MA678 homework 06

Multinomial Regression

Zara Waheed

October 10, 2021

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.

```
summary(fit_nes)
```

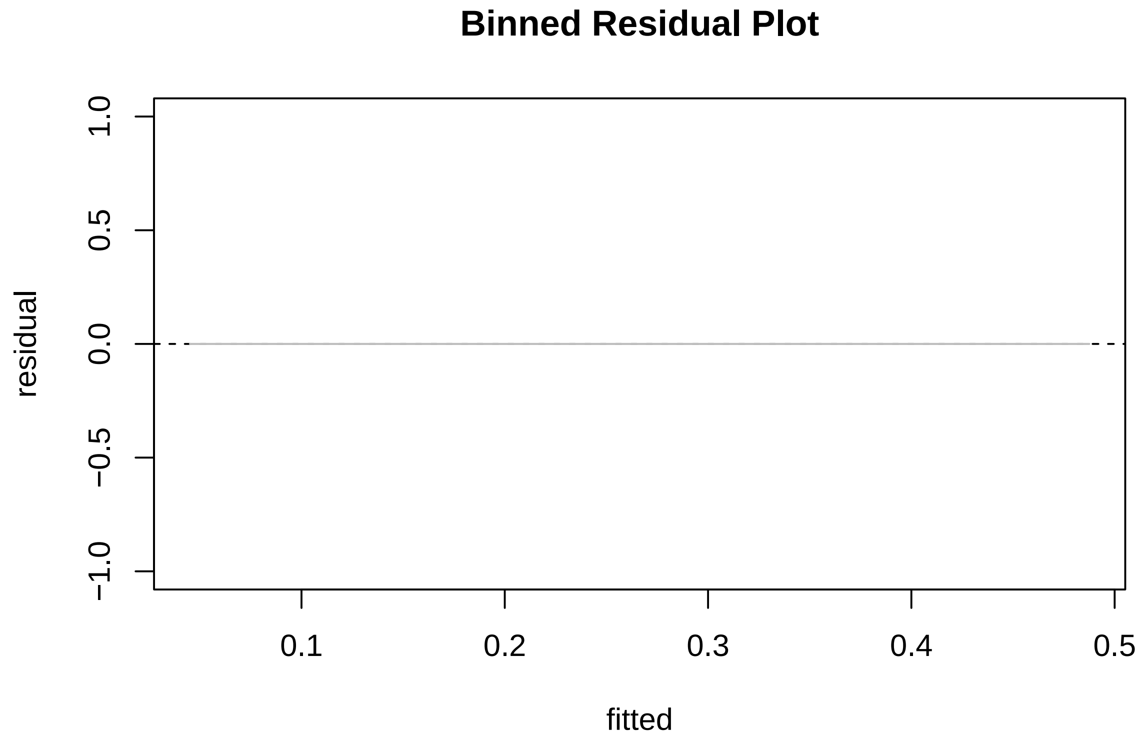
```
## Call:
## polr(formula = factor(str_partyid) ~ ideo + age + female + race +
##     educ1, data = nes, Hess = TRUE)
##
## Coefficients:
##           Value Std. Error t value
## ideo   0.001563   0.009613  0.1626
## age    0.016672   0.001013 16.4526
## female 0.114541   0.033091  3.4614
## race   0.045803   0.015549  2.9458
## educ1  0.088480   0.018779  4.7116
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -1.2345   0.0941   -13.1243
## 2|3   0.4157   0.0917     4.5353
## 3|4   1.8668   0.0932    20.0402
##
## Residual Deviance: 31437.69
## AIC: 31453.69
## (22627 observations deleted due to missingness)
```

2. Explain the results from the fitted model.

The results show that a person has a positive increase in party ID if they are older, female, more educated and have a higher value for race and ideo.

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

```
fitted = fitted(fit_nes)
resid = resid(fit_nes)
binnedplot(fitted, resid, xlab="fitted", ylab="residual",
     main="Binned Residual Plot")
```



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.

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(responses)
```

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

```
fit_responses <- stan_glm(Freq ~ Var1 + Var2, data = responses_mod, refresh = 0)

summary(fit_responses)
```

```
##
## Model Info:
```

```
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      Freq ~ Var1 + Var2
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 6
##  predictors:   4
##
## Estimates:
##               mean   sd    10%   50%   90%
## (Intercept)   15.4    9.6   4.3  15.4  26.6
## Var1vaccine   -0.9    9.9 -12.3  -0.8  10.7
## Var2moderate  -1.8   11.1 -15.0  -2.1  11.8
## Var2large     -6.6   12.1 -20.1  -6.8   7.6
## sigma         11.7    4.9   6.7  10.6  18.0
##
## Fit Diagnostics:
##           mean   sd   10%   50%   90%
## mean_PPD 12.0    7.0  3.8  12.0  20.0
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.2  1.0  2540
## Var1vaccine   0.2  1.0  2750
## Var2moderate  0.2  1.0  2400
## Var2large     0.3  1.0  2321
## sigma         0.2  1.0  1011
## mean_PPD      0.1  1.0  3074
## log-posterior 0.1  1.0   751
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

