MA678 homework 06

Multinomial Regression

Wendy Liang

Oct 22, 2020

Multinomial logit:

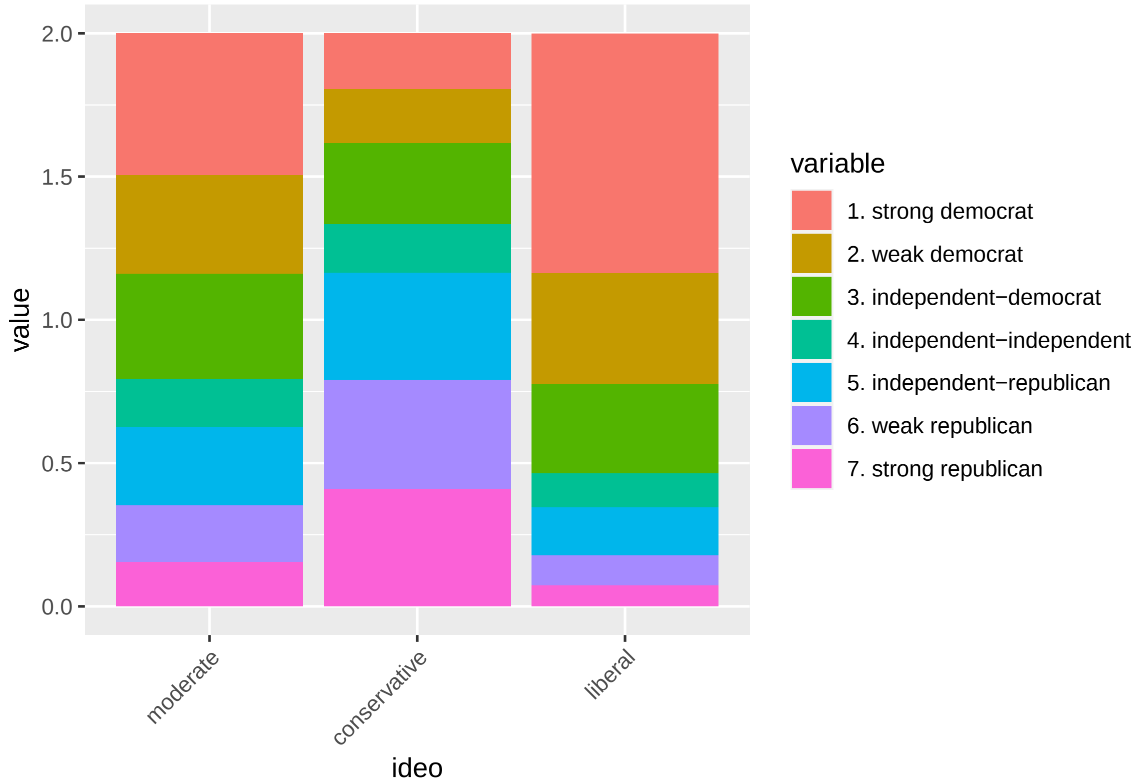
Using the individual-level survey data from the 2000 National Election Study (data in folder NES), predict party identification (which is on a five-point scale) using ideology and demographics with an ordered multinomial logit model.

1. Summarize the parameter estimates numerically and also graphically.

```
fit1 = polr(partyid7~factor(ideo)+factor(gender),data=nes52_comp)
summary(fit1)
```

```
##
## Re-fitting to get Hessian
## Call:
## polr(formula = partyid7 ~ factor(ideo) + factor(gender), data = nes52_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 t value
## 1. strong democrat|2. weak democrat                  -0.5176  0.1628    -3.1792
## 2. weak democrat|3. independent-democrat              0.2747  0.1600     1.7174
## 3. independent-democrat|4. independent-independent    1.0234  0.1667     6.1390
## 4. independent-independent|5. independent-republican  1.3914  0.1721     8.0826
## 5. independent-republican|6. weak republican          2.1535  0.1843    11.6852
## 6. weak republican|7. strong republican               3.0906  0.2047    15.0955
##
## Residual Deviance: 1975.189
## AIC: 1993.189
```

```
predx<- expand.grid(ideo = c("moderate","conservative","liberal"),gender=c("female","male"))
predy<-predict(fit1,newdata=predx,type = "p")
ggplot(melt(cbind(predx,predy),id.vars = c("gender","ideo")))+
  geom_histogram(aes(x=ideo,y=value, fill=variable),stat="identity")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
  facet_grid(~gender)
```

```
## <ggproto object: Class FacetGrid, Facet, gg>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map_data: function
##     params: list
##     setup_data: function
##     setup_params: function
##     shrink: TRUE
##     train_scales: function
##     vars: function
##     super:  <ggproto object: Class FacetGrid, Facet, gg>
```

