1)
```

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

## Call:
## polr(formula = partyid7 ~ ideo + race + age + gender, data = nes_dat
a_comp)
##
## Coefficients:
##                        Value Std. Error t value
## ideomoderate         1.00334   0.328139  3.0577
## ideoconservative     1.97643   0.180090 10.9747
## raceblack           -1.73061   0.271850 -6.3660
## raceasian            0.15497   0.546457  0.2836
## racenative american -0.10237   0.367008 -0.2789
## racehispanic        -0.58450   0.294953 -1.9817
## age                 -0.01363   0.004924 -2.7678
## genderfemale        -0.37984   0.155229 -2.4469
##
## Intercepts:
##                                                         Value   Std. Er
ror
## 1. strong democrat|2. weak democrat                     -1.4220  0.3041

## 2. weak democrat|3. independent-democrat                -0.5734  0.2961

## 3. independent-democrat|4. independent-independent       0.2165  0.2956

## 4. independent-independent|5. independent-republican  0.6097  0.2980
```

```
## 5. independent-republican|6. weak republican         1.4033  0.3028

## 6. weak republican|7. strong republican              2.3797  0.3140

##                                                     t value
## 1. strong democrat|2. weak democrat                 -4.6756
## 2. weak democrat|3. independent-democrat            -1.9366
## 3. independent-democrat|4. independent-independent   0.7324
## 4. independent-independent|5. independent-republican  2.0455
## 5. independent-republican|6. weak republican         4.6342
## 6. weak republican|7. strong republican              7.5787
##
## Residual Deviance: 1892.422
## AIC: 1920.422
## (8 observations deleted due to missingness)

coefplot(mod1)

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



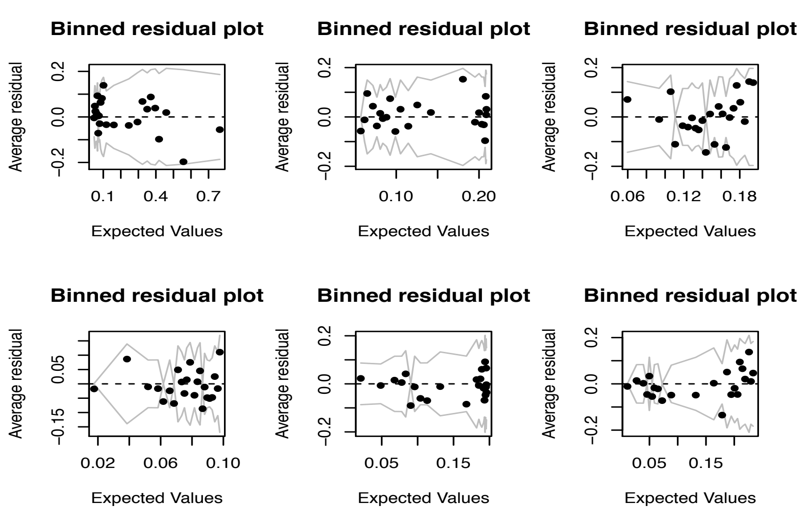
2. Explain the results from the fitted model.

```
confint(mod1)

## Waiting for profiling to be done...

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

```
##                               2.5 %       97.5 %
## ideomoderate          0.35781939  1.648405676
## ideoconservative      1.62688641  2.333204309
## raceblack            -2.27282514 -1.205143777
## raceasian            -0.92251513  1.244407549
## racenative american -0.82893228  0.616725087
## racehispanic         -1.16729016 -0.008107863
## age                  -0.02331781 -0.004001831
## genderfemale         -0.68466242 -0.075915381

# According to the confidence interval, most coefficients are significa
nt except raceasian and racenative american, which cross 0. And from th
e summary of the mod1, I realize that independent-democrat|independent-
independent and weak democrat|independent-democrat are not very signifi
cant. But, we should not ignore them.
```

