MA678 Homework 6

Yuchen Huang

11/8/2022

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.

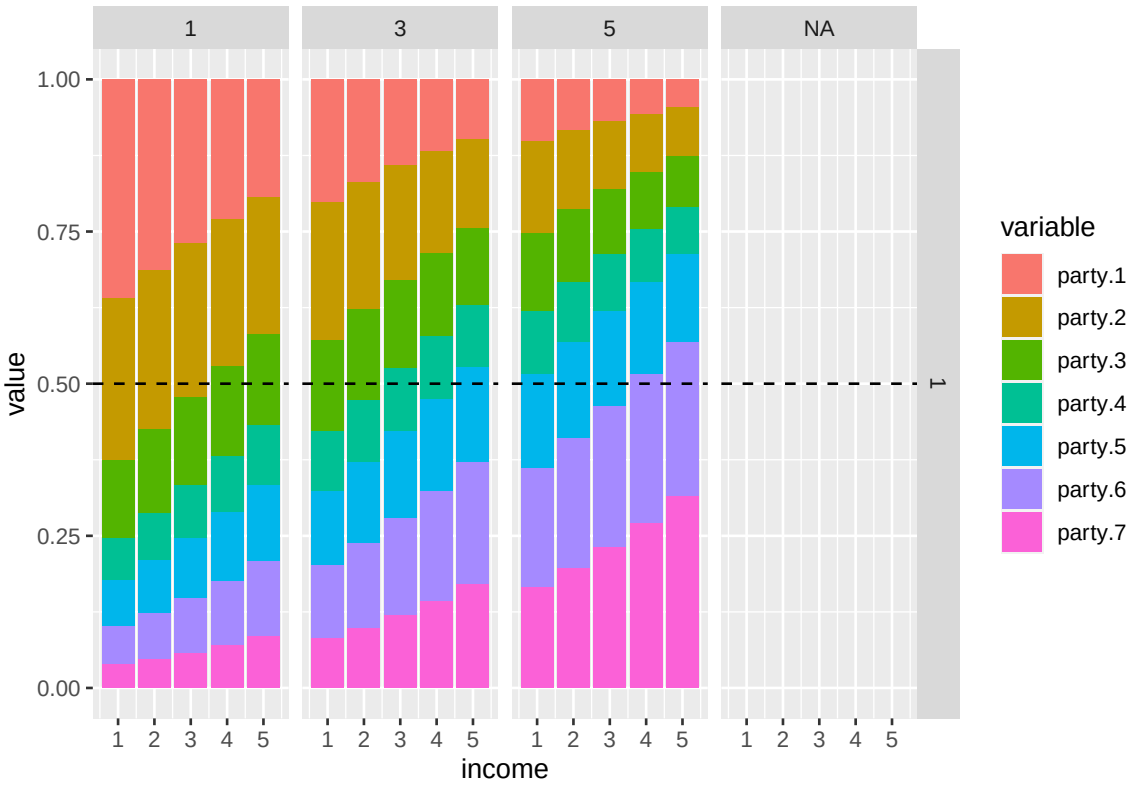
1. Summarize the parameter estimates numerically and also graphically.

```
display(fit_polr)
```

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

```
## polr(formula = ordered(partyid7) ~ ideo + female + white + income,
##     data = NES)
##        coef.est coef.se
## ideo    0.40     0.01
## female -0.10     0.03
## white   0.95     0.04
## income  0.21     0.01
## 1|2     0.98     0.07
## 2|3     2.07     0.07
## 3|4     2.68     0.07
## 4|5     3.10     0.07
## 5|6     3.74     0.07
## 6|7     4.79     0.08
## ---
## n = 12476, k = 10 (including 6 intercepts)
## residual deviance = 44943.0, null deviance is not computed by polr
```

```
predx <- expand.grid(income = unique(NES$income),
                     white = 1,female=0,ideo = unique(NES$ideo))
predy <- predict(fit_polr, newdata = predx,type="prob")
resd <- data.frame(predx[, c("income", "ideo","white")], party = predy)
ggplot(melt(resd,id.var = c("income", "ideo","white")))+
    geom_bar(position = "fill", stat = "identity") +
    aes(x=income, y = value, fill = variable) +
    facet_grid(white~ideo) +
    geom_hline(yintercept=0.5,lty=2)
```



2. Explain the results from the fitted model.

The coefficients are ideo, female, white, income, and all of them are statistically significant There are 6 Thresholds, the model estimates boundaries between the ordered categories of the dependent variable partyid7. For example, the threshold 1|2 at 0.98 is the estimated point in the latent variable where a respondent is equally likely to be in category 1 or 2. The distances between these thresholds indicate how much the latent variable needs to change to move from one category to the next. For instance, moving from category 5|6 to 6|7 requires a larger change than from 1|2 to 2|3. Overall, the model suggests that ideological orientation and identifying as white are positively associated with higher political party identification numbers, while being female is negatively associated. Income is also positively associated, suggesting those with higher income may identify with higher political party identification numbers on the provided scale. The significant coefficients and large number of observations suggest that these findings are likely robust.

