# MA678 homework 05

## Multinomial Regression

Jinfei Xue

October 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.

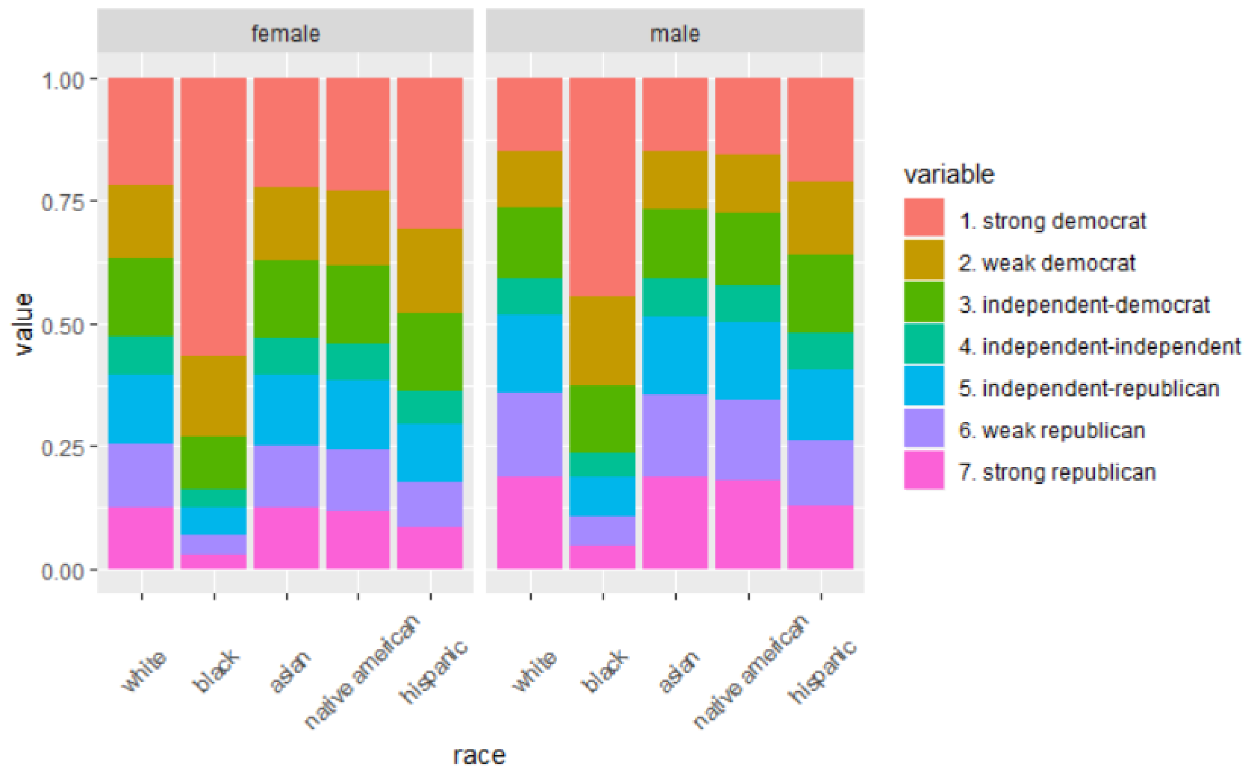
1. Summarize the parameter estimates numerically and also graphically.

```
data<-nes_data_comp[,c("partyid7","gender","race")]
data<-na.omit(data)

# Summarize the parameter estimates numerically
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

data<-data %>%
  count(partyid7,gender,race)

r_nes<-polr(partyid7~factor(gender) + factor(race) ,weights = n, data =
```

```
 data)
round(summary(r_nes)$coef,2)

##
## Re-fitting to get Hessian

##                                                        Value Std. Erro
r
## factor(gender)female                                   -0.49       0.1
5
## factor(race)black                                      -1.54       0.2
5
## factor(race)asian                                      -0.01       0.5
4
## factor(race)native american                            -0.06       0.3
5
## factor(race)hispanic                                   -0.45       0.2
8
## 1. strong democrat|2. weak democrat                    -1.76       0.1
5
## 2. weak democrat|3. independent-democrat               -1.03       0.1
3
## 3. independent-democrat|4. independent-independent     -0.38       0.1
3
## 4. independent-independent|5. independent-republican -0.07       0.1
3
## 5. independent-republican|6. weak republican            0.59       0.1
3
## 6. weak republican|7. strong republican                 1.46       0.1
5
##                                                        t value
## factor(gender)female                                     -3.18
## factor(race)black                                        -6.24
## factor(race)asian                                        -0.02
## factor(race)native american                              -0.17
## factor(race)hispanic                                     -1.64
## 1. strong democrat|2. weak democrat                     -11.91
## 2. weak democrat|3. independent-democrat                 -7.66
## 3. independent-democrat|4. independent-independent       -2.97
## 4. independent-independent|5. independent-republican   -0.56
## 5. independent-republican|6. weak republican              4.53
## 6. weak republican|7. strong republican                   9.77

newdata.nes<- expand.grid(gender=c("female","male"),race=c("white","bla
ck","asian","native american","hispanic"))
pre.nes <-predict(r_nes,newdata=newdata.nes,type="probs")

# Summarize the parameter estimates graphically
ggplot(melt(cbind(newdata.nes,pre.nes),id.vars = c("gender","race")))+
  geom_bar(stat="identity")+aes(x=race,y=value, fill=variable)+theme(ax
```

```
is.text.x = element_text(angle = 45, hjust = 0.5, vjust = 0.5)) +
  facet_grid(~gender)
```

```
                                                      Value Std. Error t value
factor(gender)female                                  -0.49       0.15   -3.18
factor(race)black                                     -1.54       0.25   -6.24
factor(race)asian                                     -0.01       0.54   -0.02
factor(race)native american                           -0.06       0.35   -0.17
factor(race)hispanic                                  -0.45       0.28   -1.64
1. strong democrat|2. weak democrat                   -1.76       0.15  -11.91
2. weak democrat|3. independent-democrat              -1.03       0.13   -7.66
3. independent-democrat|4. independent-independent    -0.38       0.13   -2.97
4. independent-independent|5. independent-republican  -0.07       0.13   -0.56
5. independent-republican|6. weak republican           0.59       0.13    4.53
6. weak republican|7. strong republican                1.46       0.15    9.77
```




