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.

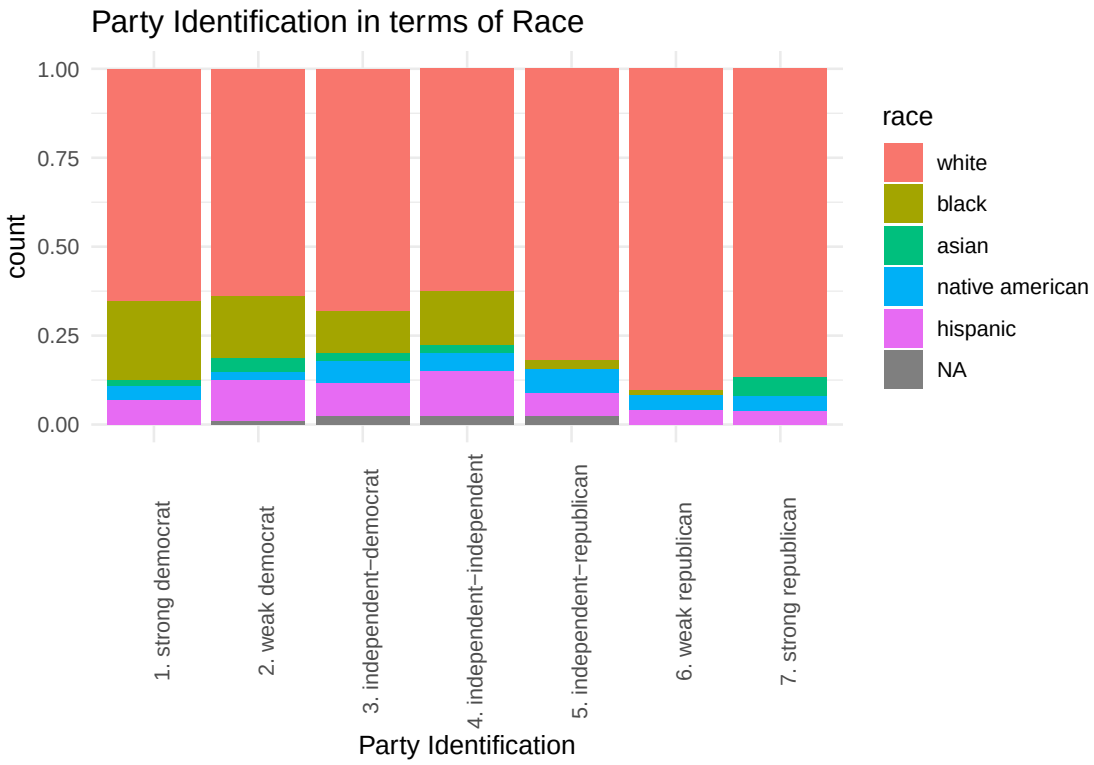
1. Summarize the parameter estimates numerically and also graphically.

```
summary(lm1)
```

