MA678 homework 05

Multinomial Regression

Tingrui Huang

Oct. 24, 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.

```
library(coefplot)
```

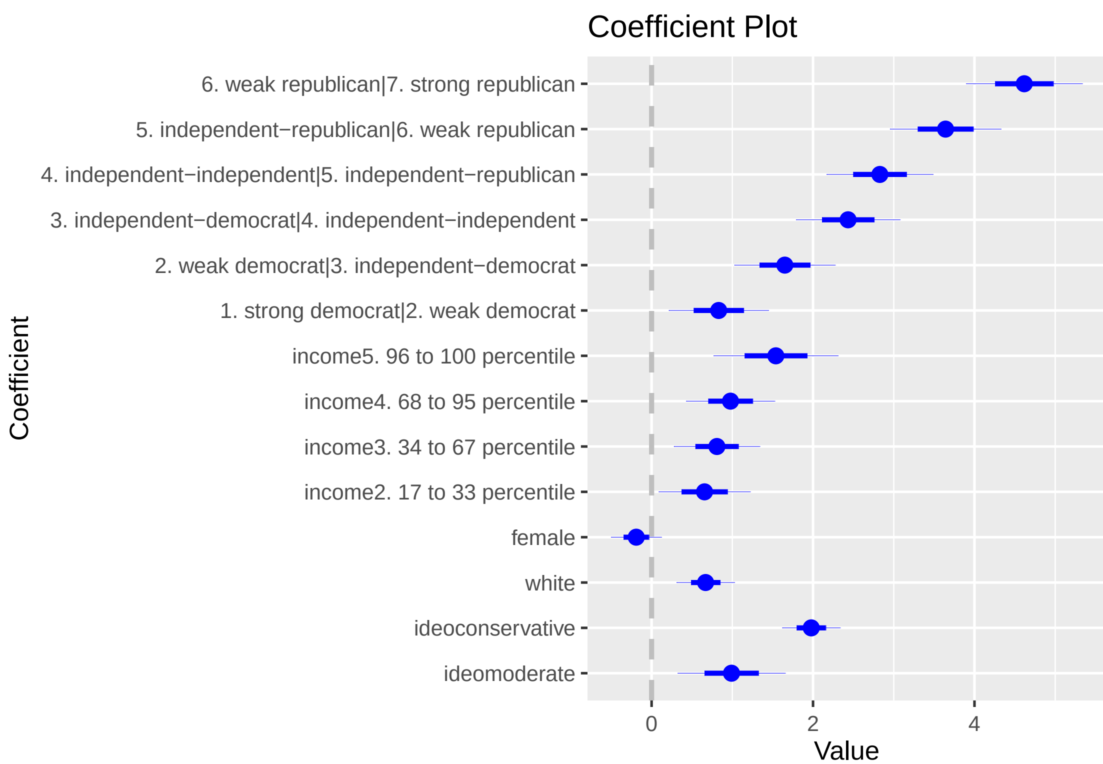
```
##
## Attaching package: 'coefplot'

## The following objects are masked from 'package:arm':
##
##     coefplot, coefplot.default, invlogit
```

```
catreg_pi <- vglm(partyid7~ideo+white+female+income, data = nes_data_comp, Hess = TRUE, family = multin
catreg_pi2 <- polr(ordered(partyid7)~ideo+white+female+income, data = nes_data_comp, Hess = TRUE)
summary(catreg_pi2)
```

```
## Call:
## polr(formula = ordered(partyid7) ~ ideo + white + female + income,
##     data = nes_data_comp, Hess = TRUE)
##
## Coefficients:
##                                Value Std. Error t value
## ideomoderate                  0.9898     0.3333   2.970
## ideoconservative              1.9778     0.1794  11.023
## white                         0.6689     0.1792   3.732
## female                       -0.1901     0.1560  -1.219
## income2. 17 to 33 percentile  0.6556     0.2834   2.313
## income3. 34 to 67 percentile  0.8090     0.2659   3.043
## income4. 68 to 95 percentile  0.9775     0.2744   3.562
## income5. 96 to 100 percentile 1.5393     0.3857   3.991
##
## Intercepts:
##                                                     Value   Std. Error
## 1. strong democrat|2. weak democrat                  0.8311  0.3087
## 2. weak democrat|3. independent-democrat             1.6504  0.3123
## 3. independent-democrat|4. independent-independent   2.4341  0.3222
## 4. independent-independent|5. independent-republican 2.8278  0.3293
## 5. independent-republican|6. weak republican         3.6415  0.3436
## 6. weak republican|7. strong republican              4.6166  0.3601
##                                                     t value
## 1. strong democrat|2. weak democrat                  2.6927
## 2. weak democrat|3. independent-democrat             5.2850
```

```
## 3. independent-democrat|4. independent-independent    7.5553
## 4. independent-independent|5. independent-republican  8.5869
## 5. independent-republican|6. weak republican         10.5983
## 6. weak republican|7. strong republican              12.8218
##
## Residual Deviance: 1936.238
## AIC: 1964.238
```

```r
# summarize parameter estimates graphically
coefplot(catreg_pi2)
```