*The graph shows that there is no significant difference between female and male. The black has the most number of strong democrat.* 2. Explain the results from the fitted model.

$log(\frac{\pi_2+\pi_3+\pi_4+\pi_5+\pi_6+\pi_7}{\pi_1}) = -1.76 - 0.49 * female - 1.54 * black - 0.01 * asian - 0.06 * native_american - 0.45 * hispanic$ *For a white male, the log odds of he is a weak democrat or independent-democrat or independent-independent or independent-republican or weak republican or strong repiblican is -1.76.*

$log(\frac{\pi_3+\pi_4+\pi_5+\pi_6+\pi_7}{\pi_1+\pi_2}) = -1.03 - 0.49 * female - 1.54 * black - 0.01 * asian - 0.06 * native_american - 0.45 * hispanic$ *For a white male, the log odds of he is an* *independent-democrat or independent-independent or independent-republican or weak republican or strong repiblican is -1.03.*

$$log(\frac{\pi_4+\pi_5+\pi_6+\pi_7}{\pi_1+\pi_2+\pi_3})$$
$$= -0.38 - 0.49 * female - 1.54 * black - 0.01 * asian - 0.06 * native_american$$
$$- 0.45 * hispanic$$

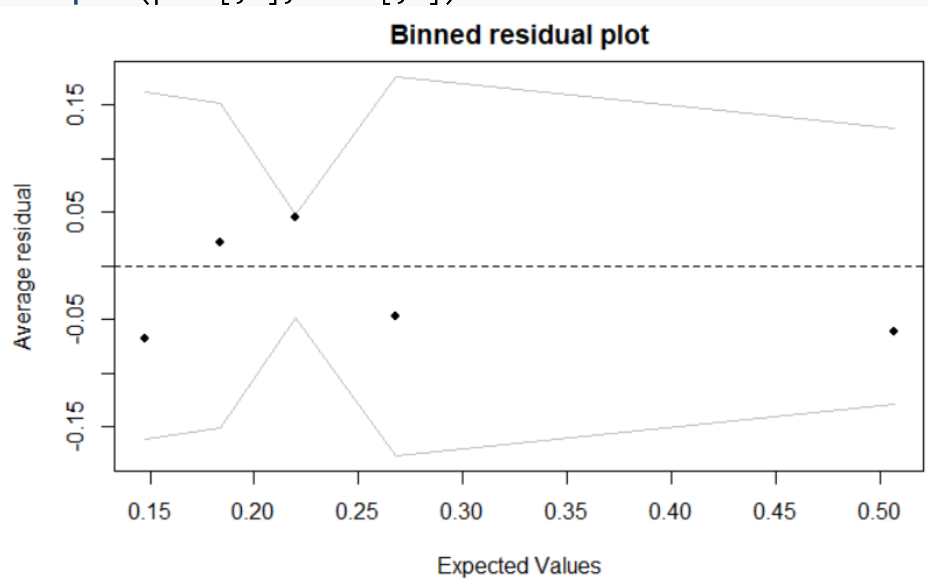
$$log(\frac{\pi_5+\pi_6+\pi_7}{\pi_1+\pi_2+\pi_3+\pi_4})$$
$$= -0.07 - 0.49 * female - 1.54 * black - 0.01 * asian - 0.06 * native_american$$
$$- 0.45 * hispanic$$

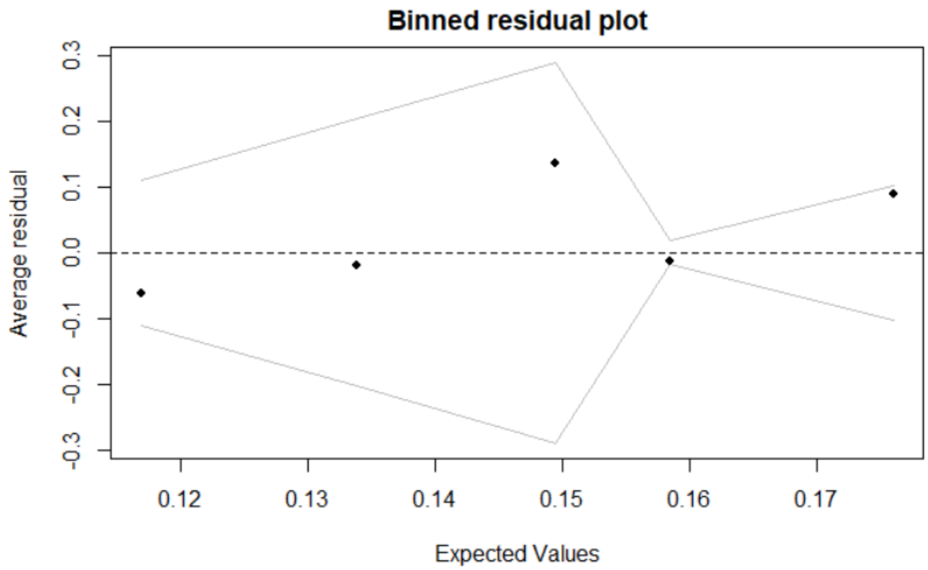
$$log(\frac{\pi_6+\pi_7}{\pi_1+\pi_2+\pi_3+\pi_4+\pi_5})$$
$$= 0.59 - 0.49 * female - 1.54 * black - 0.01 * asian - 0.06 * native_american$$
$$- 0.45 * hispanic$$