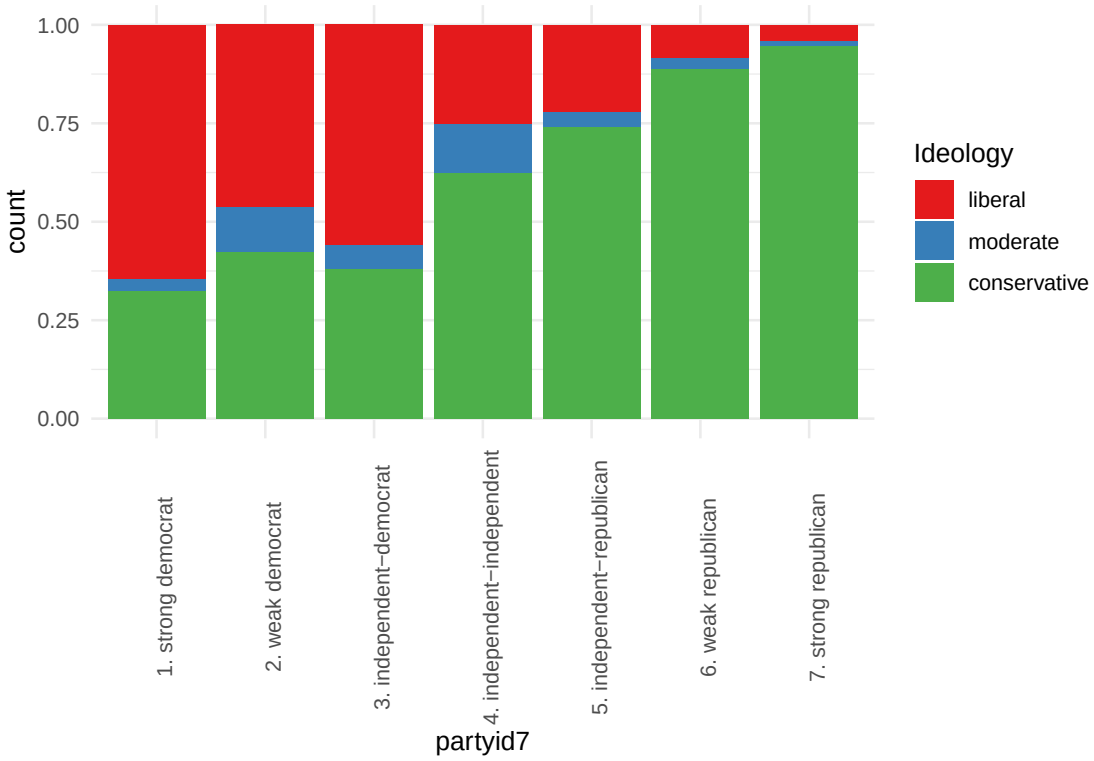
```
##
## Call:
## vglm(formula = ordered(partyid7) ~ ideo + age + gender + race +
##     income, family = cumulative(parallel = TRUE, reverse = TRUE),
##     data = nes_data_comp)
##
##
## Pearson residuals:
##                       Min      1Q  Median       3Q   Max
## logit(P[Y>=2]) -5.042  0.0871  0.1396  0.40531 2.430
## logit(P[Y>=3]) -3.516 -0.4050  0.1484  0.35255 3.323
## logit(P[Y>=4]) -3.146 -0.3932 -0.1243  0.30648 5.535
## logit(P[Y>=5]) -3.430 -0.3683 -0.1554  0.36897 7.089
## logit(P[Y>=6]) -2.525 -0.2957 -0.1424  0.41400 5.010
## logit(P[Y>=7]) -1.589 -0.2957 -0.1267 -0.07985 7.448
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                  0.598280   0.404649   1.479 0.139270
## (Intercept):2                 -0.260482   0.403841  -0.645 0.518920
## (Intercept):3                 -1.061418   0.407356  -2.606 0.009171 **
## (Intercept):4                 -1.464168   0.409883  -3.572 0.000354 ***
## (Intercept):5                 -2.275065   0.415756  -5.472 4.45e-08 ***
## (Intercept):6                 -3.265801   0.425641  -7.673 1.68e-14 ***
## ideomoderate                   1.080601   0.361231   2.991 0.002777 **
## ideoconservative               2.031104   0.183466  11.071  < 2e-16 ***
## age                           -0.012330   0.005061  -2.436 0.014844 *
## genderfemale                  -0.268804   0.159565  -1.685 0.092065 .
## raceblack                     -1.536745   0.283170  -5.427 5.73e-08 ***
## raceasian                      0.163766   0.523636   0.313 0.754473
## racenative american           -0.062166   0.371115  -0.168 0.866968
## racehispanic                  -0.529341   0.299292  -1.769 0.076954 .
## income2. 17 to 33 percentile   0.597286   0.285700   2.091 0.036563 *
## income3. 34 to 67 percentile   0.704946   0.267881   2.632 0.008499 **
## income4. 68 to 95 percentile   0.806479   0.279300   2.887 0.003883 **
## income5. 96 to 100 percentile  1.332165   0.385238   3.458 0.000544 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  6
##
## Residual deviance: 1878.696 on 3276 degrees of freedom
##
## Log-likelihood: -939.3477 on 3276 degrees of freedom
##
## Number of iterations: 6
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##                   ideomoderate              ideoconservative
##                      2.9464500                     7.6224964
##                             age                  genderfemale
##                      0.9877453                     0.7642930
##                      raceblack                     raceasian
##                      0.2150799                     1.1779383
##           racenative american                   racehispanic
##                      0.9397273                     0.5889930
##  income2. 17 to 33 percentile  income3. 34 to 67 percentile
##                      1.8171795                     2.0237377
##  income4. 68 to 95 percentile income5. 96 to 100 percentile
##                      2.2400078                     3.7892396
```