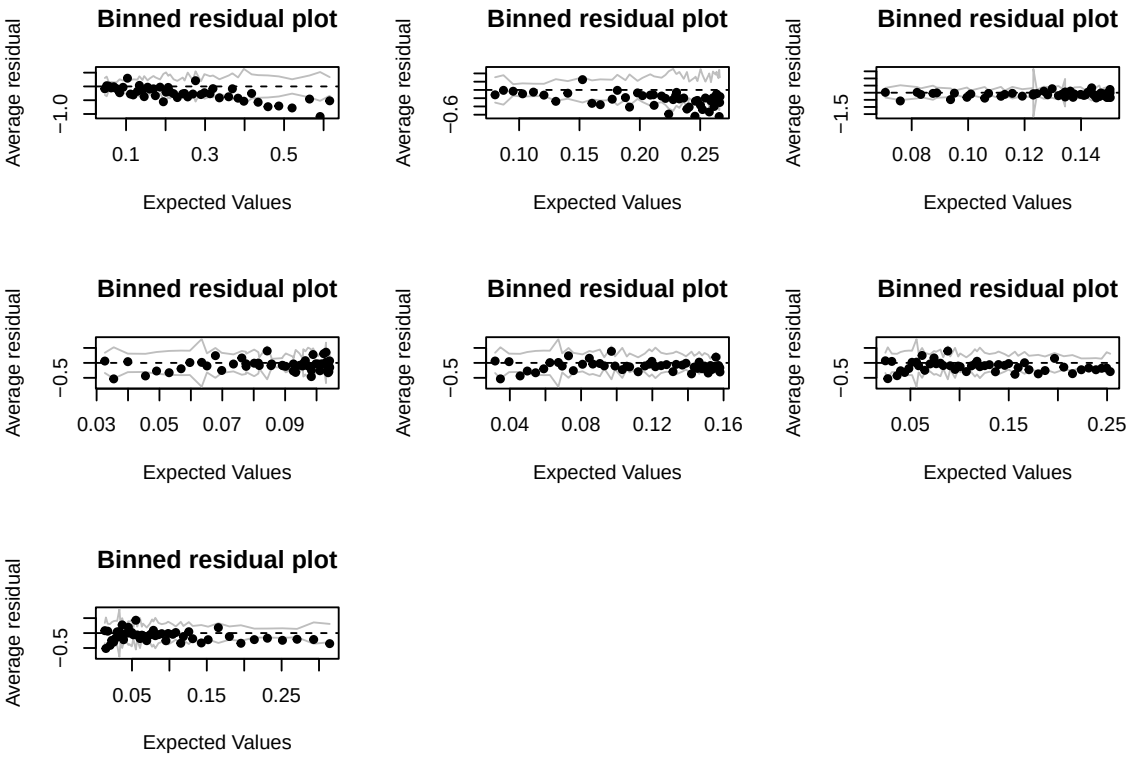
3. Use a binned residual plot to assess the fit of the model.

C7 = party - fitted(fit_polr)[,7]-3.5)

par(mfrow = c(3,3))

```
party <- NES |>
    dplyr::select(partyid7)|> na.omit()
res <- data.frame(C1 = party - fitted(fit_polr)[,1]-3.5,
                  C2 = party - fitted(fit_polr)[,2]-3.5,
                  C3 = party - fitted(fit_polr)[,3]-3.5,
                  C4 = party - fitted(fit_polr)[,4]-3.5,
                  C5 = party - fitted(fit_polr)[,5]-3.5,
                  C6 = party - fitted(fit_polr)[,6]-3.5,
                  C7 = party - fitted(fit_polr)[,7]-3.5)
par(mfrow = c(3,3))
binnedplot(fitted(fit_polr)[,1], res[,1])
binnedplot(fitted(fit_polr)[,2], res[,2])
```

```
binnedplot(fitted(fit_polr)[,3], res[,3])
binnedplot(fitted(fit_polr)[,4], res[,4])
binnedplot(fitted(fit_polr)[,5], res[,5])
binnedplot(fitted(fit_polr)[,6], res[,6])
binnedplot(fitted(fit_polr)[,7], res[,7])
```



(Optional) Choice models

Using the individual-level survey data from the election example described in Section 10.9 (data available in the folder NES),