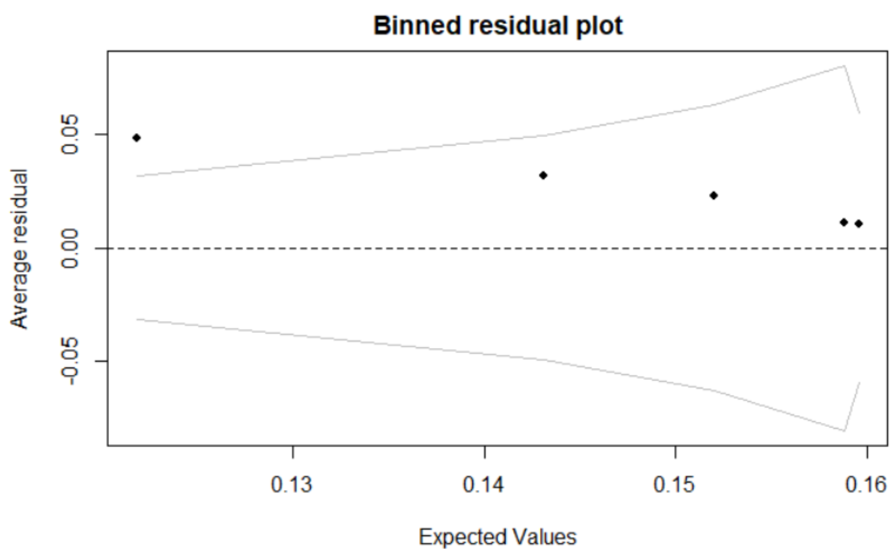
$$log(\frac{\pi_7}{\pi_1+\pi_2+\pi_3+\pi_4+\pi_5+\pi_6})$$
$$= 1.46 - 0.49 * female - 1.54 * black - 0.01 * asian - 0.06 * native_american$$
$$- 0.45 * hispanic$$

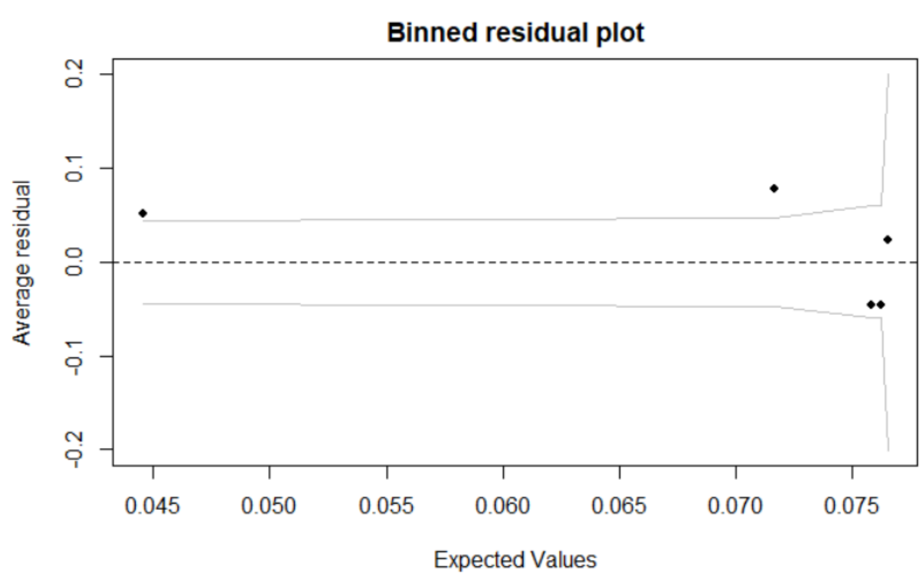
3. Use a binned residual plot to assess the fit of the model.

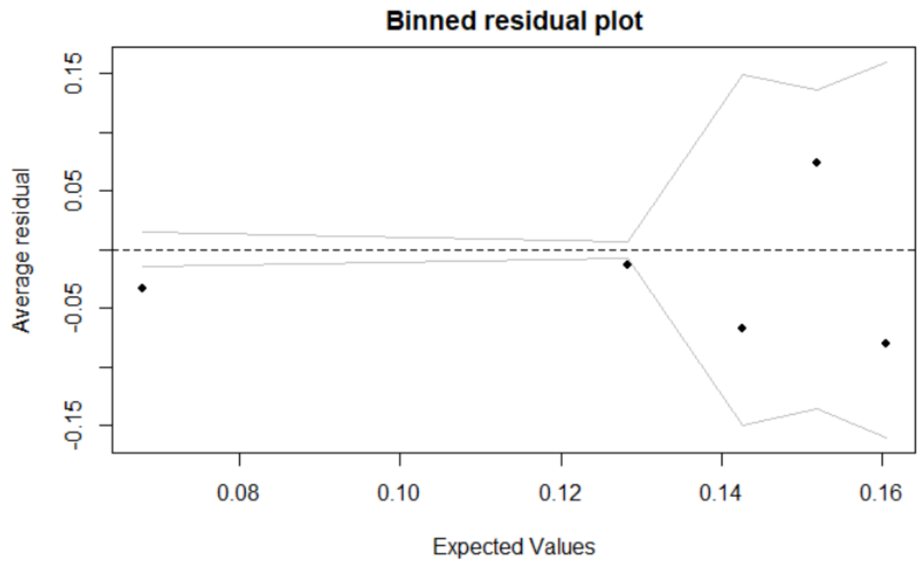
```
library(dplyr)
newdata<-dcast(data, gender + race ~ partyid7, value.var = "n")
newdata[is.na(newdata)]<-0
newdata<-newdata %>%
  mutate(total=apply(newdata[,3:9],1,sum))
newdata[,3:9]<-round(newdata[,3:9]/newdata[,"total"],2)
pred<-predict(r_nes,newdata=newdata,type="p")
resid<-newdata[,3:9]-pred

for(i in 1:7)
  binnedplot(pred[,i],resid[,i])
```