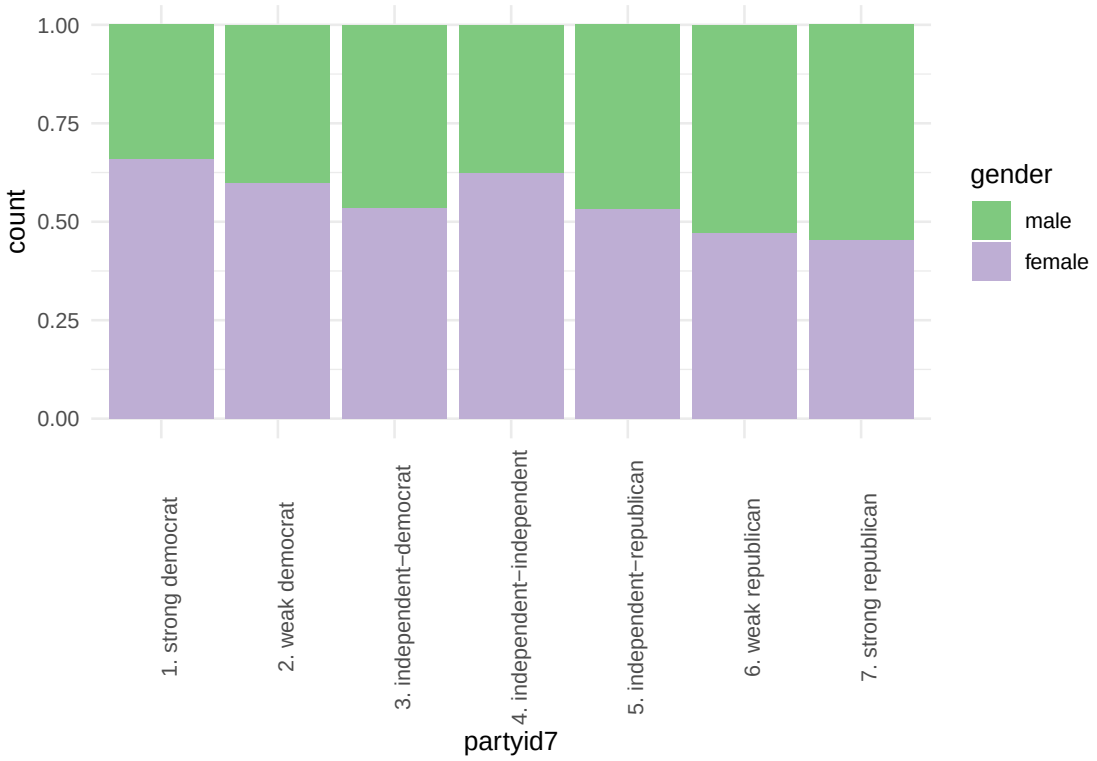
```
library(esquisse)
df <- data.frame(nes_data_comp$partyid7, nes_data_comp$ideo, nes_data_comp$age, nes_data_comp$gender, ne
colnames(df) <- paste(c("partyid7", "ideo","age","gender","race","income"))
nes_data_comp$partyid7 <- as.integer(as.character(substr(nes_data_comp$partyid7, 1,2)))

#EDA
#Party Identification vs. Race
ggplot(data = df) +
  aes(x = partyid7, fill = race) +
  geom_bar(position = "fill") +
  labs(title = "Party Identification in terms of Race",
    x = "Party Identification") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))
```




```
#Party Identification vs. Ideology
ggplot(data = df) +
  aes(x = partyid7, fill = ideo) +
  geom_bar(position = "fill") +
  scale_fill_brewer("Ideology",palette = "Set1") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))
```




```
#Party Identification vs. Gender
ggplot(data = df) +
  aes(x = partyid7, fill = gender) +
  geom_bar(position = "fill") +
  scale_fill_brewer(palette = "Accent") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))
```




```
fit <- polr(ordered(partyid7) ~ ideo + age + gender + race + income, data = nes_data_comp)
summary(fit)
```

