MA678 homework 05

Multinomial Regression

Sky Liu

Oct. 27, 2017

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.

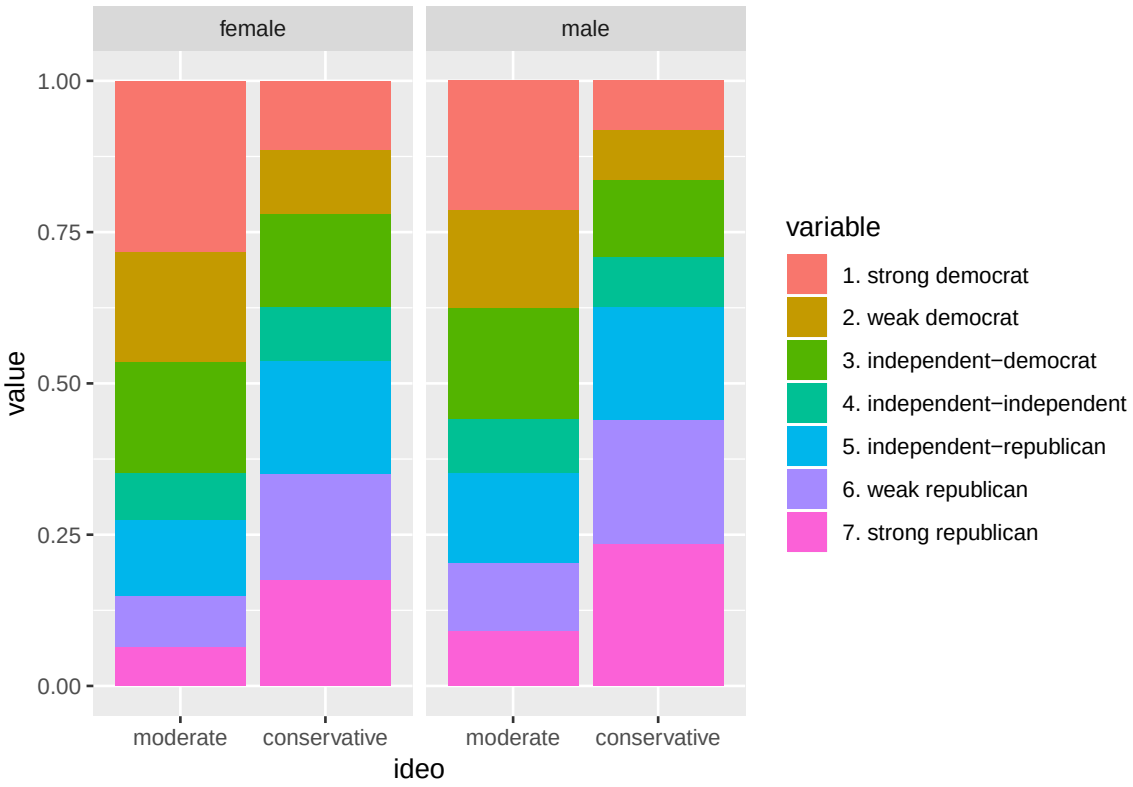
1. Summarize the parameter estimates numerically and also graphically.

```
lm1<- polr(partyid7~factor(ideo)+factor(gender),data=nes_data_comp)
summary(lm1)
```

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

```
## Call:
## polr(formula = partyid7 ~ factor(ideo) + factor(gender), data = nes_data_comp)
##
## Coefficients:
##                            Value Std. Error t value
## factor(ideo)moderate      0.7859     0.3279   2.397
## factor(ideo)conservative  1.9100     0.1747  10.932
## factor(gender)female     -0.3703     0.1526  -2.426
##
## Intercepts:
##                                                    Value   Std. Error
## 1. strong democrat|2. weak democrat               -0.5176  0.1628
## 2. weak democrat|3. independent-democrat           0.2747  0.1600
## 3. independent-democrat|4. independent-independent 1.0234  0.1667
## 4. independent-independent|5. independent-republican 1.3914  0.1721
## 5. independent-republican|6. weak republican       2.1535  0.1843
## 6. weak republican|7. strong republican            3.0906  0.2047
##                                                    t value
## 1. strong democrat|2. weak democrat               -3.1792
## 2. weak democrat|3. independent-democrat           1.7174
## 3. independent-democrat|4. independent-independent 6.1390
## 4. independent-independent|5. independent-republican 8.0826
## 5. independent-republican|6. weak republican      11.6852
## 6. weak republican|7. strong republican           15.0955
##
## Residual Deviance: 1975.189
## AIC: 1993.189
```

```
predx<- expand.grid(ideo = c("moderate","conservative"),gender=c("female","male"))
predy<-predict(lm1,newdata=predx,type = "p")
ggplot(melt(cbind(predx,predy),id.vars = c("gender","ideo")))+
  geom_bar(stat="identity")+aes(x=ideo,y=value, fill=variable)+
  facet_grid(~gender)
```