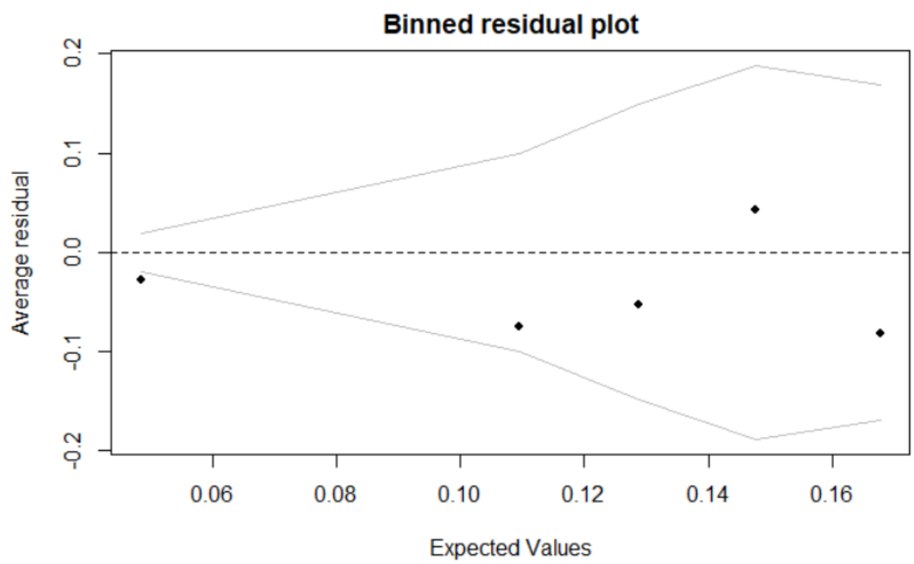


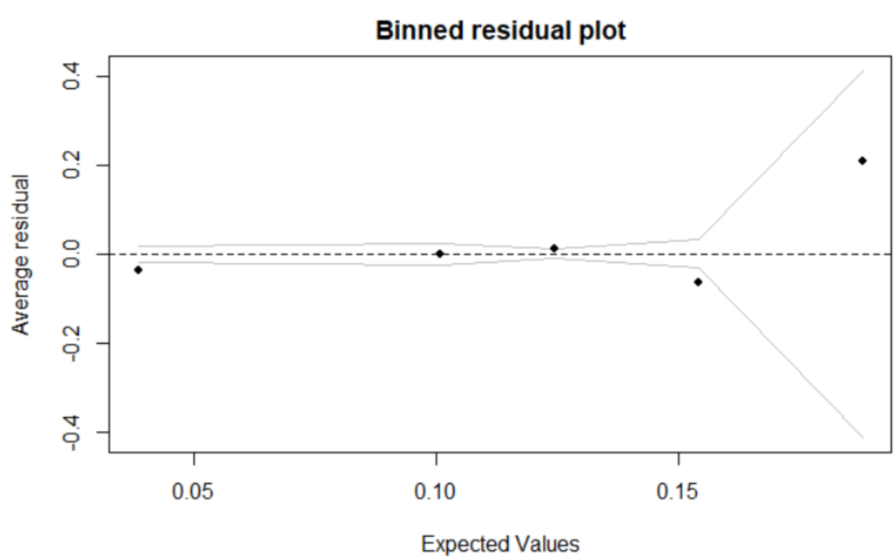

Binned residual plot
Average residual
Expected Values
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0.12
0.13
0.14
0.15
0.16
0.17


Binned residual plot
Average residual
Expected Values
0.05
0.00
-0.05
0.13
0.14
0.15
0.16


Binned residual plot
Average residual
Expected Values
0.2
0.1
0.0
-0.1
-0.2
0.045
0.050
0.055
0.060
0.065
0.070
0.075


Binned residual plot
Average residual
0.15
0.05
-0.05
-0.15
0.08
0.10
0.12
0.14
0.16
Expected Values


Binned residual plot
Average residual
0.2
0.1
0.0
-0.1
-0.2
0.06
0.08
0.10
0.12
0.14
0.16
Expected Values


Binned residual plot
Average residual
0.4
0.2
0.0
-0.2
-0.4
0.05
0.10
0.15
Expected Values

*Residual plots shows that the there are extreme residuals beyond the CI,so that the model does not fit well.*

## 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 鈥撹 cademic, vocational, or general 鈥攖 hat the students pursue in high school. The response is multinomial with three levels.

```
data(hsb)
?hsb

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

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

```
model<-nnet::multinom(prog ~ gender+race+ses+schtyp+read+write+math+sci
ence+socst , data=hsb)

## # weights:  42 (26 variable)
## initial  value 219.722458
## iter  10 value 171.814970
## iter  20 value 153.793692
## iter  30 value 152.935260
## final  value 152.935256
## converged

summary(model)