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

```
## Call:
## polr(formula = ordered(partyid7) ~ ideo + age + gender + race +
##     income, data = nes_data_comp)
##
## Coefficients:
##                                 Value Std. Error t value
## ideomoderate                  1.08056   0.333883  3.2364
## ideoconservative              2.03112   0.183480 11.0700
## age                          -0.01233   0.005034 -2.4494
## genderfemale                 -0.26880   0.158573 -1.6951
## raceblack                    -1.53675   0.275463 -5.5788
## raceasian                     0.16375   0.556582  0.2942
## racenative american          -0.06217   0.373273 -0.1666
## racehispanic                 -0.52935   0.296836 -1.7833
## income2. 17 to 33 percentile  0.59729   0.286846  2.0823
## income3. 34 to 67 percentile  0.70495   0.269530  2.6155
## income4. 68 to 95 percentile  0.80649   0.280265  2.8776
## income5. 96 to 100 percentile 1.33217   0.388133  3.4323
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -0.5983  0.4038    -1.4817
## 2|3  0.2605  0.4001     0.6511
## 3|4  1.0614  0.4023     2.6381
```

```
## 4|5  1.4642  0.4063     3.6039
## 5|6  2.2751  0.4133     5.5049
## 6|7  3.2658  0.4239     7.7041
##
## Residual Deviance: 1878.695
## AIC: 1914.695
## (8 observations deleted due to missingness)
```