2. Explain the results from the fitted model.

From the coefficient we can see that gender has no statistical significant impact on party id.

log odds of not strong democrat $= log(\frac{\pi 2+\pi 3+\pi 4+\pi 5+\pi 6+\pi 7}{\pi 1})$

$= \beta_1 moderate + \beta_2 conservative + \beta_3 female - c12$

= 0.5176 + 0.7859 * moderate+ 1.9100 * conservative - 0.3703 * female

log odds of not strong democrat nor weak democrat $= log(\frac{\pi 3+\pi 4+\pi 5+\pi 6+\pi 7}{\pi 1+\pi 2})$

$= \beta_1 moderate + \beta_2 conservative + \beta_3 female - c23$

= -0.2747 + 0.7859 * moderate+ 1.9100 * conservative - 0.3703 * female

log odds of not strong democrat, weak democrat nor independent-democrat $= log(\frac{\pi 4+\pi 5+\pi 6+\pi 7}{\pi 1+\pi 2+\pi 3})$

$= \beta_1 moderate + \beta_2 conservative + \beta_3 female - c34$

= -1.0234 + 0.7859 * moderate+ 1.9100 * conservative - 0.3703 * female

log odds of not strong democrat, weak democrat, independent-democrat nor independent-independent $= log(\frac{\pi 5+\pi 6+\pi 7}{\pi 1+\pi 2+\pi 3+\pi 4})$

$= \beta_1 moderate + \beta_2 conservative + \beta_3 female - c45$

= -1.3914 + 0.7859 * moderate+ 1.9100 * conservative - 0.3703 * female

log odds of weak republican or strong republican $= log(\frac{\pi 6+\pi 7}{\pi 1+\pi 2+\pi 3+\pi 4+\pi 5})$

$= \beta_1 moderate + \beta_2 conservative + \beta_3 female - c56$

= -2.1535 + 0.7859 * moderate+ 1.9100 * conservative - 0.3703 * female

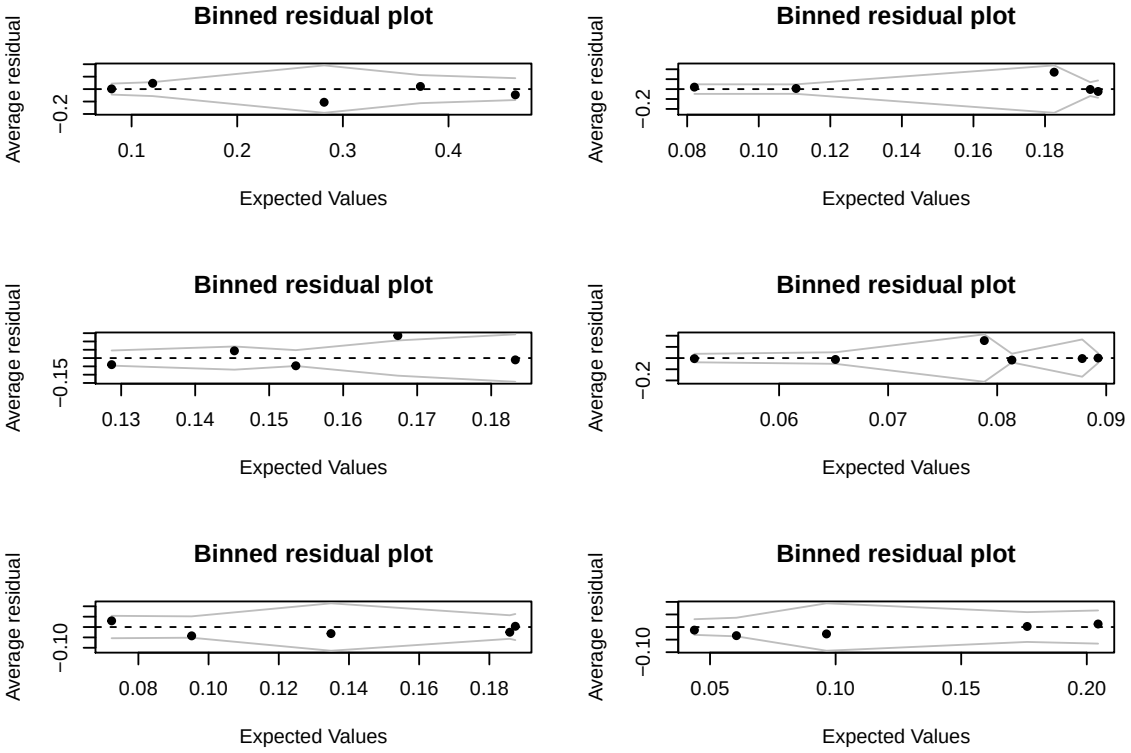
log odds of strong republican $= log(\frac{\pi 7}{\pi 1+\pi 2+\pi 3+\pi 4+\pi 5 \pi 6+})$