## Call:
## nnet::multinom(formula = prog ~ gender + race + ses + schtyp +
##     read + write + math + science + socst, data = hsb)
##
## Coefficients:
##          (Intercept)  gendermale raceasian racehispanic racewhite
## general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156
## vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881
##               seslow sesmiddle schtyppublic        read       write
## general  1.09864111 0.7029621    0.5845405 -0.04418353 -0.03627381
## vocation 0.04747323 1.1815808    2.0553336 -0.03481202 -0.03166001
##                math    science       socst
## general  -0.1092888 0.10193746 -0.01976995
## vocation -0.1139877 0.05229938 -0.08040129
##
## Std. Errors:
```

```
##          (Intercept) gendermale raceasian racehispanic racewhite 
seslow
## general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6
066763
## vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7
045772
##          sesmiddle schtyppublic       read      write       math
## general  0.5045938    0.5642925 0.03103707 0.03381324 0.03522441
## vocation 0.5700833    0.8348229 0.03422409 0.03585729 0.03885131
##             science      socst
## general  0.03274038 0.02712589
## vocation 0.03424763 0.02938212
## 
## Residual Deviance: 305.8705 
## AIC: 357.8705
```

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

```
student<-hsb[hsb[,"id"]==99,]
predict(model,student,type="p")

##  academic   general  vocation 
## 0.5076752 0.3753090 0.1170158
```

## 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.

```
model<-polr(factor(happy)~money+factor(sex)+factor(love)+factor(work),d
ata=happy)
summary(model)

## 
## 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  1.640177
## factor(sex)1  -1.02521    0.93630 -1.094956
```

```
## factor(love)2  3.45774    1.56126  2.214714
## factor(love)3  7.85068    1.85206  4.238881
## factor(work)2 -1.18913    1.68767 -0.704601
## factor(work)3  0.01566    1.58057  0.009907
## factor(work)4  1.84633    1.53697  1.201275
## factor(work)5  0.64775    2.14985  0.301298
##
## Intercepts:
##      Value   Std. Error t value
## 2|3  -0.8392  1.8387    -0.4564
## 3|4   0.0098  1.7713     0.0055
## 4|5   2.4279  2.0150     1.2050
## 5|6   4.4745  2.1064     2.1243
## 6|7   5.0676  2.1243     2.3855
## 7|8   7.3974  2.2303     3.3168
## 8|9  11.3106  2.5925     4.3628
## 9|10 13.0852  2.7917     4.6872
##
## Residual Deviance: 90.47841
## AIC: 122.4784
```

2. Interpret the parameters of your chosen model. $log(\frac{\pi_3+\ldots+\pi_{10}}{\pi_2}) = -0.84 + 0.02 * money - 1.03 * sex_1 + 3.46 * love_2 + 7.85 * love_3 - 1.19 * work_2 + 0.02 * work_3 + 1.85 * work_4 + 0.65 * work_5$

*For people whose family income is 0, and is unsatisfactory with sex and feel lonly, and has no job, log odds of that one with happy index from 3 to 10 over him or her with happy index = 2, is -0.84*

```
kable(cbind(happy[,2:5],round(fitted(model),2)))
```

| money | sex | love | work | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 36 | 0 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.68 | 0.21 | 0.06 |
| 47 | 1 | 3 | 1 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.36 | 0.54 | 0.02 | 0.00 |
| 53 | 0 | 3 | 5 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.10 | 0.75 | 0.11 | 0.03 |
| 35 | 1 | 3 | 3 | 0.00 | 0.00 | 0.01 | 0.04 | 0.04 | 0.40 | 0.50 | 0.02 | 0.00 |
| 88 | 1 | 1 | 2 | 0.45 | 0.21 | 0.30 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 175 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.37 | 0.40 | 0.22 |
| 175 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.37 | 0.40 | 0.22 |
| 45 | 0 | 2 | 3 | 0.01 | 0.01 | 0.12 | 0.41 | 0.14 | 0.27 | 0.04 | 0.00 | 0.00 |
| 35 | 1 | 2 | 2 | 0.06 | 0.07 | 0.50 | 0.29 | 0.03 | 0.04 | 0.00 | 0.00 | 0.00 |
| 55 | 1 | 1 | 4 | 0.07 | 0.08 | 0.51 | 0.28 | 0.03 | 0.03 | 0.00 | 0.00 | 0.00 |
| 40 | 0 | 2 | 3 | 0.01 | 0.01 | 0.13 | 0.42 | 0.14 | 0.25 | 0.04 | 0.00 | 0.00 |
| 45 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.10 | 0.75 | 0.11 | 0.03 |
| 45 | 1 | 3 | 3 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.37 | 0.54 | 0.02 | 0.00 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 45 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.10 | 0.75 | 0.11 | 0.03 |
| 62 | 0 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.59 | 0.28 | 0.09 |
| 44 | 1 | 2 | 3 | 0.02 | 0.02 | 0.27 | 0.47 | 0.09 | 0.12 | 0.01 | 0.00 | 0.00 |
| 85 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.70 | 0.19 | 0.05 |
| 81 | 0 | 2 | 4 | 0.00 | 0.00 | 0.01 | 0.08 | 0.06 | 0.50 | 0.33 | 0.01 | 0.00 |
| 112 | 1 | 3 | 2 | 0.00 | 0.00 | 0.00 | 0.04 | 0.03 | 0.37 | 0.53 | 0.02 | 0.00 |
| 40 | 1 | 2 | 4 | 0.00 | 0.00 | 0.06 | 0.30 | 0.15 | 0.40 | 0.08 | 0.00 | 0.00 |
| 40 | 1 | 3 | 3 | 0.00 | 0.00 | 0.01 | 0.04 | 0.03 | 0.38 | 0.52 | 0.02 | 0.00 |
| 44 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.10 | 0.75 | 0.11 | 0.03 |
| 35 | 1 | 2 | 2 | 0.06 | 0.07 | 0.50 | 0.29 | 0.03 | 0.04 | 0.00 | 0.00 | 0.00 |
| 56 | 1 | 2 | 3 | 0.01 | 0.02 | 0.23 | 0.47 | 0.10 | 0.15 | 0.02 | 0.00 | 0.00 |
| 115 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.61 | 0.27 | 0.09 |
| 44 | 1 | 2 | 4 | 0.00 | 0.00 | 0.06 | 0.29 | 0.14 | 0.41 | 0.09 | 0.00 | 0.00 |
| 50 | 1 | 3 | 3 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.35 | 0.56 | 0.02 | 0.00 |
| 45 | 0 | 2 | 4 | 0.00 | 0.00 | 0.02 | 0.14 | 0.10 | 0.52 | 0.21 | 0.00 | 0.00 |
| 41 | 0 | 3 | 5 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.12 | 0.75 | 0.09 | 0.02 |
| 50 | 0 | 3 | 5 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.11 | 0.75 | 0.10 | 0.02 |
| 85 | 1 | 1 | 1 | 0.21 | 0.17 | 0.49 | 0.11 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
| 90 | 1 | 2 | 2 | 0.02 | 0.03 | 0.34 | 0.44 | 0.07 | 0.09 | 0.01 | 0.00 | 0.00 |
| 85 | 1 | 2 | 4 | 0.00 | 0.00 | 0.03 | 0.18 | 0.12 | 0.51 | 0.16 | 0.00 | 0.00 |
| 75 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.72 | 0.17 | 0.04 |
| 70 | 0 | 2 | 3 | 0.00 | 0.01 | 0.08 | 0.35 | 0.15 | 0.35 | 0.06 | 0.00 | 0.00 |
| 0 | 0 | 2 | 2 | 0.04 | 0.05 | 0.45 | 0.36 | 0.04 | 0.05 | 0.01 | 0.00 | 0.00 |
| 31 | 0 | 2 | 4 | 0.00 | 0.00 | 0.03 | 0.17 | 0.11 | 0.51 | 0.17 | 0.00 | 0.00 |
| 60 | 1 | 3 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.74 | 0.14 | 0.03 |
| 65 | 1 | 3 | 3 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.30 | 0.61 | 0.03 | 0.01 |

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