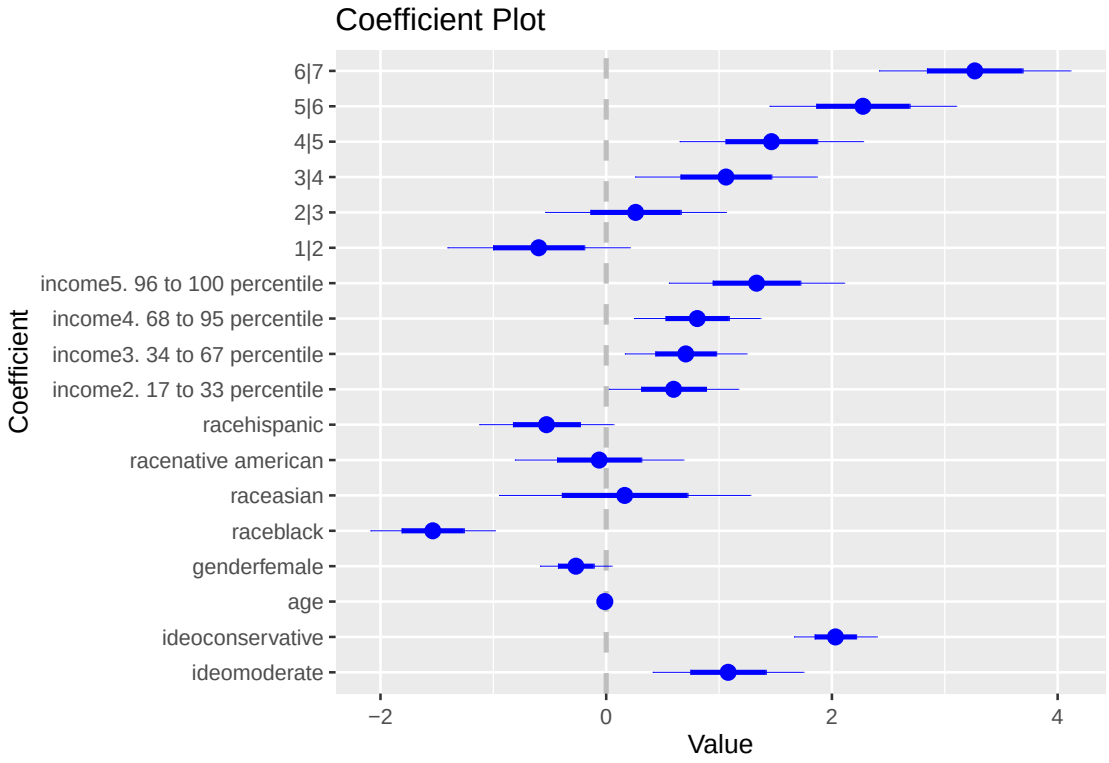
```
library(coefplot)
```

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

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

```
coefplot(fit)
```

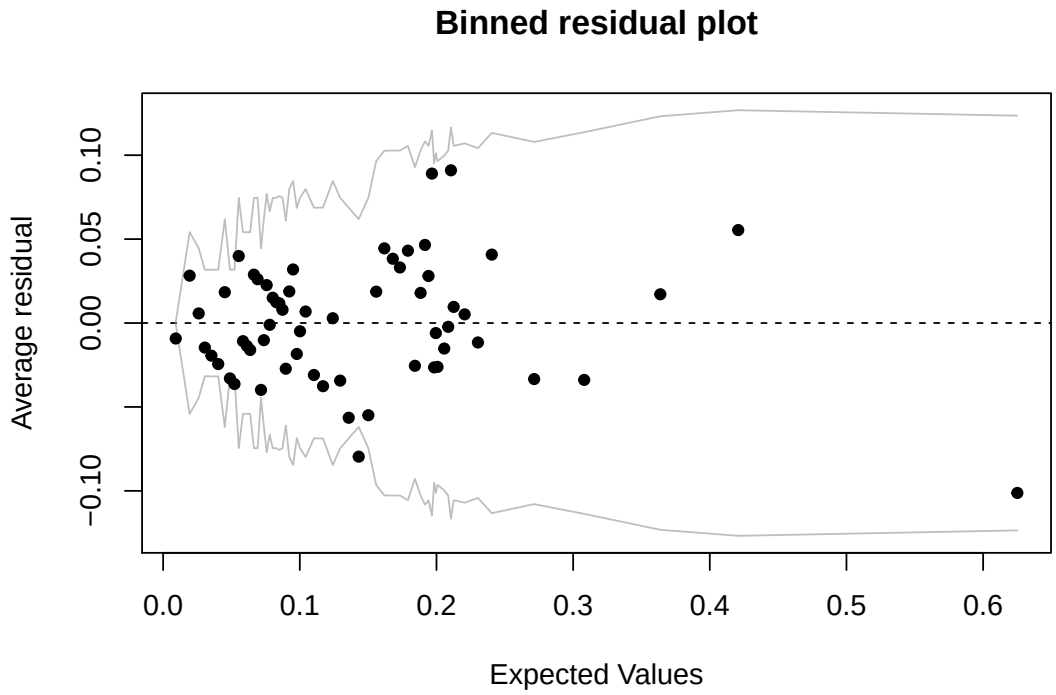
```
##
## Re-fitting to get Hessian
```



2. Explain the results from the fitted model. ***The age coefficient is close to zero and hence may not be very useful in explaining the party identification. From the plot, we can see that the coefficient estimates for gender, asian, american, and hispanic cross zero and may not be statistically significant. The coefficient for black suggests that the blacks are more likely to be supportive of Democrats than Republicans. Those with a high income level are likely to support Republicans more than Democrats. And, people with a conservative ideology are more likely to have a party identification of Republican than people with a liberal ideology.***

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

```
binnedplot(fittedvlm(lm1), resid(lm1, type = "response"))
```



The binned residual plot shows that the residuals are within the limits and the model is a good fit for the data.

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
```

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

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

```
lm2 <- multinom(prog ~ gender + race + ses + schtyp + read + write + math + science + 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(lm2)
```

```
## Call:
## 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 0.6066763
## 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.

```
pred_99 <- as.data.frame(cbind(id = hsb$id, fitted(lm2)))
pred_99 %>%
  filter(id == 99)
```

```
##   id  academic  general  vocation
## 1 99 0.5076752 0.375309 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)
```

