MA678 Homework 6

Jing Xu

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.

```
fit_1 <- polr(factor(partyid3) ~ race + age + ideo, data = nes)
summary(fit_1)
```

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

## Call:
## polr(formula = factor(partyid3) ~ race + age + ideo, data = nes)
##
## Coefficients:
##          Value Std. Error t value
## race -0.201703   0.080165  -2.516
## age  -0.004661   0.005975  -0.780
## ideo  0.578086   0.054746  10.560
##
## Intercepts:
##                           Value   Std. Error t value
## democrats|independents     1.6165  0.3945     4.0980
## independents|republicans   1.9823  0.3977     4.9839
## republicans|apolitical     7.5476  0.8120     9.2952
##
## Residual Deviance: 853.8793
## AIC: 865.8793
## (635 observations deleted due to missingness)
```

```
# plot <- melt(nes, id.vars = c("ideo", "race", "age"), variable.name = "partyid3", value.name = "prob"
#
# ggplot(data = nes) +
#   geom_bar(stat = "identity") + aes(x = race, y = mean, fill = partyid3) +
#   facet_grid(age~ideo)
```

2. Explain the results from the fitted model.

```
coef(fit_1)
```

```
##        race         age        ideo
## -0.20170256 -0.00466084  0.57808600
```

every unit increase of race or age is related to .202 or .005 decrease of log odds for partyid3 while every unit increase of ideology is associated to .578 positive change of it.

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

```
residuals(fit_1)
```

```
## NULL
```

(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.
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.

```
chisquare <- chisq.test(cell[, 2:5]) |>print()
```

```
##
##  Pearson's Chi-squared test
##
## data:  cell[, 2:5]
## X-squared = 17.648, df = 3, p-value = 0.0005199
```

```
fit_3 <- vglm(cbind(small, moderate, large) ~ treatment, family = multinomial, data = cell) |> print()
```

```
## Warning in vglm.fitter(x = x, y = y, w = w, offset = offset, Xm2 = Xm2, :
## iterations terminated because half-step sizes are very small
```

```
## Warning in vglm.fitter(x = x, y = y, w = w, offset = offset, Xm2 = Xm2, : some
## quantities such as z, residuals, SEs may be inaccurate due to convergence at a
## half-step
```

```
##
## Call:
## vglm(formula = cbind(small, moderate, large) ~ treatment, family = multinomial,
##     data = cell)
##
##
## Coefficients:
##      (Intercept):1      (Intercept):2 treatmentvaccine:1 treatmentvaccine:2
##         1.60943791         0.47000363        -2.21557372         0.02247286
##
## Degrees of Freedom: 4 Total; 0 Residual
## Residual deviance: 4.551914e-15
## Log-likelihood: -7.125072
##
## This is a multinomial logit model with 3 levels
```

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.

```
fv <- fitted.values(fit_3)
rp <- residuals(fit_3, type = 'pearson')
rd <- residuals(fit_3, type = 'deviance')
cat("the fitted value is ", fv, fill = TRUE)
```

```
## the fitted value is  0.6578947 0.1714286 0.2105263 0.5142857 0.1315789
## 0.3142857
```

```
cat("the Pearson residuals is ", rp, fill = TRUE)
```

```
## the Pearson residuals is  -1.585088e-15 7.141544e-16 1.924104e-16 -2.31607e-15
```

```
cat("the deviance residuals is ", rd, fill = TRUE)
```

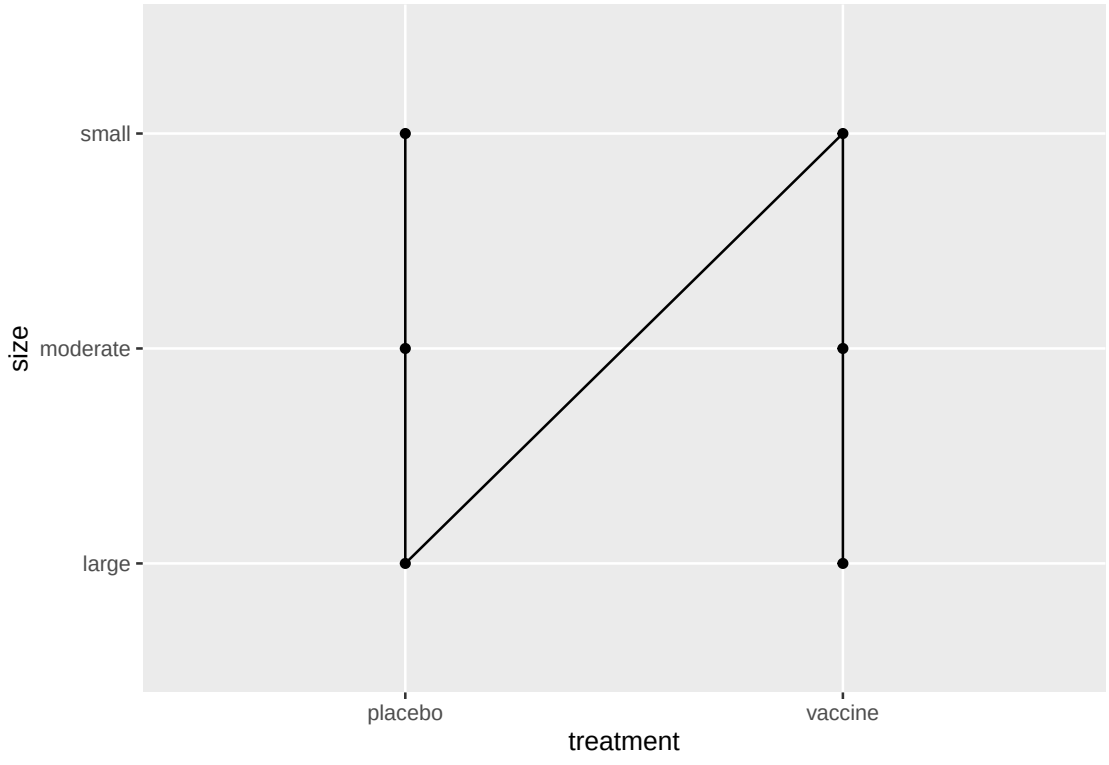
```
## the deviance residuals is
```

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.