2. Explain the results from the fitted model.

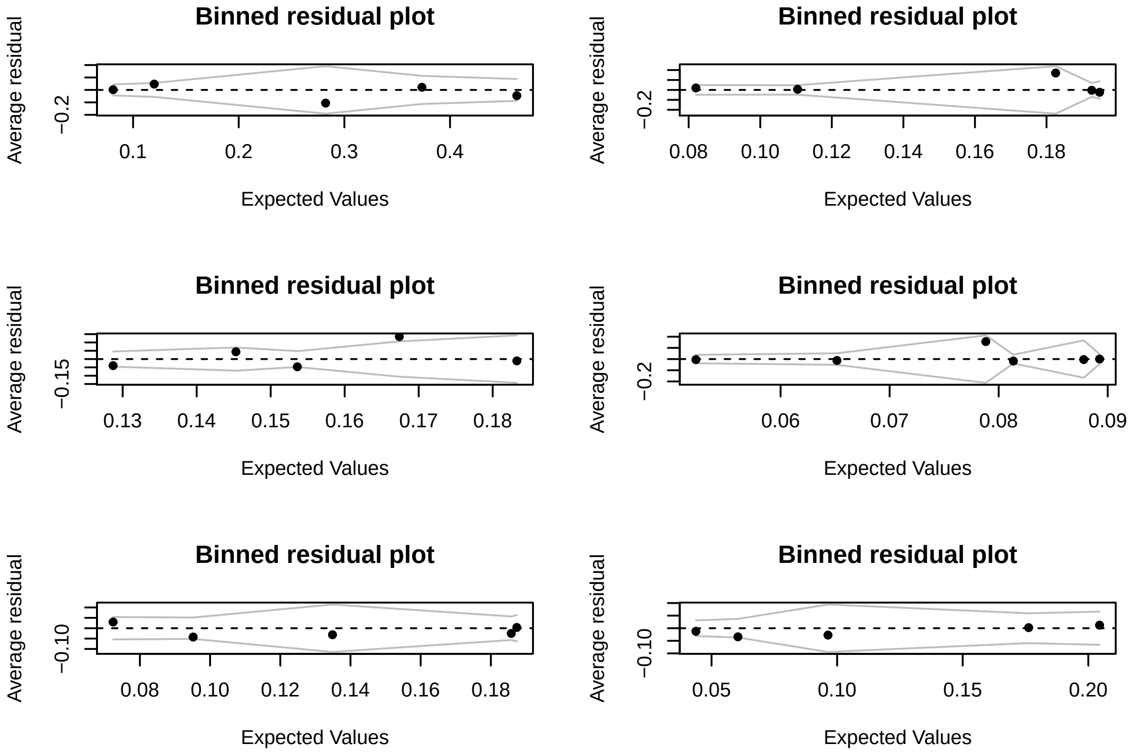
- log odds of not strong democrat $= log(\frac{\pi 2+\pi 3+\pi 4+\pi 5+\pi 6+\pi 7}{\pi 1}) = \beta_1 moderate + \beta_2 conservative + \beta_3 female - c12 = 0.5176 + 0.7859 * moderate + 1.9100 * conservative - 0.3703 * female$
- log odds of not strong democrat and no weak democrat $= log(\frac{\pi 3+\pi 4+\pi 5+\pi 6+\pi 7}{\pi 1+\pi 2}) = \beta_1 moderate + \beta_2 conservative + \beta_3 female - c23 = -0.2747 + 0.7859 * moderate + 1.9100 * conservative - 0.3703 * female$
- 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.

```
nes52_2 = cbind(partyid7=nes52_comp$partyid7, female=nes52_comp$female, ideo=nes52_comp$ideo)
nes52_2 <- data.frame(na.omit(nes52_2))
resid = model.matrix(~factor(partyid7)-1, data=nes52_2)-fitted(fit1)
par(mfrow=c(3,2))
for(i in 1:6){
  binnedplot(fitted(fit1)[,i], resid[,i])
}
```