1. Fit a logistic regression model for the choice of supporting Democrats or Republicans. Then interpret the output from this regression in terms of a utility/choice model.

```
fit_glm <- glm(vote ~ ideo+female+white+income, data = NES)
summary(fit_glm)
```

```
##
## Call:
## glm(formula = vote ~ ideo + female + white + income, data = NES)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.458199   0.016414  88.839  < 2e-16 ***
## ideo        0.008480   0.002422   3.501 0.000466 ***
## female      0.014784   0.008390   1.762 0.078092 .
## white       0.050725   0.010170   4.988 6.19e-07 ***
## income      0.055874   0.003833  14.578  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 0.2046062)
##
##     Null deviance: 2504.0  on 11950  degrees of freedom
## Residual deviance: 2444.2  on 11946  degrees of freedom
##   (29547 observations deleted due to missingness)
## AIC: 14960
##
## Number of Fisher Scoring iterations: 2
```

2. Repeat the previous exercise but now with three options: Democrat, no opinion, Republican. That is, fit an ordered logit model and then express it as a utility/choice mode

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 hemaglutinin 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.

1. Using a chi-square 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(contingency[,-1])
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency[, -1]
## X-squared = 17.648, df = 3, p-value = 0.0005199
```

```
contingency_long <- contingency |>
    pivot_longer(cols = c(small,moderate,large,Total),
                 names_to = "category",
                 values_to = "frequency")

group1 <- glm(frequency ~ category+treatment, family = poisson(), contingency_long)
group2 <- glm(frequency ~ category*treatment, family = poisson(), contingency_long)
AIC(group1);AIC(group2)
```

```
## [1] 64.64834
```

```
## [1] 52.00582
```

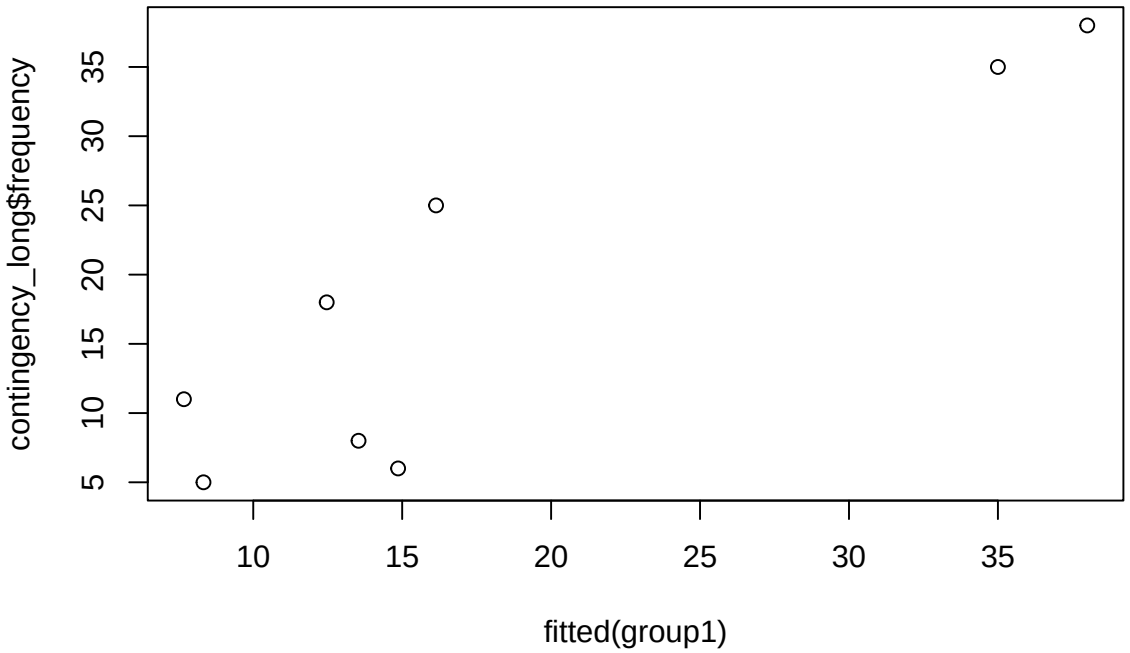
```
anova(group1, group2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: frequency ~ category + treatment
## Model 2: frequency ~ category * treatment
```

```
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         3     18.642
## 2         0      0.000  3   18.642 0.0003241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. For the model corresponding to the hypothesis of homogeneity 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.

```
plot(fitted(group1), contingency_long$frequency)
```




```
resid(group1, type = "pearson")
```

```
##             1             2             3             4             5
##  2.206329e+00 -1.504324e+00 -1.153435e+00  1.152653e-15 -2.298942e+00
##             6             7             8
##  1.567470e+00  1.201852e+00  1.201036e-15
```

```
resid(group1, type = "deviance")
```

```
##             1             2             3             4             5
##  2.040115e+00 -1.629720e+00 -1.246900e+00  5.161914e-08 -2.615460e+00
##             6             7             8
##  1.468817e+00  1.127679e+00  3.650024e-08
```