$= \beta_1 moderate + \beta_2 conservative + \beta_3 female - c67$

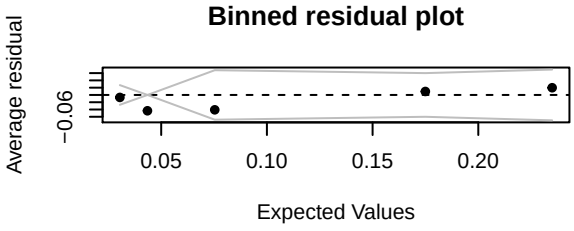
= -3.0906 + 0.7859 * moderate+ 1.9100 * conservative - 0.3703 * female

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

```
nesdata <- cbind(partyid=nes_data_comp$partyid7, female=nes_data_comp$female, ideo=nes_data_comp$ideo)
nesdata <- data.frame(na.omit(nesdata))
resid <- model.matrix(~factor(partyid)-1, data=nesdata)-fitted(lm1)
par(mfrow=c(3,2))
for(i in 1:6){
  binnedplot(fitted(lm1)[,i], resid[,i])
}
```




```
binnedplot(fitted(lm1)[,7], resid[,7])
```



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).

```
lm2 <- multinom(prog~read+write+math+science+race,data = hsb,trace=FALSE,HESS=TRUE)
summary(lm2)
```

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

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

```
predict(lm2,newdata=hsb[hsb$id==99,],type="p")
```

```
##  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.

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

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

```
lm3 <-polr(factor(happy)~money+factor(sex)+factor(love)+factor(work),data=happy)
summary(lm3)
```

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

```
## Call:
```

```
## polr(formula = factor(happy) ~ money + factor(sex) + factor(love) +
##     factor(work), data = happy)
##
## Coefficients:
##                   Value Std. Error  t value
## money           0.01783    0.01087  1.64024
## factor(sex)1   -1.02504    0.93629 -1.09479
## factor(love)2   3.45757    1.56121  2.21467
## factor(love)3   7.85036    1.85200  4.23885
## factor(work)2 -1.18912    1.68765 -0.70460
## factor(work)3   0.01574    1.58056  0.00996
## factor(work)4   1.84630    1.53696  1.20127
## factor(work)5   0.64794    2.14983  0.30139
##
## Intercepts:
##       Value   Std. Error t value
## 2|3   -0.8390  1.8387    -0.4563
## 3|4    0.0100  1.7713     0.0056
## 4|5    2.4280  2.0149     1.2050
## 5|6    4.4745  2.1063     2.1243
## 6|7    5.0675  2.1243     2.3856
## 7|8    7.3973  2.2303     3.3168
## 8|9   11.3105  2.5925     4.3628
## 9|10 13.0849  2.7916     4.6872
##
## Residual Deviance: 90.47841
## AIC: 122.4784
```

2. Interpret the parameters of your chosen model. $log\frac{\pi 3+...+\pi 10}{\pi 1+\pi 2} = 0.84+0.0178???money???1.025???sex_1+3.46???love_2+7.85???love_3???1.19???work_2+0.02???work_3+1.85???work_4+0.65???work_5$ Log odds of lonely people who are unsatisfactory with sex, with 0 family income, that has happy index from 3 to 10 over the ones with happy index = 2, is 0.84

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

```
kable(
predict(lm3,newdata=data.frame(love=1,sex=0,work=1,money=30),type="probs"))
```

	x
2	0.2020073
3	0.1697176
4	0.4973909
5	0.1118011
6	0.0084460
7	0.0095918
8	0.0010243
9	0.0000174
10	0.0000035

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
lm4 <- polr(policy~sex+year,data=uncviet,weights = y,Hess = TRUE)
summary(lm4)
```

```
## 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
```

The probability of policy is not A is $exp(1.11???0.65???sexmale + 1.18???yearGrad + 0.40???yearJunior + 0.54???yearSenior + 0.13???yearSoph)$

A male has opinions B,C or D is 48% (1- exp(-0.65)) lower than a female, holding other variable constant.