```
df <- data.frame(treatment = c(rep('placebo',3),rep('vaccine',3)),
                         size = rep(c('small','moderate','large'),2),
                         value = c(25,8,5,6,18,11))%>%mutate(
                           treatment = as.factor(treatment),
                           size = as.factor(size)
                         )
fit_4 <- polr(size ~  treatment, Hess = T, data = df)
summary(fit_4)
```

```
## Call:
## polr(formula = size ~ treatment, data = df, Hess = T)
##
## Coefficients:
##                       Value Std. Error   t value
## treatmentvaccine 1.802e-13        1.5 1.202e-13
##
## Intercepts:
##                 Value   Std. Error t value
## large|moderate -0.6931  1.1456    -0.6050
## moderate|small  0.6931  1.1456     0.6050
##
## Residual Deviance: 13.18335
## AIC: 19.18335
```

```
ggplot(data = df, aes(x = treatment, y = size, group = 1)) +
  geom_point() +
  geom_line()
```



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).

```
fit_5 <- multinom(prog ~ gender + race + ses + schtyp + read + write + math + science + socst, hsb, tra
summary(fit_5)
```

```
## 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.

```
predict(fit_5,type="probs")[99,]
```

```
##  academic   general  vocation
## 0.1939578 0.2830642 0.5229780
```

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_6 <- polr(factor(happy) ~ money + sex + love + work, Hess = T, data = happy)
summary(fit_6)
```

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

```
coef(fit_6)
```

```
##      money        sex       love       work
##  0.0224593 -0.4734361  3.6076415  0.8875138
```

every unit changing in money, sex, love, work is related to .022, -.473, 3.608. .888 difference of log odds for happy.

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

```
predict(fit_6, newdata = data.frame(love = 0,sex = 0,work = 0,money = 30),type = "probs")
```

```
##            2            3            4            5            6            7
## 9.918136e-01 5.151715e-03 2.828232e-03 1.727561e-04 1.402708e-05 1.707206e-05
##            8            9           10
## 2.514648e-06 4.984512e-08 1.086838e-08
```

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)
df <- uncviet %>%
  group_by(sex,year) %>%
  summarise(y = sum(y))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

```
head(df)
```

```
## # A tibble: 6 x 3
## # Groups:   sex [2]
##   sex    year       y
##   <fct>  <fct>  <dbl>
## 1 Female Fresh     77
## 2 Female Grad     187
## 3 Female Junior   167
## 4 Female Senior   101
## 5 Female Soph      50
## 6 Male   Fresh    439
```

```
fit_7 <- glm(y ~ sex + year, family = poisson(link = 'log') , data = df)
summary(fit_7)
```

```
##
## Call:
## glm(formula = y ~ sex + year, family = poisson(link = "log"),
##     data = df)
##
## Deviance Residuals:
```

```
##       1        2        3        4        5        6        7        8
## -1.9526   0.6060   4.9084  -0.4330  -4.7018   0.8921  -0.2913  -2.5630
##       9       10
##  0.2045   2.0134
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.55837    0.05778  78.893   <2e-16 ***
## sexMale      1.48324    0.04591  32.305   <2e-16 ***
## yearGrad     0.62809    0.05452  11.521   <2e-16 ***
## yearJunior   0.15415    0.05999   2.569   0.0102 *
## yearSenior   0.09953    0.06076   1.638   0.1014
## yearSoph    -0.04763    0.06301  -0.756   0.4497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1627.776  on 9  degrees of freedom
## Residual deviance:   62.113  on 4  degrees of freedom
## AIC: 146.78
##
## Number of Fisher Scoring iterations: 4
```

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.

```
fit_8 <- multinom(status ~ year, Hess = T, weights = Freq, data = pneumo)
```

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

```
summary(fit_8)
```

```
## Call:
## multinom(formula = status ~ year, data = pneumo, weights = Freq,
##     Hess = T)
##
## 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
```

```
predict(fit_8, data.frame(year = 25), type = 'probs')
```

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

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

```
fit_9 <- polr(status ~ year, Hess = T, weights = Freq, data = pneumo)
summary(fit_9)
```

```
## Call:
## polr(formula = status ~ year, data = pneumo, weights = Freq,
##     Hess = T)
##
## 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
```

```
predict(fit_9, data.frame(year = 25), type = 'probs')
```

```
##       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.

```
df <- pneumo %>% filter(status=="normal" | status=="mild")
```

4. Compare the three analyses.