Contingency table and ordered logit model

In a prospective study of a new living attenuated recombinant vaccine for influenza, patients were randomly allocated to two groups, one of which was given the new vaccine and the other a saline placebo. The responses were titre levels of hemagglutinin inhibiting antibody found in the blood six weeks after vaccination; they were categorized as “small”, “medium” or “large”.

treatment	small	moderate	large	Total
placebo	25	8	5	38
vaccine	6	18	11	35

The cell frequencies in the rows of table are constrained to add to the number of subjects in each treatment group (35 and 38 respectively). We want to know if the pattern of responses is the same for each treatment group.

```
##         treatment level count
## placebo        25     8     5
## vaccine         6    18    11
```

1. Using a chisqure test and an appropriate log-linear model, test the hypothesis that the distribution of responses is the same for the placebo and vaccine groups.

```
chisq.test(datatable)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  datatable
## X-squared = 17.648, df = 2, p-value = 0.0001472
```

2. For the model corresponding to the hypothesis of homogeniety of response distributions, calculate the fitted values, the Pearson and deviance residuals, and the goodness of fit statistics X^2 and D. Which of the cells of the table contribute most to X^2 and D? Explain and interpret these results.
3. Re-analyze these data using ordered logit model (use `polr`) to estiamte the cut-points of a latent continuous response varaible and to estimate a location shift between the two treatment groups. Sketch a rough diagram to illustrate the model which forms the conceptual base for this analysis.

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)
```

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

```
fit2 = multinom(prog~read+write+math+science+race,data = hsb,trace=FALSE,HESS=TRUE)
summary(fit2)
```

```
## Call:
## multinom(formula = prog ~ read + write + math + science + race, 
##     data = hsb, trace = FALSE, HESS = TRUE)
## 
## Coefficients:
##          (Intercept)        read       write       math    science  raceasian
## general     4.924957 -0.05388450 -0.03946933 -0.1071044 0.09229507 1.11489221
## vocation    8.777829 -0.05594167 -0.06281609 -0.1253231 0.05262485 0.08636574
##          racehispanic   racewhite
## general   -0.60687283 -0.01313942
## vocation   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(fit2,newdata=hsb[hsb$id==99,],type="probs")
```

```
##  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)
```

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

```
fit3 <-polr(factor(happy)~money+factor(sex)+factor(love)+factor(work),data=happy)
summary(fit3)
```

```
##
## 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. $log\frac{\pi3+...+\pi10}{\pi1+\pi2} = 0.84 + 0.0178money - 1.025sex_1 + 3.46love_2 + 7.85love_3 - 1.19work_2 + 0.02work_3 + 1.85work_4 + 0.65work_5$ For lonely people who are unsatisfactory with sex, with 0 family income, with happy index from 3 to 10 over the ones with happy index = 2, the log odd 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(fit3,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)
fit4 = polr(policy~sex+year,data=uncviet,weights = y,Hess = TRUE)
summary(fit4)
```

```
## 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
```

P(policy != A) is $exp(1.11 + 0.65sexmale + 1.18yearGrad + 0.40?yearJunior + 0.54yearSenior + 0.13yearSoph)$

Holding other variable constant

- A male has opinions B,C or D is 1- exp(-0.65)=48% lower than a female.
- A grad student has opinions B,C or D is exp(1.177)-1=224% higher than a freshman, holding other variable constant.
- A junior student has opinions B,C or D is exp(0.396)-1=49% higher than a freshman.
- A senior student has opinions B,C or D is exp(0.5444)-1=72% higher than a freshman.
- A sophomore student has opinions B,C or D is exp(1.1315)-1=14% 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.

