# MA678 homework 05

Multinomial Regression

*Your Name*

*September 2, 2017*

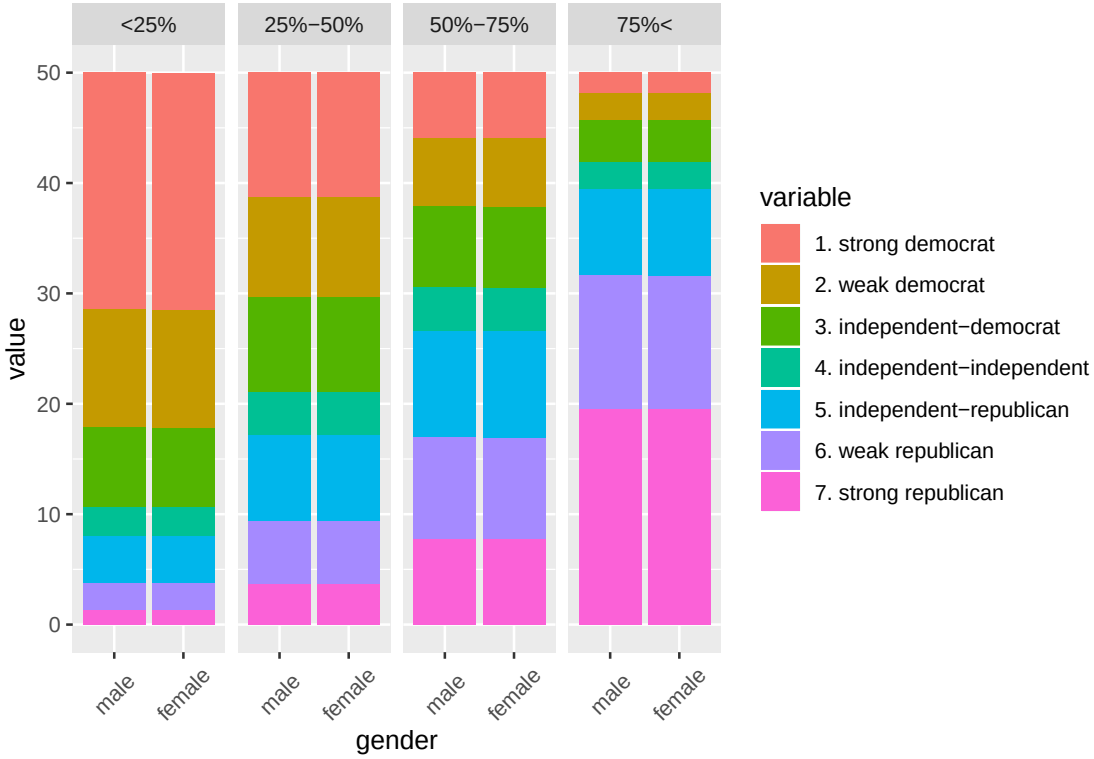
## Multinomial logit:

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.**

```
## Call:
## polr(formula = partyid7 ~ income + ideology + gender + race +
##     south, data = data, weights = n)
##
## Coefficients:
##                                Value Std. Error  t value
## income2. 17 to 33 percentile   0.168152     0.2287  0.73539
## income3. 34 to 67 percentile   0.224638     0.2240  1.00294
## income4. 68 to 95 percentile   0.396625     0.2279  1.74000
## income5. 96 to 100 percentile  0.889344     0.3107  2.86275
## ideology25%-50%                1.073798     0.1561  6.87682
## ideology50%-75%                1.930326     0.1851 10.42779
## ideology75%<                   3.272192     0.1950 16.78182
## genderfemale                  -0.004038     0.1200 -0.03364
## raceblack                     -1.680274     0.2199 -7.64110
## raceasian                     -0.242245     0.3557 -0.68102
## racenative american            0.076474     0.3277  0.23338
## racehispanic                  -0.298200     0.2697 -1.10555
## south1                         0.312260     0.1377  2.26719
##
## Intercepts:
##                                                      Value   Std. Error
## 1. strong democrat|2. weak democrat                 -0.2354  0.2412
## 2. weak democrat|3. independent-democrat             0.7314  0.2411
## 3. independent-democrat|4. independent-independent   1.5070  0.2450
## 4. independent-independent|5. independent-republican 1.8735  0.2484
## 5. independent-republican|6. weak republican         2.7475  0.2566
## 6. weak republican|7. strong republican              3.8489  0.2696
##                                                     t value
## 1. strong democrat|2. weak democrat                 -0.9761
## 2. weak democrat|3. independent-democrat             3.0344
## 3. independent-democrat|4. independent-independent   6.1506
## 4. independent-independent|5. independent-republican 7.5428
## 5. independent-republican|6. weak republican        10.7062
## 6. weak republican|7. strong republican             14.2760
##
## Residual Deviance: 3108.609
## AIC: 3146.609
```

```
##                                    2.5 %     97.5 %
## income2. 17 to 33 percentile  -0.27889261  0.6182341
## income3. 34 to 67 percentile  -0.21320816  0.6656069
## income4. 68 to 95 percentile  -0.04887078  0.8454597
## income5. 96 to 100 percentile  0.28286939  1.5019610
## ideology25%-50%                0.76896407  1.3812809
## ideology50%-75%                1.56951076  2.2954285
## ideology75%<                   2.89340971  3.6579955
## genderfemale                  -0.23924149  0.2314588
## raceblack                     -2.11750910 -1.2543077
## raceasian                     -0.94295846  0.4596835
## racenative american           -0.56706482  0.7222998
## racehispanic                  -0.82961030  0.2301452
## south1                         0.04246987  0.5825987
```