3. Re-analyze these data using ordered logit model (use `polr`) to estimate the cut-points of a latent continuous response variable 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.

```
polr(ordered(category) ~ treatment+frequency, contingency_long)
```

```
## Call:
## polr(formula = ordered(category) ~ treatment + frequency, data = contingency_long)
##
## Coefficients:
## treatmentvaccine        frequency
```

```
##      -0.07140478       0.22685710
##
## Intercepts:
## large|moderate moderate|small    small|Total
##       1.809619       3.626806       6.699904
##
## Residual Deviance: 14.05397
## AIC: 24.05397
```

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).

```
fit_multi <- multinom(prog ~ gender+race+ses+schtyp+read+write+math+science+socst, data = hsb, trace = F
summary(fit_multi)
```

```
## Call:
## multinom(formula = prog ~ gender + race + ses + schtyp + read +
##     write + math + science + socst, data = hsb, trace = FALSE)
##
## Coefficients:
##          (Intercept)  gendermale raceasian racehispanic racewhite     seslow
## general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156 1.09864111
## vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881 0.04747323
##          sesmiddle schtyppublic        read       write       math    science
## general  0.7029621    0.5845405 -0.04418353 -0.03627381 -0.1092888 0.10193746
## vocation 1.1815808    2.0553336 -0.03481202 -0.03166001 -0.1139877 0.05229938
##                socst
## general  -0.01976995
## vocation -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    science
## general  0.5045938    0.5642925 0.03103707 0.03381324 0.03522441 0.03274038
## vocation 0.5700833    0.8348229 0.03422409 0.03585729 0.03885131 0.03424763
##              socst
## general  0.02712589
## vocation 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.

```
id99 <- hsb[hsb$id == 99,]
predict(fit_multi, newdata = id99, type = "probs")
```

```
##  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)
```

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

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

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

```
## polr(formula = factor(happy) ~ money + love + sex + work, data = happy)
##       coef.est coef.se
## money  0.02     0.01
## love   3.61     0.80
## sex   -0.47     0.79
## work   0.89     0.41
## 2|3    5.47     1.99
## 3|4    6.47     1.92
## 4|5    9.16     2.17
## 5|6   10.97     2.32
## 6|7   11.51     2.37
## 7|8   13.54     2.67
## 8|9   17.29     3.15
## 9|10  19.01     3.33
## ---
## n = 39, k = 12 (including 8 intercepts)
## residual deviance = 94.9, null deviance is not computed by polr
```

```
happy <- happy |>
    mutate(Money = scale(money, center = F),
           Work = work-3,
           Love = love-2)
fit_happyc <- polr(factor(happy) ~ Money+Love+sex+Work, data = happy)
display(fit_happyc)
```

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

```
## polr(formula = factor(happy) ~ Money + Love + sex + Work, data = happy)
##       coef.est coef.se
## Money  1.62     0.77
## Love   3.61     0.80
## sex   -0.47     0.79
## Work   0.89     0.41
## 2|3   -4.41     1.57
```

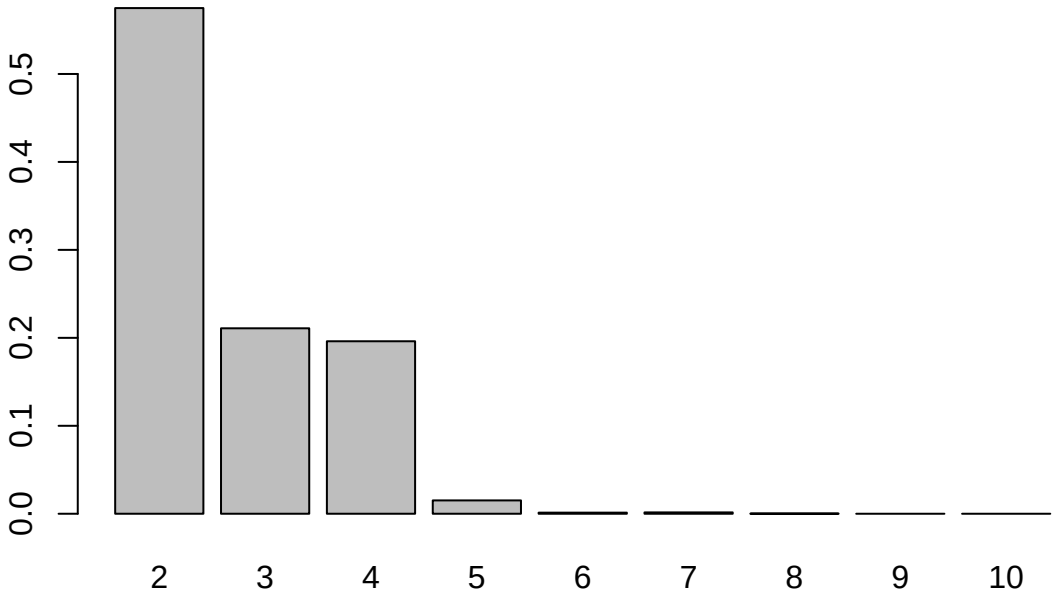
```
## 3|4   -3.41     1.37
## 4|5   -0.72     0.95
## 5|6    1.09     0.84
## 6|7    1.63     0.85
## 7|8    3.67     1.02
## 8|9    7.41     1.49
## 9|10   9.13     1.81
## ---
## n = 39, k = 12 (including 8 intercepts)
## residual deviance = 94.9, null deviance is not computed by polr
```

2. Interpret the parameters of your chosen model

By observing the coefficients, we can say that money, love, work can improve the happiness. The p-value of sex is not statistically significant, and the sign of sex coefficient doesn't make sense, so we can remove this predictor.

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