```r
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.

```r
fit5_a = 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
summary(fit5_a)
```

```
## 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
```

```
pre1 = predict(fit5_a,newdata=data.frame(year=25),type = "probs")
```

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

```
fit5_b = polr(status~year,weights=Freq,data=pneumo)
summary(fit5_b)
```

```
##
## 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
```

```
pre2 = predict(fit5_b,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.

```
pneumo3=pneumo
pneumo3$status = as.character(pneumo3$status)
pneumo3$status[9:24] = "abnormal"
pneumo3$status = as.factor(pneumo3$status)
pneumo_abnormal = pneumo3[9:24, ]

fit5_normal = glm( status ~ year, data = pneumo3,
                   family = binomial(link = "logit"), weights = Freq)
fit5_abnormal = glm( status ~ year, data = pneumo_abnormal,
                     family = binomial(link = "logit"), weights = Freq)

normal=predict (fit5_normal, newdata=data.frame(year=25), type = "response")
severe=predict (fit5_abnormal, newdata=data.frame(year=25), type = "response") *
  (1-predict (fit5_normal, newdata=data.frame(year=25), type = "response"))
mild = (1-predict (fit5_abnormal, newdata=data.frame(year=25), type = "response")) *
  (1-predict (fit5_normal, newdata=data.frame(year=25), type = "response"))

pre3=c(normal,mild,severe)
```

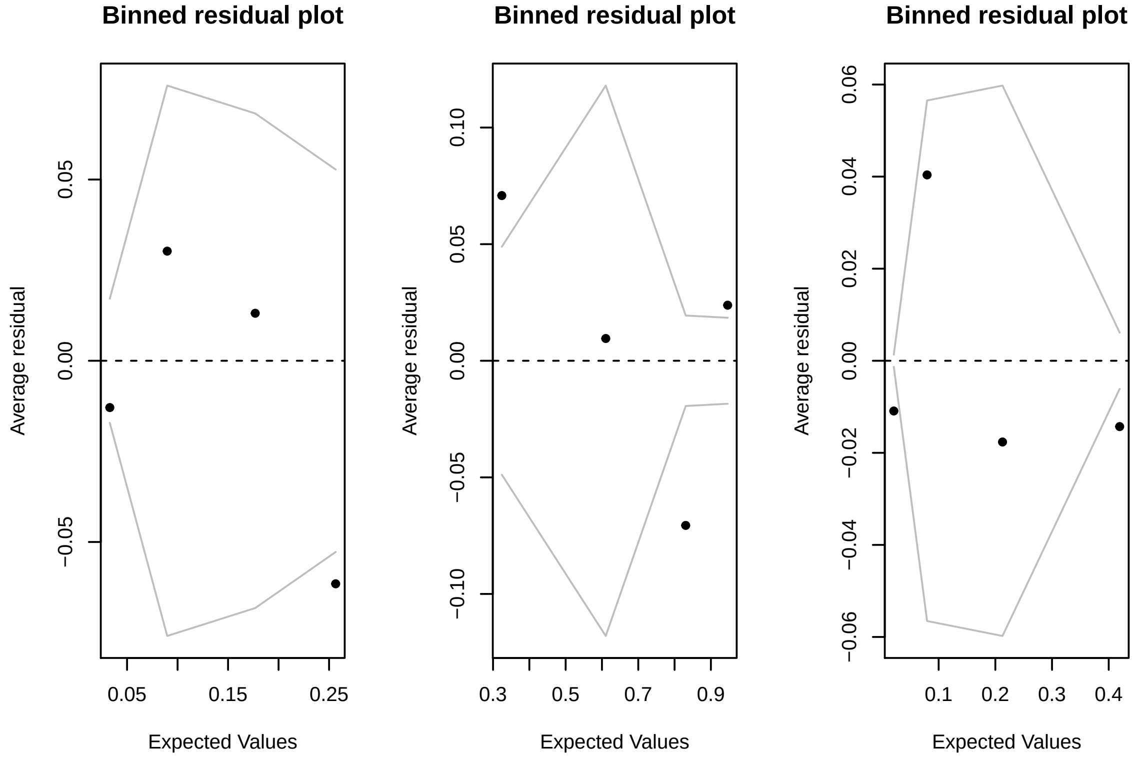
4. Compare the three analyses.