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

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

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

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

```
confint(lm3)
```

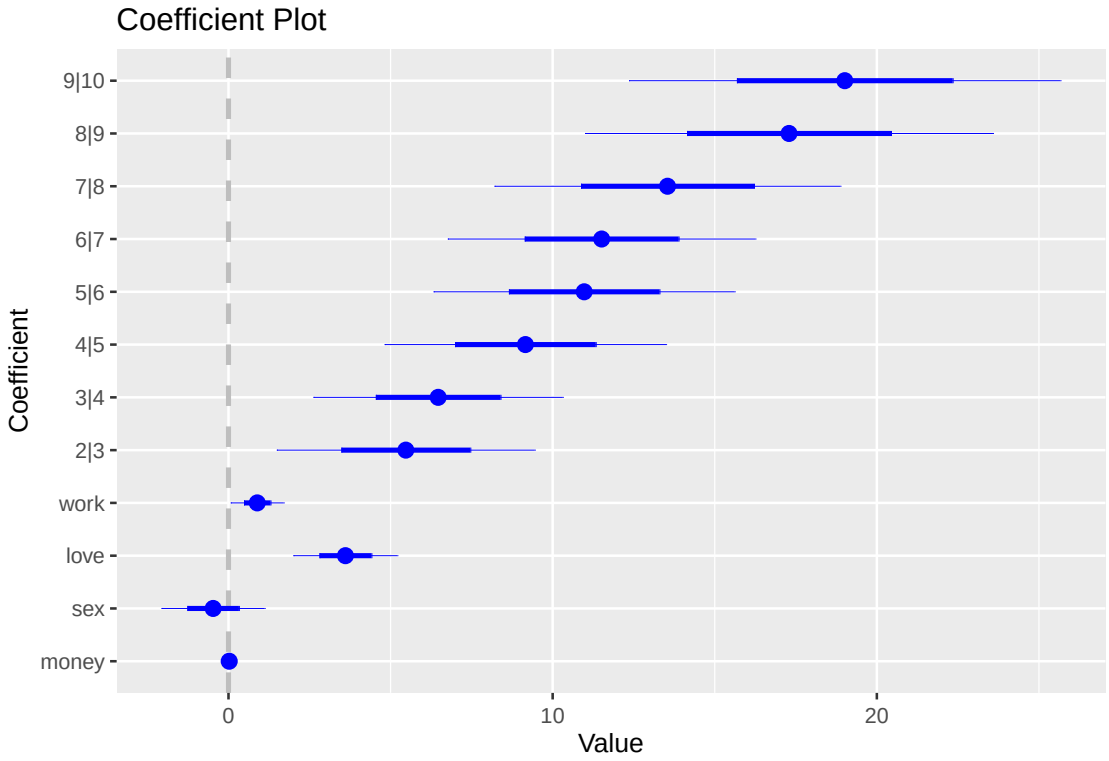
```
## Waiting for profiling to be done...
```

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

```
##              2.5 %     97.5 %
## money  0.002276811 0.04490097
## sex   -2.068912555 1.07918378
## love   2.168908595 5.37172931
## work   0.123787532 1.74622976
```

```
coefplot(lm3)
```

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



The lower bound of the confidence interval for the money coefficient is very close to zero and the confidence interval for the sex coefficient crosses zero, implying that these two coefficients are not statistically significant. The coefficient for love implies that people who have a deep feeling of belonging and caring are happier than those who are lonely.

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

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)
library(reshape)
```

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

## The following objects are masked from 'package:reshape2':
##
##     colsplit, melt, recast

## The following object is masked from 'package:data.table':
```

```
##
##     melt

## The following object is masked from 'package:Matrix':
##
##     expand
```

```
df <- data.frame(uncviet)
viet <- untable(df, num=df[,1])
df_viet <- within(viet, rm(y))

lm4 <- vglm(ordered(policy) ~ sex + year, data = df_viet, family = cumulative)
summary(lm4)
```

```
##
## Call:
## vglm(formula = ordered(policy) ~ sex + year, family = cumulative,
##     data = df_viet)
##
##
## Pearson residuals:
##                       Min      1Q  Median     3Q    Max
## logit(P[Y<=1]) -1.231 -0.4365 -0.2827 1.1754 3.7097
## logit(P[Y<=2]) -1.434 -0.8960 -0.2128 0.6081 2.4048
## logit(P[Y<=3]) -5.893  0.1478  0.2129 0.3992 0.6212
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -1.4504     0.1530  -9.479  < 2e-16 ***
## (Intercept):2  -0.2819     0.1253  -2.250  0.02444 *
## (Intercept):3   3.5492     0.2728  13.009  < 2e-16 ***
## sexMale:1       1.0198     0.1356   7.524 5.33e-14 ***
## sexMale:2       0.9419     0.1009   9.335  < 2e-16 ***
## sexMale:3      -0.4968     0.1832  -2.712  0.00669 **
## yearGrad:1     -1.2172     0.1306  -9.321  < 2e-16 ***
## yearGrad:2     -1.1100     0.1145  -9.694  < 2e-16 ***
## yearGrad:3     -1.4657     0.2349  -6.240 4.37e-10 ***
## yearJunior:1   -0.4137     0.1326  -3.119  0.00181 **
## yearJunior:2   -0.3913     0.1244  -3.145  0.00166 **
## yearJunior:3   -0.4022     0.2782  -1.446  0.14821
## yearSenior:1   -0.3886     0.1321  -2.942  0.00327 **
## yearSenior:2   -0.5890     0.1253  -4.703 2.57e-06 ***
## yearSenior:3   -0.8561     0.2607  -3.283  0.00103 **
## yearSoph:1     -0.1771     0.1334  -1.328  0.18430
## yearSoph:2     -0.1150     0.1312  -0.877  0.38059
## yearSoph:3     -0.1438     0.3004  -0.479  0.63210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
## Residual deviance: 7667.101 on 9423 degrees of freedom
```

```
##
## Log-likelihood: -3833.551 on 9423 degrees of freedom
##
## Number of iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):3'
##
## Exponentiated coefficients:
##    sexMale:1    sexMale:2    sexMale:3   yearGrad:1   yearGrad:2
##    2.7727402    2.5647309    0.6085033    0.2960568    0.3295518
##   yearGrad:3 yearJunior:1 yearJunior:2 yearJunior:3 yearSenior:1
##    0.2309188    0.6612297    0.6761688    0.6688370    0.6780385
## yearSenior:2 yearSenior:3   yearSoph:1   yearSoph:2   yearSoph:3
##    0.5548585    0.4248218    0.8376635    0.8913502    0.8660240
```