2. Explain the results from the fitted model.

```r
catreg_pi2 <- polr(partyid7~ideo+white+female+income, data = nes_data_comp, Hess = TRUE)
summary(catreg_pi2)
```

```
## Call:
## polr(formula = partyid7 ~ ideo + white + female + income, data = nes_data_comp,
##     Hess = TRUE)
##
## Coefficients:
##                                 Value Std. Error t value
## ideomoderate                   0.9898     0.3333   2.970
## ideoconservative               1.9778     0.1794  11.023
## white                          0.6689     0.1792   3.732
## female                        -0.1901     0.1560  -1.219
## income2. 17 to 33 percentile   0.6556     0.2834   2.313
## income3. 34 to 67 percentile   0.8090     0.2659   3.043
## income4. 68 to 95 percentile   0.9775     0.2744   3.562
## income5. 96 to 100 percentile  1.5393     0.3857   3.991
```

```
##
## Intercepts:
##                                                    Value   Std. Error
## 1. strong democrat|2. weak democrat                 0.8311  0.3087
## 2. weak democrat|3. independent-democrat            1.6504  0.3123
## 3. independent-democrat|4. independent-independent  2.4341  0.3222
## 4. independent-independent|5. independent-republican 2.8278 0.3293
## 5. independent-republican|6. weak republican        3.6415  0.3436
## 6. weak republican|7. strong republican             4.6166  0.3601
##                                                    t value
## 1. strong democrat|2. weak democrat                 2.6927
## 2. weak democrat|3. independent-democrat            5.2850
## 3. independent-democrat|4. independent-independent  7.5553
## 4. independent-independent|5. independent-republican 8.5869
## 5. independent-republican|6. weak republican       10.5983
## 6. weak republican|7. strong republican            12.8218
##
## Residual Deviance: 1936.238
## AIC: 1964.238
```

```r
confint(catreg_pi2)
```

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

```
##                                     2.5 %    97.5 %
## ideomoderate                    0.3339910 1.6447802
## ideoconservative                1.6294238 2.3331253
## white                           0.3188670 1.0219168
## female                         -0.4960088 0.1156843
## income2. 17 to 33 percentile    0.1025110 1.2148676
## income3. 34 to 67 percentile    0.2907121 1.3342866
## income4. 68 to 95 percentile    0.4426589 1.5197427
## income5. 96 to 100 percentile   0.7872815 2.3019601
```

From the above result we can see the "female" predictor is not statistically significant. Other variables such as "ideo", "white" and "income" are statistically significant. "ideo": the liberal is set as a baseline. Comparing to liberal, people have moderate and conservative ideology tend to have independent or republican party identification. For people with moderate ideo, the log odds of supporting republican will be increase by 0.989 and people with conservative ideo, the log odds will increase 1.977. "white": the coefficient is positive means white people tend to be more republican-friendly comparing with other race. "income": income level of "0-16" is set as baseline. Generally speaking, the more a people make the more this people will be friendly to republican.