A grad student has opinions B,C or D is 224% (exp(1.177)-1) higher than a freshman, holding other variable constant.

A junior student has opinions B,C or D is 49% (exp(0.396)-1) higher than a freshman, holding other variable constant.

A senior student has opinions B,C or D is 72% (exp(0.5444)-1) higher than a freshman, holding other variable constant.

A sophemore student has opinions B,C or D is 14% (exp(1.1315)-1) higher than a freshman, holding other variable constant.

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")
head(pneumo)
```

```
##   Freq status year
## 1   98 normal  5.8
## 2   51 normal 15.0
## 3   34 normal 21.5
## 4   35 normal 27.5
## 5   32 normal 33.5
## 6   23 normal 39.5
```

```
?pneumo
```

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

```
lm6_nominal<-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
```

```
summary(lm6_nominal)
```

```
## 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
```

```
pred1<-predict(lm6_nominal,newdata=data.frame(year=25),type = "probs")
```

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

```
lm6_ornidal<-polr(status~year,weights=Freq,data=pneumo)
summary(lm6_ornidal)
```

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

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

```
pred2<-predict(lm6_ornidal,newdata=data.frame(year=25),type = "probs")
```

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.

```
n_pneumo <- pneumo
n_pneumo$status <- as.character(n_pneumo$status)
n_pneumo$status[9:24] <- "abnormal"
n_pneumo$status <- as.factor(n_pneumo$status)
abn_pneumo <- pneumo[-1:-8, ]
lm6_normal <- glm( status ~ year, data = n_pneumo, family = binomial(link = "logit"), weights = Freq)
lm6_abnormal <- glm( status ~ year, data = abn_pneumo, family = binomial(link = "logit"), weights = Fre