```
money<-as.data.frame(30)
sex<-as.data.frame(0)
love<-as.data.frame(1)
work<-as.data.frame(1)
sample<-cbind(money,sex,love,work)
colnames(sample)<-c("money","sex","love","work")
pred<-predict(model,sample,type="p")
kable(pred)
```

| | x |
|---|---|
| 2 | 0.2019701 |

| | |
|---|---|
| 3 | 0.1697169 |
| 4 | 0.4974261 |
| 5 | 0.1118049 |
| 6 | 0.0084455 |
| 7 | 0.0095913 |
| 8 | 0.0010242 |
| 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
model<-polr(policy~sex+year,weights=y,data=uncviet)
summary(model)

##
## 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
```

$$\frac{P(policy\ isB\ or\ C\ or\ D)}{P(policy\ is\ A)} = exp(-1.11 - 0.65 * sex_male + 1.18 * year_Grad + 0.40 * year_Junior + 0.54 * year_Senior + 0.13 * year_Soph)$$

*With same year of school, the odds of a male has opinions B (follow the present policy) or C (withdraw troops to strong points and open negotiations on elections involving the Viet Cong) or D (immediate withdrawal of all U.S. troops) is exp(-0.65) times of the same odds of a female.*

*With same gender, the odds of a Graduate has opinions B (follow the present policy) or C (withdraw troops to strong points and open negotiations on elections involving the Viet Cong) or D (immediate withdrawal of all U.S. troops) is exp(1.18) times of the same odds of a Freshman.*

*With same gender, the odds of a Junior student has opinions B (follow the present policy) or C (withdraw troops to strong points and open negotiations on elections involving the Viet Cong) or D (immediate withdrawal of all U.S. troops) is exp(0.40) times of the same odds of a Freshman.*

*With same gender, the odds of a Senior student has opinions B (follow the present policy) or C (withdraw troops to strong points and open negotiations on elections involving the Viet Cong) or D (immediate withdrawal of all U.S. troops) is exp(0.54) times of the same odds of a Freshman.*

*With same gender, the odds of a Soph student has opinions B (follow the present policy) or C (withdraw troops to strong points and open negotiations on elections involving the Viet Cong) or D (immediate withdrawal of all U.S. troops) is exp(0.13) times of the same odds of a Freshman.*

$$\frac{P(policy\ isC\ or\ D)}{P(policy\ is\ A)} = exp(-0.01 - 0.65 * sex_male + 1.18 * year_Grad + 0.40 * year_Junior + 0.54 * year_Senior + 0.13 * year_Soph)$$

$$\frac{P(policy\ isD)}{P(policy\ is\ A\ or\ B\ or\ C)} = exp(2.44 - 0.65 * sex_male + 1.18 * year_Grad + 0.40 * year_Junior + 0.54 * year_Senior + 0.13 * year_Soph)$$

## 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.

```r
library(faraway)
data(pneumo,package="faraway")
?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
model_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(model_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

miner<-as.data.frame(25)
colnames(miner)<-"year"
predict(model_nominal,newdata=miner,type="p")

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

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

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