```
happy_pred <- predict(fit_happy, newdata = list(money = 30, sex = 0, work = 1, love = 1), type = "prob")
barplot(happy_pred)
```



Newspaper survey on Vietnam War

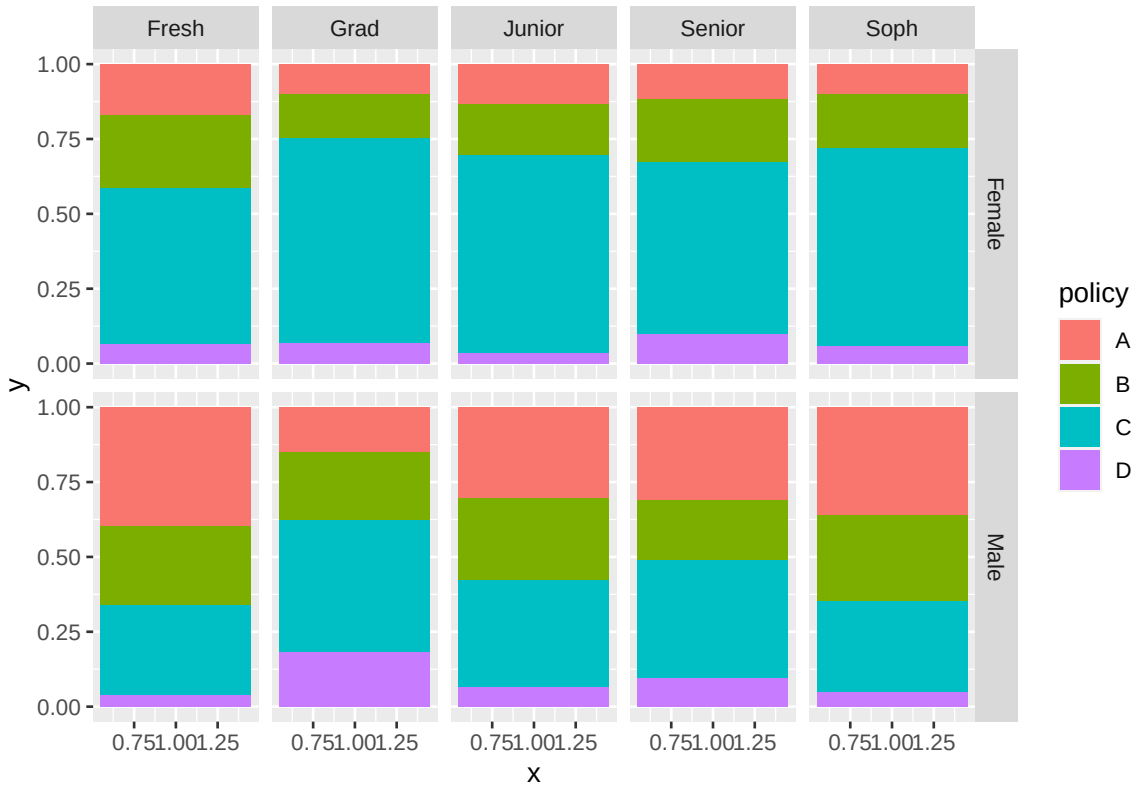
A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset `uncviet`. Treat the opinion as the response and the sex and year as predictors. Build a proportional odds model, giving an interpretation to the estimates.