normal<-predict (lm6_normal, newdata=data.frame(year=25), type = "response")
severe<-predict (lm6_abnormal, newdata=data.frame(year=25), type = "response") *(1-predict (lm6_normal,
mild <- (1-predict (lm6_abnormal, newdata=data.frame(year=25), type = "response")) *(1-predict (lm6_nor
pred3<-c(mild, normal, severe)
```

4. Compare the three analyses.

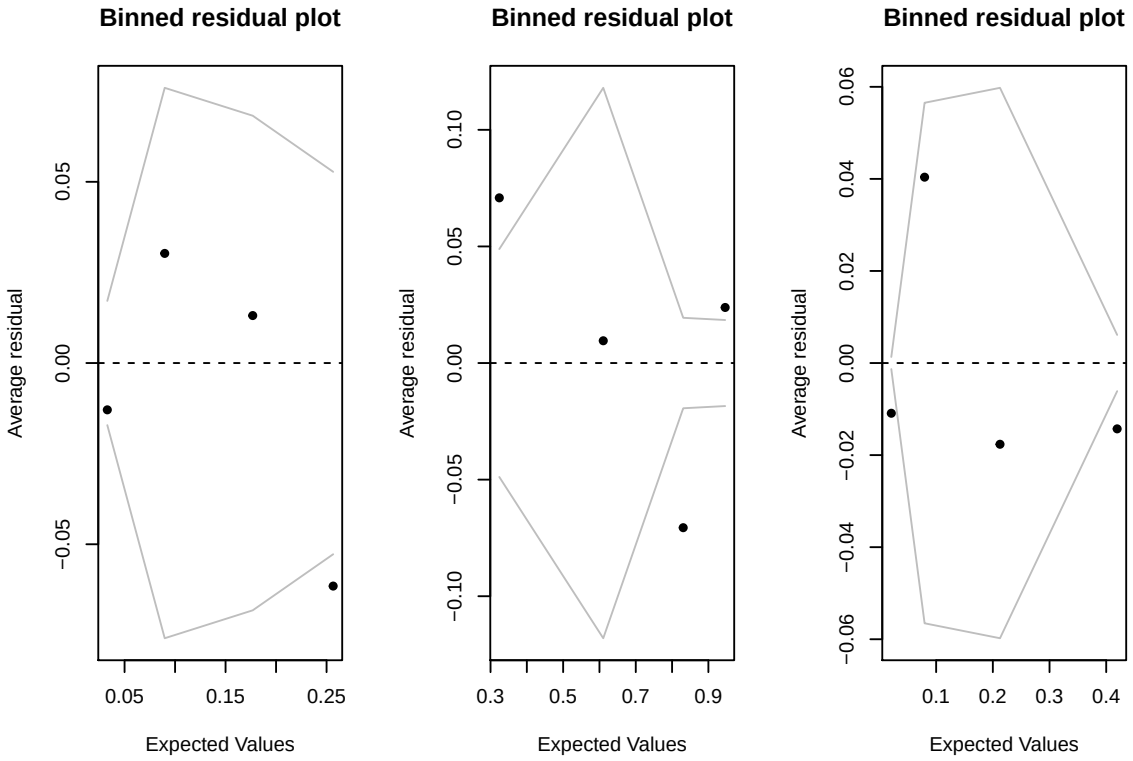
```
kable(rbind(pred1,pred2,pred3))
```

	mild	normal	severe
pred1	0.0914882	0.827787	0.0807248
pred2	0.0965236	0.781728	0.1217484
pred3	0.0966500	0.826299	0.0770510

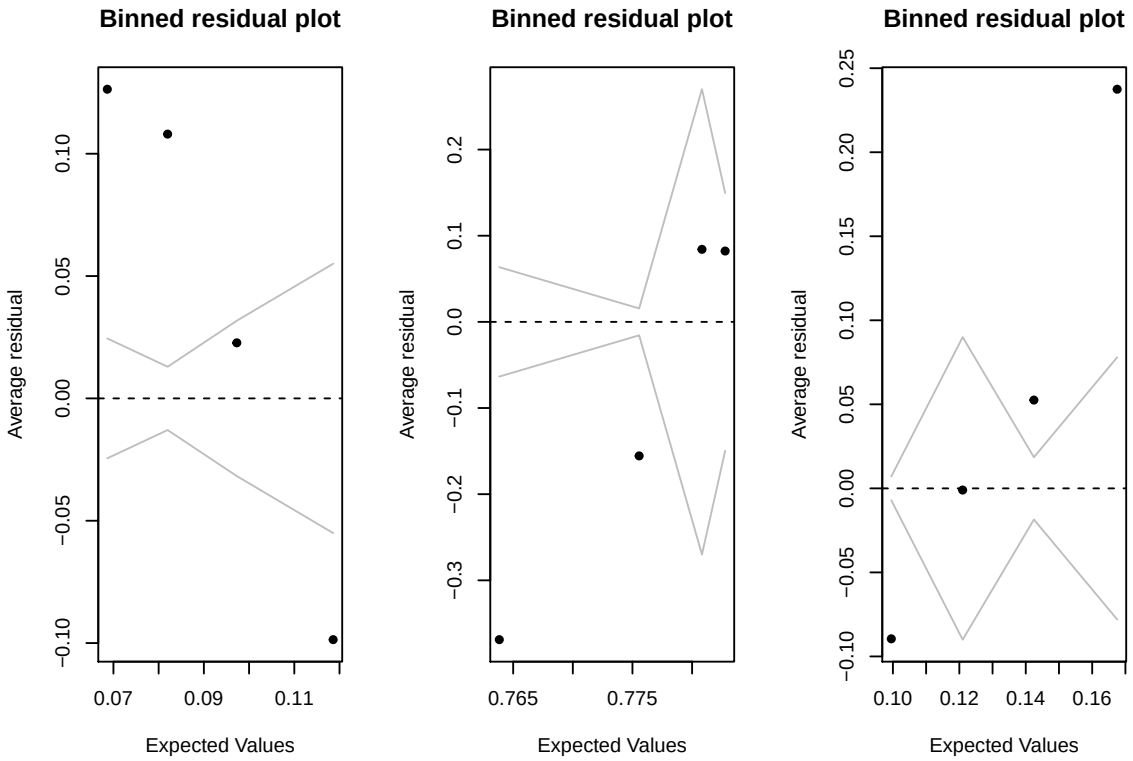
```
new_pneumo<-dcast(pneumo, year ~ status, value.var = "Freq")
new_pneumo<-new_pneumo %>%mutate(total=apply(new_pneumo[,2:4],1,sum))
new_pneumo[,2:4]<-round(new_pneumo[,2:4]/new_pneumo[,"total"],2)

pred11<-predict(lm6_nominal,newdata=new_pneumo,type="p")
resid1<-new_pneumo[,2:4]-pred11
par(mfrow=c(1,3))
```

```
for(i in 1:3){
  binnedplot(pred11[,i],resid1[,i])
  }
```




```
pred22<-predict(lm6_ornidal,newdata=new_pneumo,type="p")
resid2<-new_pneumo[,2:4]-pred22
par(mfrow=c(1,3));for(i in 1:3){binnedplot(pred22[,i],resid2[,i])}
```



From the binnedplot we can see that the second model does not fit well.

From the probability prediction table we can see that the first and the second model are similar.

(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

name	description
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.