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)
?hsb
```

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

```
fit_hsb <- multinom(factor(prog) ~ gender + race + read + write + math + science + socst , data = hsb, h
```

```
## # weights:  33 (20 variable)
## initial  value 219.722458
## iter  10 value 181.360808
## iter  20 value 162.567073
## final  value 162.546157
## converged
```

```
summary(fit_hsb)
```

```
## Call:
## multinom(formula = factor(prog) ~ gender + race + read + write +
##     math + science + socst, data = hsb, hess = TRUE)
##
## Coefficients:
##          (Intercept) gendermale  raceasian racehispanic   racewhite        read
## general     5.409757 -0.1803008  1.0011763  -0.64382594 -0.02092716 -0.04400951
## vocation    9.713345 -0.3481200 -0.2585968   0.06412298  0.39775210 -0.03075144
##                write       math    science       socst
## general  -0.03540744 -0.1057670 0.09809156 -0.02789520
## vocation -0.04902747 -0.1150006 0.05697860 -0.06813879
##
## Std. Errors:
##          (Intercept) gendermale raceasian racehispanic racewhite       read
## general     1.609632  0.4394142  1.017025    0.8753965 0.6972532 0.03012985
## vocation    1.740986  0.4721241  1.377469    0.8096209 0.7072538 0.03215976
##               write       math    science      socst
## general  0.03292687 0.03433111 0.03193672 0.02630079
## vocation 0.03362717 0.03711884 0.03282764 0.02664999
##
## Residual Deviance: 325.0923
## AIC: 365.0923
```

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

```
hsb_id99 <- hsb %>%
  filter(id == 99)

predict(fit_hsb, newdata = hsb_id99, type = "probs")
```

```
##  academic   general  vocation
## 0.4115755 0.4462236 0.1422009
```

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.

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

2. Interpret the parameters of your chosen model.

```
summary(fit_happy)
```

```
## Call:
## polr(formula = factor(happy) ~ money + sex + love + work, data = happy,
##     Hess = TRUE)
##
## Coefficients:
##          Value Std. Error t value
## money  0.02246    0.01066  2.1064
## sex   -0.47344    0.79498 -0.5955
## love   3.60765    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     5.4972
## 9|10 19.0112  3.3270     5.7142
##
## Residual Deviance: 94.86029
## AIC: 118.8603
```

According to the model, an increase in the indicators for money, love and work have a positive effect on happiness. An active sex life seems to have a negative effect on happiness.

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

```
# since person 37 seems to fit the criteria best

person37 <- happy %>% filter(money == 31)

predict(fit_happy, newdata = person37, type = "probs")
```

```
##            2            3            4            5            6            7
## 0.0024955743 0.0042427121 0.0841815471 0.2891954772 0.1323198481 0.3766788405
##            8            9           10
## 0.1079545065 0.0024054581 0.0005260361
```

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

fit_uncviet <- polr(policy ~ sex + year, data = uncviet)
```

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
```

```
## Help on topic 'pneumo' was found in the following packages:
##
##   Package               Library
##   VGAM                  /Library/Frameworks/R.framework/Versions/4.1/Resources/library
##   faraway               /Library/Frameworks/R.framework/Versions/4.1/Resources/library
##
##
## Using the first match ...
```

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.

```
fit_pneumo1 <- multinom(status ~ year, data = pneumo, Hess = TRUE)
```

```
## # weights:  9 (4 variable)
## initial  value 26.366695
## final  value 26.366695
## converged
```

```
# Make prediction

pneumo_25 <- data.frame(year=25)
predict(fit_pneumo1, newdata = pneumo_25, type = "probs")
```

```
##      mild    normal    severe
## 0.3333333 0.3333333 0.3333333
```

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

```
fit_pneumo2 <- polr(status ~ year, data = pneumo, Hess = TRUE)

# Make prediction

predict(fit_pneumo2, newdata = pneumo_25, type = "probs")
```

```
##      mild    normal    severe
## 0.3333333 0.3333333 0.3333333
```

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
# Fit the model for whether he has the disease

pneumo$normal <- pneumo$status=="normal"

fit1 <- multinom(normal ~ year, data = pneumo, Hess = TRUE)
```

```
## # weights:  3 (2 variable)
## initial  value 16.635532
## final  value 15.276340
## converged
```

```r
# Fit the model for whether they have a severe or moderate case

pneumo_sick <- pneumo %>%
  filter(status == 'severe' | status == 'mild')

fit2 <- multinom(status ~ year, data = pneumo_sick, Hess = TRUE)
```

```
## Warning in multinom(status ~ year, data = pneumo_sick, Hess = TRUE): group
## 'normal' is empty
```

```
## # weights:  3 (2 variable)
## initial  value 11.090355
## final  value 11.090355
## converged
```

```r
  #stan_glm(re78 ~ treat + age + married + sample + educ_cat4 + educ + black, data=lalonde, subset = re
```

4. Compare the three analyses.

The first and second analyses give the same results. The third analysis is more in depth as it looks at the values in different stages so it might be the best course to take out of the three.