```
data(uncviet)

uncviet_wider <- uncviet |>
    pivot_wider(names_from = policy,
                values_from = y)

fit_unc <- vglm(cbind(A,B,C,D) ~ sex+year, family = cumulative(parallel = T), uncviet_wider)
```

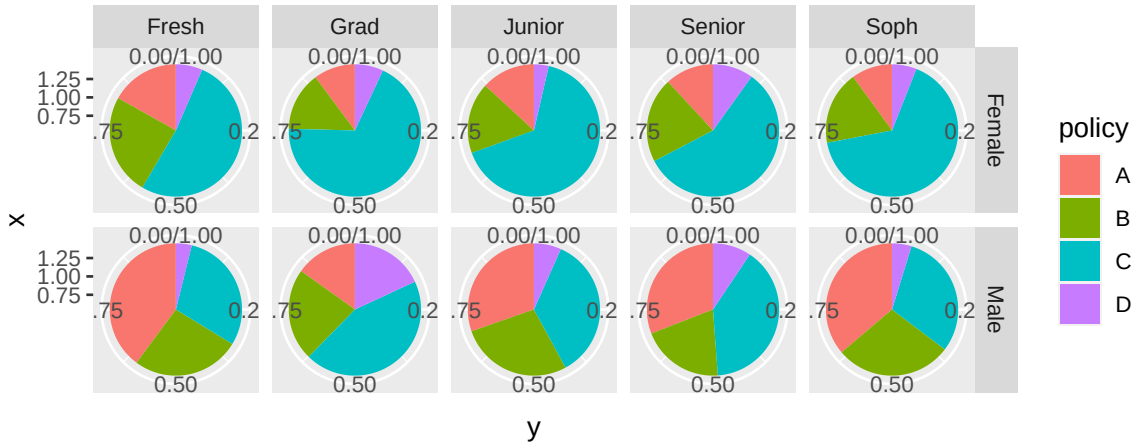
```r
summary(fit_unc)
```

```
##
## Call:
## vglm(formula = cbind(A, B, C, D) ~ sex + year, family = cumulative(parallel = T),
##     data = uncviet_wider)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -1.10979    0.11220  -9.891  < 2e-16 ***
## (Intercept):2 -0.01305    0.11069  -0.118 0.906167
## (Intercept):3  2.44170    0.12118  20.149  < 2e-16 ***
## sexMale        0.64703    0.08720   7.420 1.17e-13 ***
## yearGrad      -1.17699    0.10238 -11.496  < 2e-16 ***
## yearJunior    -0.39642    0.11054  -3.586 0.000335 ***
## yearSenior    -0.54439    0.11165  -4.876 1.08e-06 ***
## yearSoph      -0.13150    0.11532  -1.140 0.254141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3])
##
## Residual deviance: 112.0238 on 22 degrees of freedom
##
## Log-likelihood: -131.8698 on 22 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##    sexMale   yearGrad yearJunior yearSenior   yearSoph
##  1.9098677  0.3082047  0.6727233  0.5801928  0.8767760
```

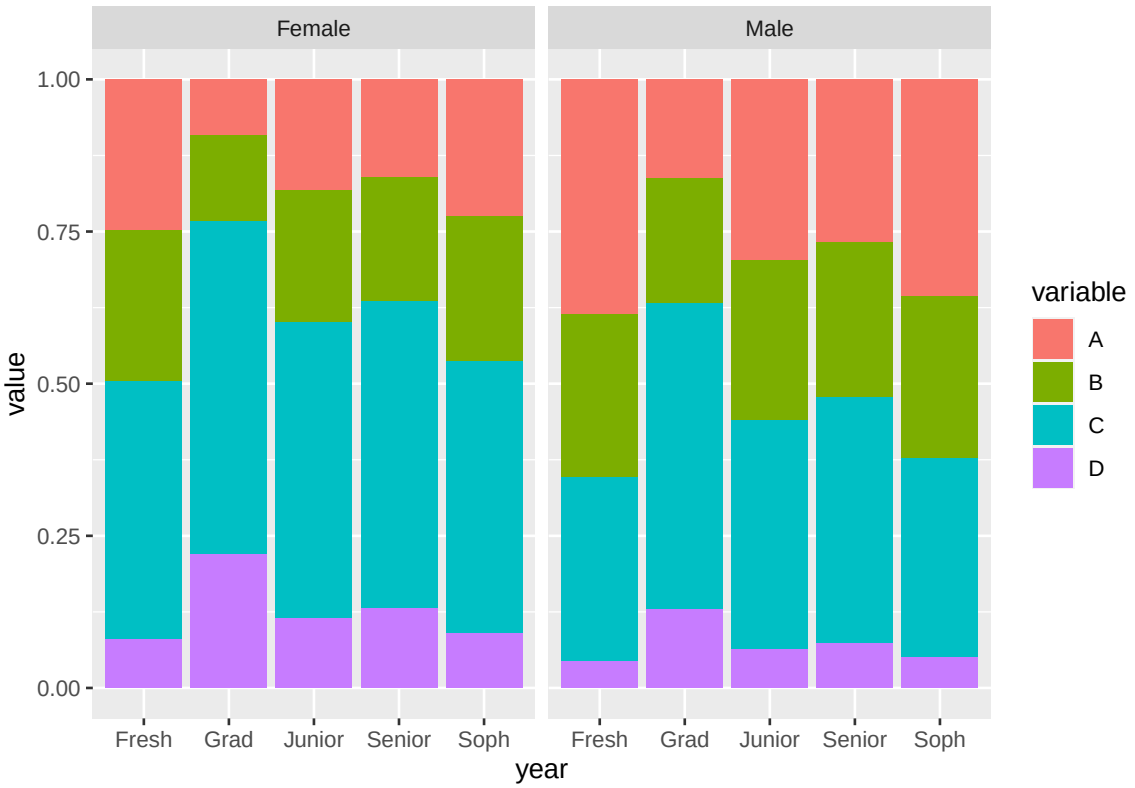
```r
uncviet |>
    ggplot()+
    geom_bar(position = "fill", stat = "identity")+
    aes(x = 1, y = y, fill = policy) +
    facet_grid(sex~year)
```