First in order to incorporate and interprte the effet of ideology, I divided it into 4 categories, We can see that, the possibility of belonging to each party identifications, there is only a slight difference between Male and Female in each categories of ideology.

**2. Explain the results from the fitted model.**

Because the responding variable is treated as ordinal, so the cumulative model can be interpreted as the distribution of Y, which is

$$P(Y > y_i) = logit^{-1}(X\beta_i)$$

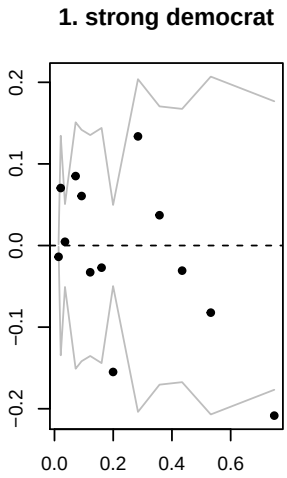
We can also interpret it as odds, which is:

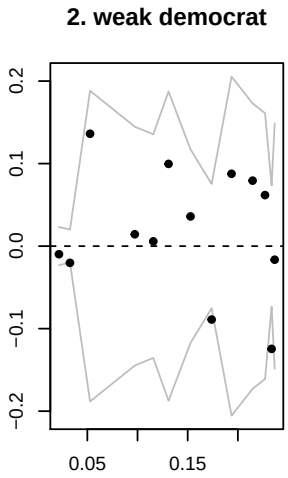
$$\log \frac{P(Y > y_i)}{P(Y \leqslant y_i)} = X\beta$$

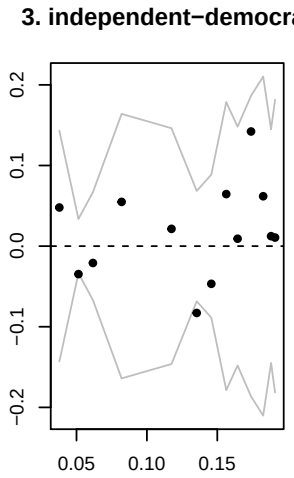
- $\log \frac{P(Y=2,3,4,5,6,7)}{P(Y=1)} = -0.2473 + X\beta$

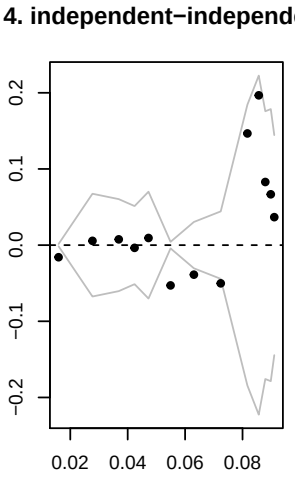
- $\log \frac{P(Y=3,4,5,6,7)}{P(Y=1,2)} = 0.7236 + X\beta$
- $\log \frac{P(Y=4,5,6,7)}{P(Y=1,2,3)} = 1.5005 + X\beta$
- $\log \frac{P(Y=5,6,7)}{P(Y=1,2,3,4)} = 1.8673 + X\beta$
- $\log \frac{P(Y=6,7)}{P(Y=1,2,3,4,5)} = 2.7418 + X\beta$
- $\log \frac{P(Y=7)}{P(Y=1,2,3,4,5,6)} = 3.8434 + X\beta$