```r
predict(model_ornidal,newdata=miner,type="p")
```

```
##       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
data_normal<-pneumo[pneumo$status=="normal",]
data_mild<-pneumo[pneumo$status=="mild",]
data_severe<-pneumo[pneumo$status=="severe",]
data_abnormal<-rbind(data_mild,data_severe)
data_abnormal$status<-rep("abnormal",dim(data_abnormal)[1])
data_h1<-rbind(data_normal,data_abnormal)
data_h1$status<-factor(data_h1$status)
model_h1<-multinom(status~year,weights=Freq, data=data_h1)
```

```
## # weights:  3 (2 variable)
## initial  value 257.157604
## final  value 152.963516
## converged
```

```r
summary(model_h1)
```

```
## Call:
## multinom(formula = status ~ year, data = data_h1, weights = Freq)
##
## Coefficients:
##                   Values  Std. Err.
## (Intercept) -3.96635181 0.41892897
## year         0.09626924 0.01236388
##
## Residual Deviance: 305.927
## AIC: 309.927
```

```r
data_h2<-rbind(data_mild,data_severe)
data_h2$status<-factor(data_h2$status)
model_h2<-multinom(status~year,weights=Freq,data=data_h2)
```

```
## # weights:  3 (2 variable)
## initial  value 56.838069
## final  value 55.444195
## converged
```

```
summary(model_h2)
```

```
## Call:
## multinom(formula = status ~ year, data = data_h2, weights = Freq)
##
## Coefficients:
##                   Values  Std. Err.
## (Intercept) -1.11342251 0.86248390
## year         0.03547178 0.02350152
##
## Residual Deviance: 110.8884
## AIC: 114.8884
```

```
predict(model_h1,miner,type="p")
```

```
##         1
## 0.1737011
```

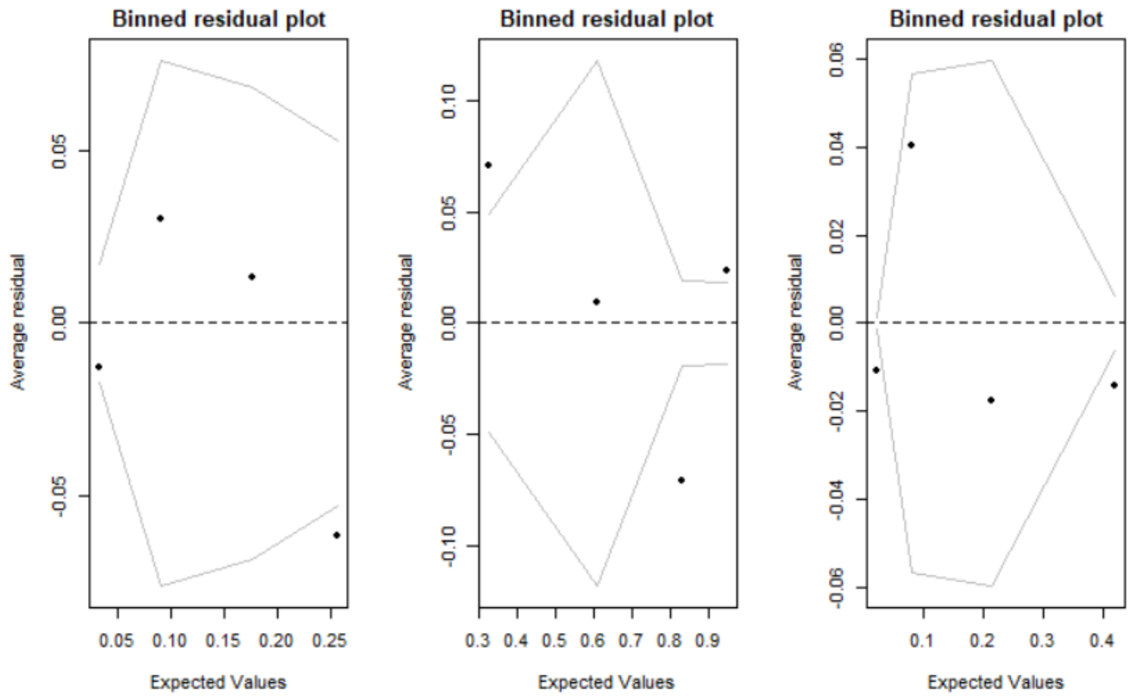
```
predict(model_h2,miner,type="p")
```