```
uncviet |>
    ggplot() +
    geom_bar(position = "fill", stat = "identity")+
    aes(x = 1, y = y, fill = policy) +
    facet_grid(sex~year) +
    coord_polar("y", start = 0)
```




```
predx <- expand.grid(sex = levels(uncviet$sex), year = levels(uncviet$year))
predy <- (predict(fit_unc, newdata = predx, type = "response"))
ggplot(melt(data.frame(predx, predy), id.vars = c("sex", "year")))+
    geom_bar(stat = "identity") +
    aes(x = year, y = value, fill = variable)+
    facet_grid(~sex)
```



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.

```
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.

lines(years, predict[,1], col = “red”)

lines(years, predict[,2], col = “blue”)

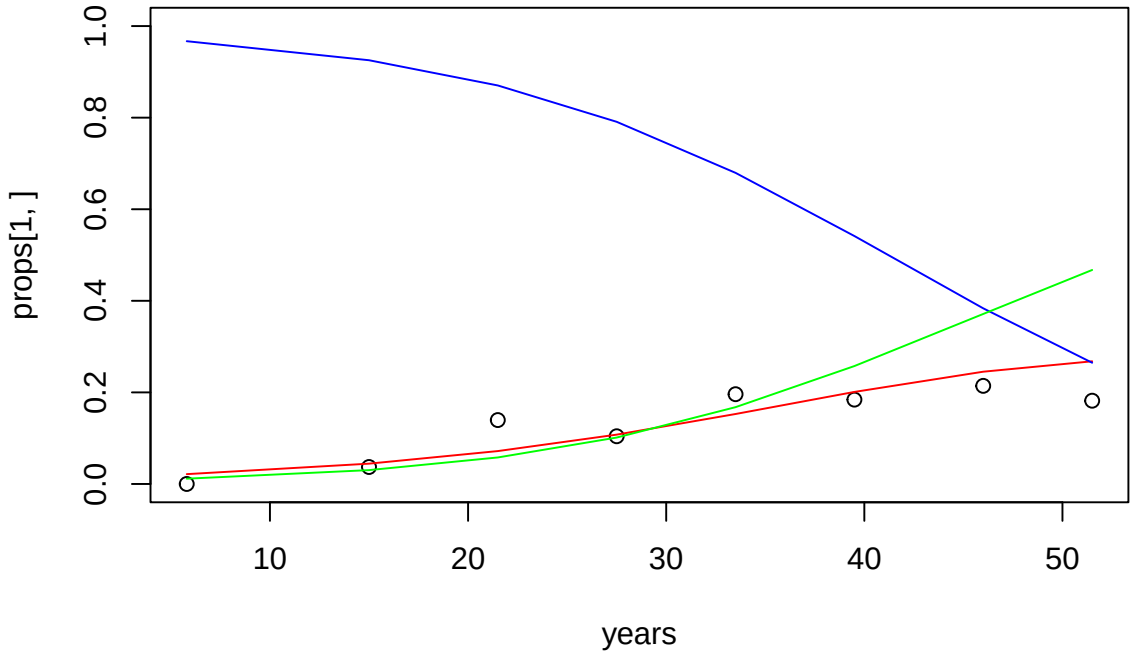
lines(years, predict[,3], col = “green”)

```
counts <- xtabs(Freq ~ status+year, pneumo)
years <- pneumo$year[1:8]
fit <- multinom(t(counts) ~ years, trace = F)
summary(fit)
```

```
## Call:
## multinom(formula = t(counts) ~ years, trace = F)
##
## Coefficients:
##        (Intercept)       years
## normal   4.2916723 -0.08356506
## severe  -0.7681706  0.02572027
##
```

```
## Std. Errors:
##        (Intercept)      years
## normal   0.5214110 0.01528044
## severe   0.7377192 0.01976662
##
## Residual Deviance: 417.4496
## AIC: 425.4496
```

```
predict <- predict(fit, newdata = list(year = years), type = "probs")
props <- prop.table(counts,2)
plot(years, props[1,], ylim = c(0,1))
lines(years, predict[,1], col = "red")
lines(years, predict[,2], col = "blue")
lines(years, predict[,3], col = "green")
```