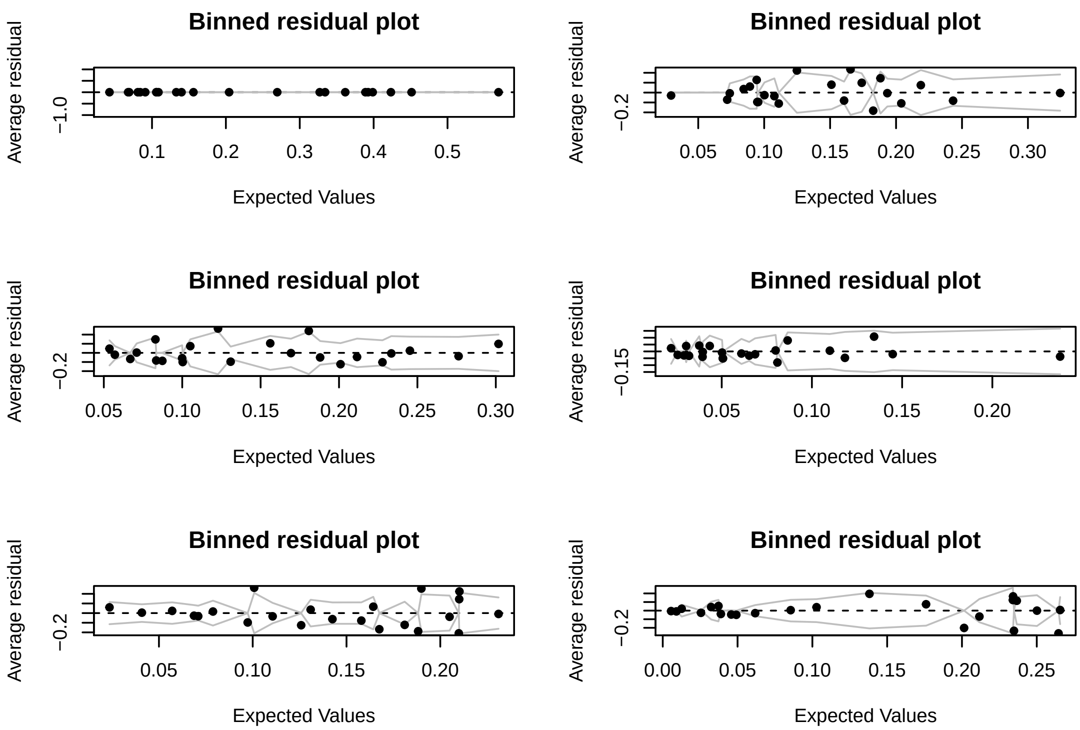
3. Use a binned residual plot to assess the fit of the model.

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------

## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.7.7
## v readr   1.1.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0

## -- Conflicts ---------------------------------------------------------------------------------
## x dplyr::between()   masks data.table::between()
## x tidyr::expand()    masks Matrix::expand()
## x tidyr::fill()      masks VGAM::fill()
```

```
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x dplyr::recode()    masks car::recode()
## x dplyr::select()    masks MASS::select()
## x purrr::some()      masks car::some()
## x purrr::transpose() masks data.table::transpose()
```

```
nes_resid <- nes_data_comp %>% select(partyid7,ideo,white,female,income) %>% na.omit() %>% as.data.fram
nes_resid_m <- model.matrix(~factor(partyid7),data=nes_resid)-fitted(catreg_pi)
nes_resid_m[,1] <- (nes_resid$partyid7==1)*1

par(mfrow=c(3,2))
for (i in 1:6) {
  binnedplot(fitted(catreg_pi)[,i],nes_resid_m[,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 programa`academic, vocational, or general`athat the students pursue in high school. The response is multinomial with three levels.

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

```
## starting httpd help server ... done
```

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

```r
catreg_hs <- polr(prog~gender+race+ses+schtyp+read+write+math+science+socst,data = hsb, Hess = TRUE)
summary(catreg_hs)
```

```
## Call:
## polr(formula = prog ~ gender + race + ses + schtyp + read + write +
##     math + science + socst, data = hsb, Hess = TRUE)
##
## Coefficients:
##                  Value Std. Error  t value
## gendermale    -0.22195    0.34843 -0.63700
## raceasian     -0.18565    0.84315 -0.22019
## racehispanic -0.03453    0.63825 -0.05411
## racewhite      0.20342    0.53649  0.37916
## seslow         0.25620    0.46229  0.55420
## sesmiddle      0.88021    0.39523  2.22705
## schtyppublic   1.21460    0.48390  2.51001
## read          -0.03077    0.02389 -1.28810
## write         -0.02559    0.02548 -1.00436
## math          -0.09354    0.02623 -3.56584
## science        0.04922    0.02385  2.06382
## socst         -0.05027    0.02053 -2.44828
##
## Intercepts:
##                   Value   Std. Error t value
## academic|general -6.2633  1.3957    -4.4876
## general|vocation -4.8251  1.3613    -3.5444
##
## Residual Deviance: 321.3928
## AIC: 349.3928
```