```
kable(rbind(pre1,pre2,pre3=pre3))
```

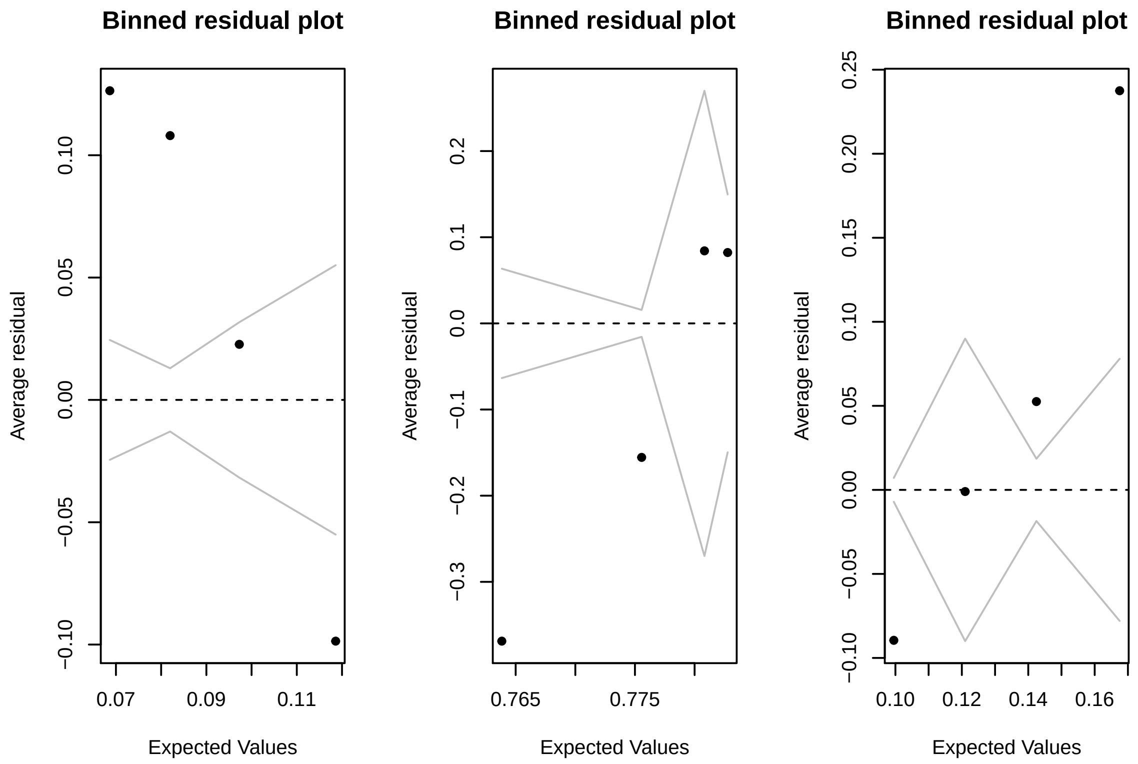
	mild	normal	severe
pre1	0.0914882	0.827787	0.0807248

	mild	normal	severe
pre2	0.0965236	0.781728	0.1217484
pre3	0.8262990	0.173701	0.0000000

```
pneumo4 = dcast(pneumo, year ~ status, value.var = "Freq")
pneumo4 %<>% mutate(total=apply(pneumo4[,2:4],1,sum))
pneumo4[,2:4] = round(pneumo4[,2:4]/pneumo4[,"total"],2)
pre1_1=predict(fit5_a,newdata=pneumo4,type="p")
resid1=pneumo4[,2:4]-pre1_1
par(mfrow=c(1,3))
for(i in 1:3){
  binnedplot(pre1_1[,i],resid1[,i])
  }
```



```
pre2_2<-predict(fit5_b,newdata=pneumo4,type="p")
resid2=pneumo4[,2:4]-pre2_2
par(mfrow=c(1,3))
for(i in 1:3){
  binnedplot(pre2_2[,i],resid2[,i])
  }
```



- The result of the first two models are similar. But they both don't fit well.