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

```
data<-na.omit(data.frame(cbind(partyid7=nes_data_comp$partyid7,ideo=nes
_data_comp$ideo, race=nes_data_comp$race, age=nes_data_comp$age, gender
=nes_data_comp$gender)))
pr<-model.matrix(~factor(partyid7)-1, data=data)-fitted(mod1)
pred=fitted(mod1)
##pr <-data.frame(data[,colnames(pred)]/rowSums(data[,colnames(pred)])-
pred)
par(mfrow=c(2,3))
for(i in 1:6) binnedplot(pred[,i],pr[,i])
```



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).

```
library(nnet)
mod2<-multinom(prog~read+write+math+science+socst+schtyp+gender, data=h
sb,trace=FALSE)
summary(mod2)

## Call:
## multinom(formula = prog ~ read + write + math + science + socst +
##     schtyp + gender, data = hsb, trace = FALSE)
##
## Coefficients:
##          (Intercept)        read       write       math    science
## general     4.435877 -0.04393702 -0.02751721 -0.0991316 0.09637668
## vocation    7.820754 -0.03504340 -0.04124208 -0.1149067 0.06515018
##                socst schtyppublic gendermale
## general  -0.03247733    0.6513792 -0.2051378
## vocation -0.06684787    1.7281319 -0.3444629
##
## Std. Errors:
##          (Intercept)       read      write       math    science
## general     1.605532 0.02989198 0.03258346 0.03379351 0.03088801
## vocation    1.850841 0.03247272 0.03378273 0.03725463 0.03140748
##               socst schtyppublic gendermale
## general  0.02589797    0.5417632  0.4377327
## vocation 0.02735284    0.8082408  0.4801419
##
## Residual Deviance: 322.9444
## AIC: 354.9444
```

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

```
library(tidyverse)

## ─ Attaching packages ───────────────────────── tidyverse 1.2.1 ─

## ✓ tibble  1.4.2     ✓ purrr   0.2.5
## ✓ tidyr   0.8.1     ✓ dplyr   0.7.6
## ✓ readr   1.1.1     ✓ stringr 1.3.1
## ✓ tibble  1.4.2     ✓ forcats 0.3.0
```

```
## ─ Conflicts ──────────────────────────── tidyverse_conflicts() ─
## ✖ dplyr::between()   masks data.table::between()
## ✖ tidyr::expand()    masks Matrix::expand()
## ✖ tidyr::fill()      masks VGAM::fill()
## ✖ dplyr::filter()    masks stats::filter()
## ✖ dplyr::first()     masks data.table::first()
## ✖ dplyr::lag()       masks stats::lag()
## ✖ dplyr::last()      masks data.table::last()
## ✖ dplyr::recode()    masks car::recode()
## ✖ dplyr::select()    masks MASS::select()
## ✖ purrr::some()      masks car::some()
## ✖ purrr::transpose() masks data.table::transpose()

newdata<-filter(hsb,id=="99")
predict(mod2,newdata=newdata, type="probs")

##  academic   general  vocation
## 0.3662839 0.4596583 0.1740577
```