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

```r
predict(catreg_hs,hsb[hsb$id==99,],type="probs")
```

```
##  academic   general  vocation
## 0.5818527 0.2724298 0.1457174
```

Happiness

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

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

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

```r
library(nnet)
catreg_hp <- polr(factor(happy)~money+sex+love+work, data = happy,Hess = TRUE)
summary(catreg_hp)
```

```
## Call:
## polr(formula = factor(happy) ~ money + sex + love + work, data = happy,
##     Hess = TRUE)
##
## 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(catreg_hp)
```

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

##              2.5 %     97.5 %
## money  0.002276811 0.04490097
## sex   -2.068912555 1.07918378
## love   2.168908594 5.37172930
## work   0.123787533 1.74622976
```

Among the four predictors, the sex is not a statistically significant predictor. "money": money has a positive coefficient which means the more money a person make the higher happy score this person will get. For every additional thousand dollars a person make, the log odds of getting higher happy score will increase 0.02. "love" and "work": both of these predictors have positive coefficient so that for every unit increase in love and work, the log odds of getting a higher happy score will increase by their coefficients.

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

```
##            2            3            4            5            6
## 5.749090e-01 2.108352e-01 1.960955e-01 1.515274e-02 1.250661e-03
##            7            8            9           10
## 1.526345e-03 2.252149e-04 4.465201e-06 9.736115e-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.

```
data(uncviet)
?uncviet
catreg_viet <- polr(policy~sex+year, weights = y,data = uncviet, Hess = TRUE)
summary(catreg_viet)
```

```
## Call:
## polr(formula = policy ~ sex + year, data = uncviet, weights = y,
##     Hess = TRUE)
##
## 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
```

"sexMale": comparing to female, male students tend to be more aggressive on the Vietnam war. The log odds of getting higher response level is less for male comparing with female by 0.64. "year": take fresh as a baseline, all other levels has relatively mild opinion on Viewnam war. In general, students at higher level tend to have higher log odds of getting a mild opinion comparing to student at lower level.

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
View(pneumo)
```

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

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

```r
predict(catreg_pne,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.

```r
catreg_pne2 <- polr(status~year, weights = Freq, data = pneumo, Hess = TRUE)
summary(catreg_pne2)
```

```
## Call:
## polr(formula = 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
```

```r
predict(catreg_pne2,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.

```r
pneumo2 <- pneumo %>% mutate(disease=ifelse(pneumo$status=="normal",0,1)) # normal = 0, mild & severe =
ifdisease <- glm(disease~year, data = pneumo2,weights=Freq)
summary(ifdisease)
```

```
## Warning in summary.glm(ifdisease): observations with zero weight not used
## for calculating dispersion

##
## Call:
## glm(formula = disease ~ year, data = pneumo2, weights = Freq)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1159  -0.8509   0.8882   1.5095   2.0451
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.099464   0.159865  -0.622   0.5409
## year         0.013687   0.005855   2.338   0.0299 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.508414)
##
##     Null deviance: 63.876  on 21  degrees of freedom
## Residual deviance: 50.168  on 20  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 2
```

```r
pneumo3 <- pneumo2 %>% filter(disease==1) %>% mutate(level=ifelse(status=="mild",0,1))
levelofs <- glm(level~year,data = pneumo3,weights=Freq)
ifd_pred <- predict(ifdisease, newdata=data.frame(year=25),type="response")
no_disease <- 1-ifd_pred
level_pred <- predict(levelofs,data.frame(year=25),type="response")
level_mild <- (1 - level_pred)*ifd_pred
level_severe <- level_pred*ifd_pred
cbind(no_disease,level_mild,level_severe)
```

```
##   no_disease level_mild level_severe
## 1  0.7572788  0.1348025    0.1079187
```

4. Compare the three analyses.

```r
summary(catreg_pne)
```

```
## 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
```

```r
summary(catreg_pne2)
```

```
## Call:
## polr(formula = 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
```

```r
summary(ifdisease,ifd_pred)
```

```
##
## Call:
## glm(formula = disease ~ year, data = pneumo2, weights = Freq)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1159  -0.8509   0.8882   1.5095   2.0451
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.099464   0.049729  -2.000   0.0455 *
## year         0.013687   0.001821   7.515 5.69e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2427212)
##
##     Null deviance: 63.876  on 21  degrees of freedom
## Residual deviance: 50.168  on 20  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 2
```

The first model has smaller residual deviance and AIC than the second model.

(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

name	description
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.