```
##         1
## 0.4435842
```

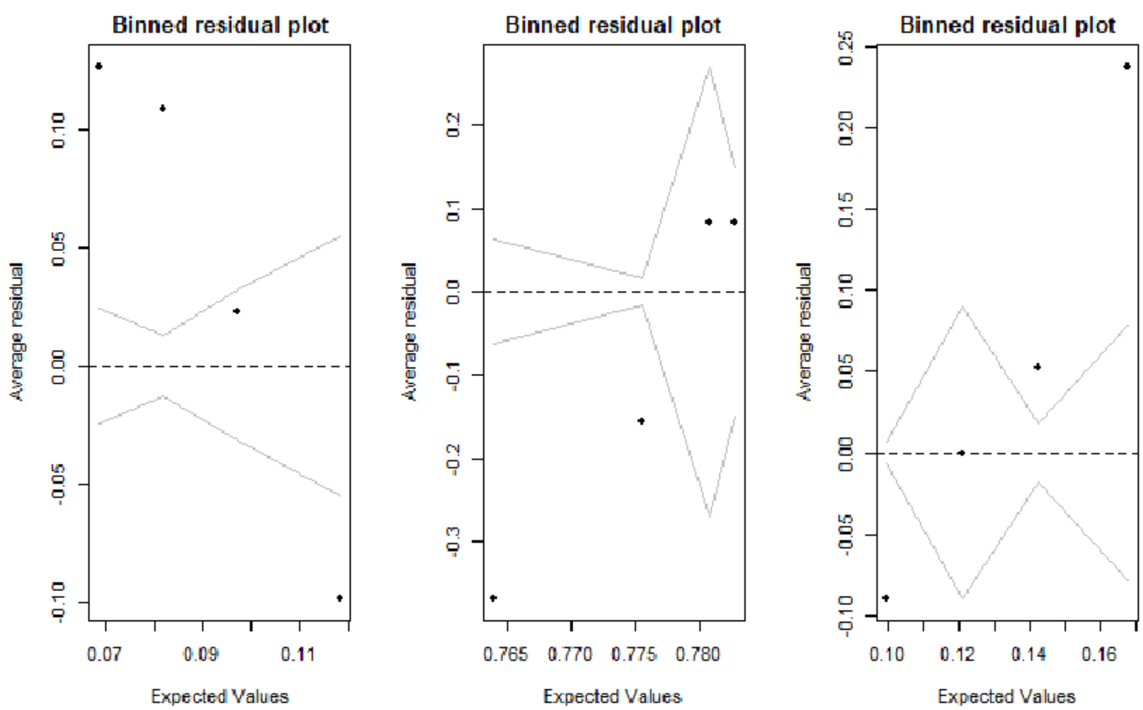
*It shows that the predicted probability of this miner is abnormal is 0.17, and the probability of this miner is normal is 0.83. Then we go on for the second model, the probability of his disease is mild is* $(1-0.44)*0.17=0.1$, *and the probability of his disease is severe is.* $0.44*0.17=0.07$

4. Compare the three analyses.

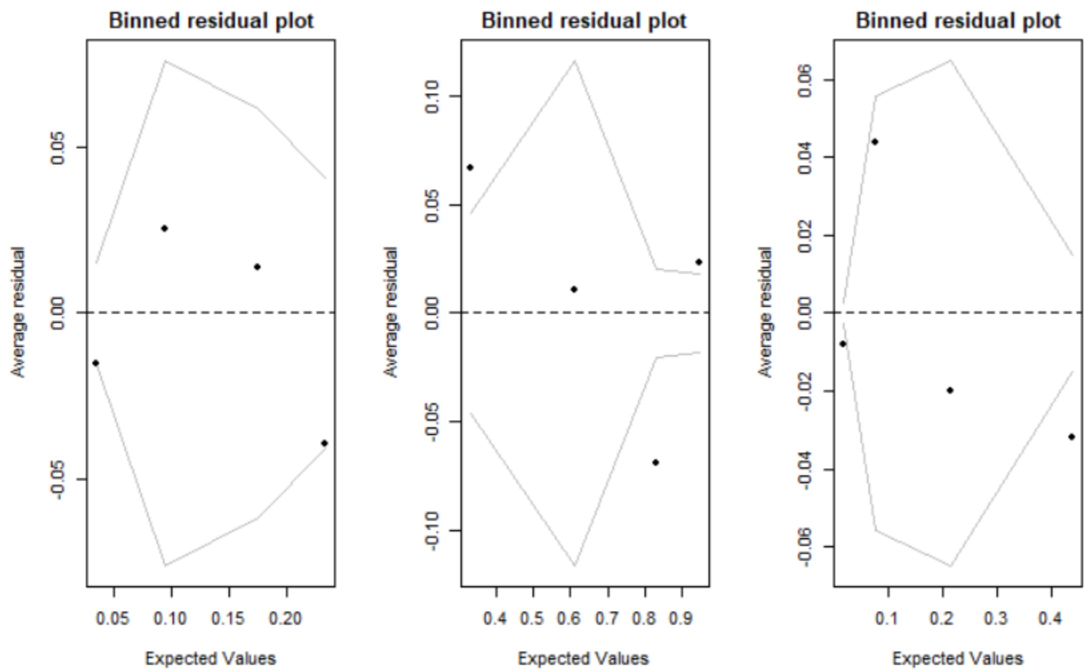
```
#residual plot1
pneumo_new<-dcast(pneumo, year ~ status, value.var = "Freq")
pneumo_new<-pneumo_new %>%
  mutate(total=apply(pneumo_new[,2:4],1,sum))
pneumo_new[,2:4]<-round(pneumo_new[,2:4]/pneumo_new[,"total"],2)
pred1<-predict(model_nominal,newdata=pneumo_new,type="p")
resid1<-pneumo_new[,2:4]-pred1

par(mfrow=c(1,3))
for(i in 1:3)
  binnedplot(pred1[,i],resid1[,i])
```




```
#residual plot2
pred2<-predict(model_ornidal,newdata=pneumo_new,type="p")
resid2<-pneumo_new[,2:4]-pred2

par(mfrow=c(1,3))
for(i in 1:3)
  binnedplot(pred2[,i],resid2[,i])
```



```
#residual plot3
p_abnormal<-predict(model_h1,pneumo_new,type="p")
p_normal<-1-p_abnormal
p_severe<-p_abnormal*predict(model_h2,pneumo_new,type="p")
p_mild<-p_abnormal*(1-predict(model_h2,pneumo_new,type="p"))
pred3<-cbind(p_mild,p_normal,p_severe)
resid3<-pneumo_new[,2:4]-pred3

par(mfrow=c(1,3))
for(i in 1:3)
  binnedplot(pred3[,i],resid3[,i])
```




*The first and third analysis have similar binned residual plots.*

*And most residuals of 1,3 analysis fall in the CI, but the 2 analysis have some extreme residuals, so that the 1,3 analysis are better.*

## (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.