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.

```
library(nnet)
mod3<-multinom(happy~money+sex+love+work,data=happy,trace=FALSE)
summary(mod3)

## Call:
## multinom(formula = happy ~ money + sex + love + work, data = happy,
##     trace = FALSE)
##
## Coefficients:
##    (Intercept)    money        sex        love      work
## 3     95.34718 8.207436   47.53907 -121.974570 -83.37503
## 4    108.15356 6.448616  126.62067 -144.128459 -19.45492
## 5    103.43665 6.504835   17.26409  -89.127605 -18.72953
## 6    -56.16590 6.632862  -29.02349   -9.832305 -20.71266
## 7     23.22477 6.557004   16.88244  -51.573474 -17.95274
## 8    -94.78326 6.586027  -39.30831    6.352961 -17.99809
## 9   -213.80005 6.596051   16.52875  -14.640589  13.18524
## 10  -149.75016 4.278169 -142.97178   95.281252 -45.18974
##
## Std. Errors:
```

```
##    (Intercept)       money          sex        love        work
## 3  0.325485656 27.67909907 3.254857e-01 0.328049326 0.328049314
## 4  0.799797489  4.62411816 7.997975e-01 1.571938267 1.210306606
## 5  0.746595285  4.62366481 1.682088e+00 1.493190570 0.865609268
## 6  2.378685936  4.62378421 4.087284e+00 1.201823879 1.931873302
## 7  1.620528738  4.62359628 1.651471e+00 0.829190888 0.703236473
## 8  1.228910839  4.62360881 1.892117e+00 1.036603126 0.814853390
## 9  0.103510063  4.62362103 1.035101e-01 0.310530190 0.414040255
## 10 0.001547399  0.06183351 9.923873e-08 0.004642197 0.007414509
##
## Residual Deviance: 53.40729
## AIC: 133.4073
```

2. Interpret the parameters of your chosen model.

```
confint(mod3)

## , , 3
##
##                  2.5 %     97.5 %
## (Intercept)   94.70924   95.98512
## money        -46.04260   62.45747
## sex           46.90113   48.17701
## love        -122.61753 -121.33160
## work         -84.01799  -82.73207
##
## , , 4
##
##                   2.5 %     97.5 %
## (Intercept)  106.585985  109.72113
## money         -2.614489   15.51172
## sex          125.053096  128.18824
## love        -147.209401 -141.04752
## work         -21.827074  -17.08276
##
## , , 5
##
##                  2.5 %    97.5 %
## (Intercept) 101.973346 104.89995
## money        -2.557381  15.56705
## sex          13.967262  20.56093
## love        -92.054205 -86.20101
## work        -20.426092 -17.03297
##
## , , 6
##
##                  2.5 %     97.5 %
## (Intercept) -60.828037 -51.503760
## money        -2.429589  15.695312
## sex         -37.034425 -21.012564
## love        -12.187836  -7.476773
```

```
## work        -24.499065 -16.926260
##
## , , 7
##
##                  2.5 %    97.5 %
## (Intercept)  20.048588  26.40094
## money        -2.505078  15.61909
## sex          13.645613  20.11926
## love        -53.198659 -49.94829
## work        -19.331061 -16.57442
##
## , , 8
##
##                  2.5 %     97.5 %
## (Intercept) -97.191883 -92.374641
## money        -2.476079  15.648134
## sex         -43.016792 -35.599830
## love          4.321256   8.384666
## work        -19.595174 -16.401007
##
## , , 9
##
##                  2.5 %     97.5 %
## (Intercept) -214.00293 -213.59717
## money         -2.46608   15.65818
## sex           16.32588   16.73163
## love         -15.24922  -14.03196
## work          12.37374   13.99675
##
## , , 10
##
##                   2.5 %      97.5 %
## (Intercept) -149.753197 -149.747131
## money          4.156978    4.399361
## sex         -142.971776 -142.971776
## love          95.272153   95.290350
## work         -45.204268  -45.175204
```

```
## Form the confidence intervals, I realize that in different happiness
 scale, the significances of coefficients are different. But, generally
 money is the most significant predictor. To interpret the parameter, l
et's take a look on happiness scale>5. The interpretation is logit(happ
y>6)=6.63money-29.02sex-9.83love-20.71work+56.17
```

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(mod3,newdata=data.frame(love=1,sex=0,work=1,money=30),type="pro
b")
```