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

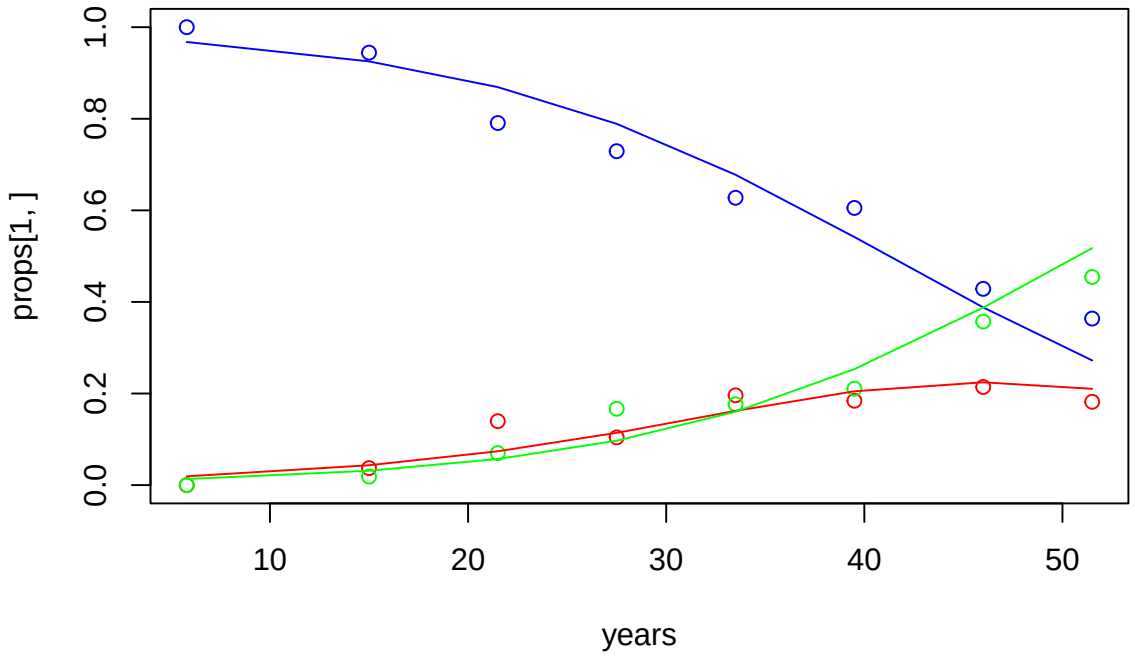
```
pneumo2 <- data.frame(status = rep(pneumo$status, pneumo$Freq), year = rep(pneumo$year, pneumo$Freq))
pneumo2$status <- ordered(pneumo2$status, levels=c("normal", "mild", "severe"))
library(MASS)
omod <- polr(status ~ year, pneumo2)
summary(omod)
```

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

```
## Call:
## polr(formula = status ~ year, data = pneumo2)
##
## Coefficients:
##       Value Std. Error t value
## year 0.0959    0.01194   8.034
##
## Intercepts:
##             Value   Std. Error t value
```

```
## normal|mild  3.9558  0.4097     9.6558
## mild|severe  4.8690  0.4411    11.0383
##
## Residual Deviance: 416.9188
## AIC: 422.9188
```

```
plot(years, props[1,], col="red", ylim=c(0,1))
points(years, props[2,], col="blue")
points(years, props[3,], col="green")
fitted <- predict(omod, newdata=list(year=years), type="probs")
lines(years, fitted[,1], col="blue")
lines(years, fitted[,2], col="red")
lines(years, fitted[,3], col="green")
```



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<- data.frame(normal=pneumo[pneumo$status == "normal","Freq"], disease=pneumo[pneumo$status == "
binmodw <- glm(cbind(disease,normal) ~ year, data=pneumo3,family=binomial)
binmodd <- glm(cbind(severe, mild) ~ year, data=pneumo3, family = binomial)

predict(binmodw,data=pneumo3,type="response")
```

```
##         1         2         3         4         5         6         7
## 0.03204667 0.07430865 0.13049793 0.21099340 0.32271286 0.45916195 0.61349640
##         8
## 0.72938688
```

4. Compare the three analyses.

```
predict(fit, newdata=list(years=25), type="probs")
```

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

```r
predict(omod, newdata=list(year=25), type="probs")
```

```
##     normal       mild     severe
## 0.82610096 0.09601474 0.07788430
```