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
```

```
##    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
## 7    12 normal 46.0
## 8     4 normal 51.5
## 9     0   mild  5.8
## 10    2   mild 15.0
## 11    6   mild 21.5
## 12    5   mild 27.5
## 13   10   mild 33.5
## 14    7   mild 39.5
## 15    6   mild 46.0
## 16    2   mild 51.5
## 17    0 severe  5.8
## 18    1 severe 15.0
## 19    3 severe 21.5
## 20    8 severe 27.5
## 21    9 severe 33.5
## 22    8 severe 39.5
## 23   10 severe 46.0
## 24    5 severe 51.5
```

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.

```
lm5 <- vglm(status ~ year, data = pneumo, family = multinomial)
summary(lm5)
```

```
##
## Call:
## vglm(formula = status ~ year, family = multinomial, data = pneumo)
##
##
## Pearson residuals:
##                     Min 1Q Median    3Q   Max
## log(mu[,1]/mu[,3])  -1 -1 -0.366 1.366 1.366
## log(mu[,2]/mu[,3])  -1 -1 -0.366 1.366 1.366
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -1.305e-15  1.143e+00       0        1
## (Intercept):2 -1.831e-15  1.143e+00       0        1
## year:1         4.075e-17  3.420e-02       0        1
## year:2         6.835e-17  3.420e-02       0        1
##
## Number of linear predictors:  2
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 52.7334 on 44 degrees of freedom
##
## Log-likelihood: -26.3667 on 44 degrees of freedom
##
## Number of iterations: 1
##
## No Hauck-Donner effect found in any of the estimates
##
## Reference group is level  3  of the response
```

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

```
lm6 <- vglm(ordered(status) ~ year, data = pneumo, family = cumulative)
summary(lm6)
```

```
##
## Call:
## vglm(formula = ordered(status) ~ year, family = cumulative, data = pneumo)
##
##
## Pearson residuals:
##                       Min     1Q Median    3Q   Max
## logit(P[Y<=1]) -1.000 -1.000 -0.366 1.366 1.366
## logit(P[Y<=2]) -1.366 -1.366  0.366 1.000 1.000
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -6.931e-01  9.894e-01  -0.701    0.484
## (Intercept):2  6.931e-01  9.894e-01   0.701    0.484
## year:1         4.814e-18  2.962e-02   0.000    1.000
```

```
## year:2        -1.315e-17  2.962e-02   0.000    1.000
##
## Number of linear predictors:  2
##
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
##
## Residual deviance: 52.7334 on 44 degrees of freedom
##
## Log-likelihood: -26.3667 on 44 degrees of freedom
##
## Number of iterations: 1
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
## year:1 year:2
##      1      1
```

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.

4. Compare the three analyses.

(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

name	description
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.