```
##             2            3            4            5            6 
## 1.476969e-83 2.134248e-24 1.301533e-23 1.000000e+00 8.507098e-35 
##            7            8            9           10 
## 3.097618e-18 5.706471e-44 4.207186e-91 4.243541e-71
```

newspaper survey on Vietnam War

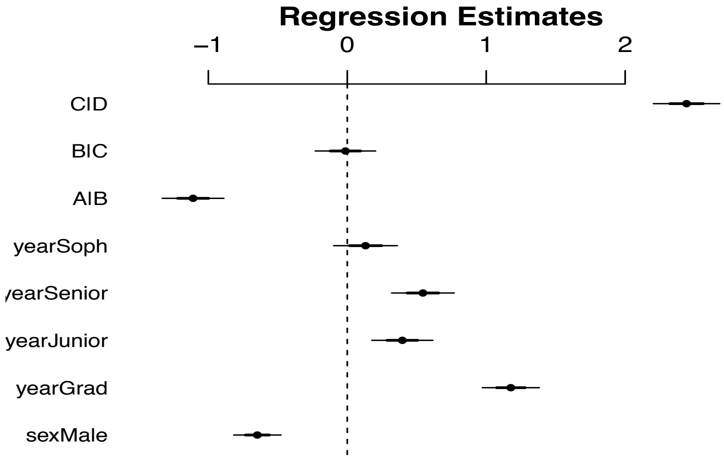
A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset `uncviet`. Treat the opinion as the response and the sex and year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

```r
data(uncviet)
?uncviet
mod4<-polr(policy~sex+year,data=uncviet,weights=y)
summary(mod4)
```

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

## Call:
## polr(formula = policy ~ sex + year, data = uncviet, weights = y)
## 
## Coefficients:
##               Value Std. Error t value
## sexMale     -0.6470    0.08499  -7.613
## yearGrad     1.1770    0.10226  11.510
## yearJunior   0.3964    0.10972   3.613
## yearSenior   0.5444    0.11248   4.840
## yearSoph     0.1315    0.11460   1.148
## 
## Intercepts:
##     Value    Std. Error t value 
## A|B  -1.1098   0.1107   -10.0210
## B|C  -0.0130   0.1086    -0.1202
## C|D   2.4417   0.1194    20.4455
## 
## Residual Deviance: 7757.056 
## AIC: 7773.056
```

```r
coefplot(mod4)
```

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



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")
?pneumo

## Help on topic 'pneumo' was found in the following packages:
##
##   Package                 Library
##   faraway                 /Library/Frameworks/R.framework/Versions/3.5
/Resources/library
##   VGAM                    /Library/Frameworks/R.framework/Versions/3.5
/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.

```
mod5<-multinom(formula=status~Freq+year,data=pneumo, family=multinomial,
 Hess=TRUE)

## # weights:  12 (6 variable)
## initial  value 26.366695
## iter  10 value 14.853565
## iter  20 value 14.131230
## iter  30 value 13.978859
```

```
## final  value 13.978664
## converged
```

```r
summary(mod5)
```

```
## Call:
## multinom(formula = status ~ Freq + year, data = pneumo, Hess = TRUE,

##     family = multinomial)
##
## Coefficients:
##        (Intercept)       Freq         year
## normal -39.5225538 0.87075608  0.687479973
## severe  -0.1093733 0.07164446 -0.008502992
##
## Std. Errors:
##        (Intercept)      Freq       year
## normal   41.194433 0.7308474 0.74063844
## severe    1.165409 0.1646184 0.03941746
##
## Residual Deviance: 27.95733
## AIC: 39.95733
```

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

```r
mod55<-polr(status~Freq+year,data=pneumo, Hess=TRUE)
summary(mod55)
```

```
## Call:
## polr(formula = status ~ Freq + year, data = pneumo, Hess = TRUE)
##
## Coefficients:
##          Value Std. Error t value
## Freq 0.0011645    0.01712 0.06800
## year 0.0008784    0.02897 0.03033
##
## Intercepts:
##               Value   Std. Error t value
## mild|normal   -0.6431  1.1625    -0.5532
## normal|severe  0.7435  1.1652     0.6381
##
## Residual Deviance: 52.72871
## AIC: 60.72871
```

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.

```r
pneumo$disease<-1
```

4. Compare the three analyses.

(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
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.