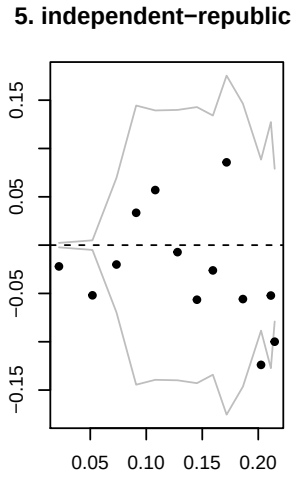
**3. Use a binned residual plot to assess the fit of the model.**

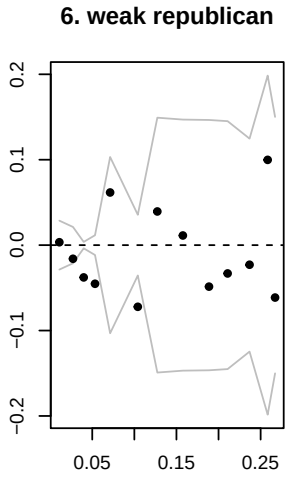


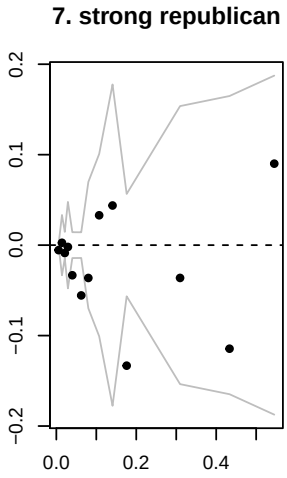



# High School and Beyond

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.

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

```
## # 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
```

```
## Call:
## nnet::multinom(formula = prog ~ gender + race + ses + schtyp +
##     read + write + math + science + socst, data = data)
##
## 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 0.6066763
## vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772
##          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.**

```
##  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`.

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

```
## Call:
## polr(formula = factor(happy) ~ money + factor(sex) + factor(love) +
##     factor(work), data = data, weights = n)
##
## Coefficients:
##                   Value Std. Error   t value
## money           0.01783    0.01087  1.640269
## factor(sex)1   -1.02504    0.93628 -1.094799
## factor(love)2   3.45763    1.56121  2.214710
## factor(love)3   7.85032    1.85199  4.238856
## factor(work)2 -1.18913    1.68764 -0.704613
## factor(work)3   0.01566    1.58055  0.009908
## factor(work)4   1.84616    1.53694  1.201193
## factor(work)5   0.64759    2.14981  0.301232
##
## Intercepts:
##       Value   Std. Error t value
## 2|3   -0.8388  1.8386    -0.4562
## 3|4    0.0101  1.7713     0.0057
## 4|5    2.4280  2.0149     1.2050
## 5|6    4.4745  2.1063     2.1243
## 6|7    5.0675  2.1242     2.3855
## 7|8    7.3972  2.2302     3.3168
## 8|9   11.3103  2.5925     4.3628
## 9|10 13.0848  2.7916     4.6872
##
## Residual Deviance: 90.47841
## AIC: 122.4784
```

**2. Interpret the parameters of your chosen model.**

Because the responding variable is treated as ordinal, so the cumulative model can be interpreted as the distribution of Y, which is

$$P(Y > y_i) = logit^{-1}(X\beta_i)$$

We can also interpret it as odds, which is:

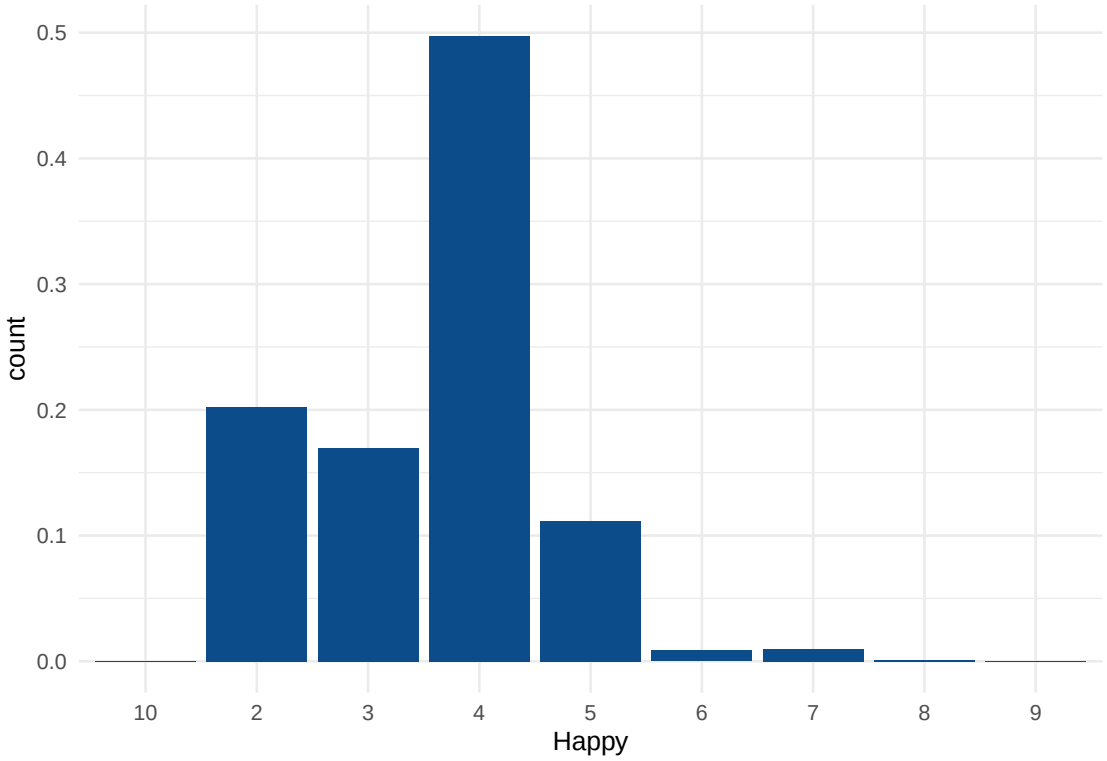
$$\log \frac{P(Y > y_i)}{P(Y \leqslant y_i)} = X\beta$$

- $\log \frac{P(Y=3,\ldots,10)}{P(Y=2)} = -0.8388 + X\beta$
- $\log \frac{P(Y=4,\ldots,10)}{P(Y=3)} = 0.0101 + X\beta$
- $\log \frac{P(Y=5,\ldots,10)}{P(Y=4)} = 2.4280 + X\beta$
- $\ldots$

**3. Predict the happiness distribution for subject whose parents earn $30,000 a year,**

who is lonely, not sexually active and has no job.

| | Probability | Happy |
|---|---|---|
| 2 | 0.2020282 | 2 |
| 3 | 0.1697189 | 3 |
| 4 | 0.4973656 | 4 |
| 5 | 0.1118034 | 5 |
| 6 | 0.0084461 | 6 |
| 7 | 0.0095925 | 7 |
| 8 | 0.0010244 | 8 |
| 9 | 0.0000174 | 9 |
| 10 | 0.0000035 | 10 |




## 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.

```
## Call:
## polr(formula = policy ~ sex + year, data = data, 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
```

Because the responding variable is treated as ordinal, so the cumulative model can be interpreted as the distribution of Y, which is

$$P(Y \geqslant y_i) = logit^{-1}(X\beta_i)$$

We can further calculate the density of each levels of Y by:

$$P(Y = y_i) = P(Y > y_i) - P(Y > y_{i-1})$$

- $\frac{P(Y=B,C,D)}{P(Y=A)} = exp\{-1.1098 + X\beta\}$
- $\frac{P(Y=C,D)}{P(Y=A,B)} = exp\{-0.0130 + X\beta\}$
- $\frac{P(Y=D)}{P(Y=A,B,C)} = exp\{2.4417 + X\beta\}$

## 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.

**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.**

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

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

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

```
##       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.**

```
## glm(formula = cbind(ill, n - ill) ~ year, family = binomial,
##     data = data)
##             coef.est coef.se
## (Intercept) -3.97     0.42
## year         0.10     0.01
## ---
```

```
##   n = 8, k = 2
##   residual deviance = 11.6, null deviance = 97.6 (difference = 86.0)
```

```
## glm(formula = cbind(severe, ill - severe) ~ year, family = binomial,
##     data = data)
##             coef.est coef.se
## (Intercept) -1.11     0.86
## year         0.04     0.02
## ---
##   n = 7, k = 2
##   residual deviance = 1.7, null deviance = 4.1 (difference = 2.3)
```

```
##        1
## 0.173701
```

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

```
## [1] 0.07705102
```

```
## [1] 0.09664998
```

```
##     Normal       Mild     Severe
## 1 0.826299 0.09664998 0.07705102
```

**4. Compare the three analyses.**

| | model 1 | model 2 | model 3 |
|---|---|---|---|
| Normal | 0.8277870 | 0.7817280 | 0.826299 |
| Mild | 0.0914882 | 0.0965236 | 0.096650 |
| Severe | 0.0807248 | 0.1217484 | 0.077051 |

We can see that the predictions result from model 1 and 3 are almost the same, while model has give higher possibility to severe disease.