Multinomial Logistic regression

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Multinomial Outcome

- People's occupational choices might be influenced by their parents' occupations and their own education level. We can study the relationship of one's occupation choice with education level and father's occupation. The occupational choices will be the outcome variable which consists of categories of occupations.
- A biologist may be interested in food choices that alligators make. Adult alligators might have different preferences from young ones. The outcome variable here will be the types of food, and the predictor variables might be size of the alligators and other environmental variables.
- Entering high school students make program choices among general program, vocational program and academic program. Their choice might be modeled using their writing score and their social economic status.

Alternative Methods

- Multiple Logistic Regression Analyses

Alternative Methods

- Multiple Logistic Regression Analyses
  - Each analysis is potentially run on a different set of samples

Alternative Methods

- Multiple Logistic Regression Analyses
  - Each analysis is potentially run on a different set of samples
  - Without constraining the logistic regression models, we can end up with the probability of choosing all possible outcome categories greater than 1.
- Collapsing the number categories to two then perform logistic regression

Alternative Methods

- Multiple Logistic Regression Analyses
  - Each analysis is potentially run on a different set of samples
  - Without constraining the logistic regression models, we can end up with the probability of choosing all possible outcome categories greater than 1.
- Collapsing the number categories to two then perform logistic regression
  - This approach suffers from loss of information and changes the original research questions to very different ones.

Examples: Career Choices

- The data set contains variables on 200 students. The outcome variable is prog, program type: general, vocational, academic.
- The predictor variables are social economic status, ses, a three-level categorical variable and writing score, write, a continuous variable.

Multinomial Logistic Regression

$$
\begin{aligned}
log[\frac{P(general)}{P(academic)}] &= \beta_{10} + \beta_{11} I(SES=2) + \beta_{12} I(SES=3) + \beta_{13} write \\
log[\frac{P(vocational)}{P(academic)}] &= \beta_{20} + \beta_{21} I(SES=2) + \beta_{22} I(SES=3) + \beta_{23} write
\end{aligned}
$$

- The ratio of the probability of choosing one outcome category over the probability of choose the baseline category is often referred as relative risk (sometimes referred as odds).
- Relative risk ratio

$$
\frac{\frac{P(general|X=x+1)}{P(academic|X=x+1)}}{\frac{P(general|X=x)}{P(academic|X=x)}}
$$

$$\begin{aligned} log[\frac{P(general)}{P(academic)}] &= \beta_{10} + \beta_{11} I(SES = 2) + \beta_{12} I(SES = 3) + \beta_{13} write \\ log[\frac{P(vocational)}{P(academic)}] &= \beta_{20} + \beta_{21} I(SES = 2) + \beta_{22} I(SES = 3) + \beta_{23} write \end{aligned}$$

- β_{13}: A one-unit increase in the variable **write** is associated with β_{13} increase in the log relative risk (odds) of being in general program versus academic program.

Multinomial Logistic Regression

$$\begin{aligned} log[\frac{P(general)}{P(academic)}] &= \beta_{10} + \beta_{11} I(SES = 2) + \beta_{12} I(SES = 3) + \beta_{13} write \\ log[\frac{P(vocational)}{P(academic)}] &= \beta_{20} + \beta_{21} I(SES = 2) + \beta_{22} I(SES = 3) + \beta_{23} write \end{aligned}$$

- β_{13}: A one-unit increase in the variable **write** is associated with β_{13} increase in the log relative risk (odds) of being in general program versus academic program.
- β_{23} A one-unit increase in the variable **write** is associated with β_{23} increase in the log relative risk (odds) of being in vocational program versus academic program.

Example:Career Choices

```
> dat<-read.dta(file=careerurl)
> dat[1:3,]
   id female    ses schtyp     prog read write math science socst       honors awards cid
1  45 female    low public vocation   34    35   41      29    26 not enrolled      0   1
2 108   male middle public  general   34    33   41      36    36 not enrolled      0   1
3  15   male   high public vocation   39    39   44      26    42 not enrolled      0   1

> with(ml, table(ses, prog))
        prog
ses      general academic vocation
  low         16       19       12
  middle      20       44       31
  high         9       42        7
```

Examples for Multinomial Logistic Regression

```
> ml$prog2 <- relevel(ml$prog, ref = "academic")
> test <- multinom(prog2 ~ ses + write, data = ml)
# weights:  15 (8 variable)
initial  value 219.722458
iter  10 value 179.982880
final  value 179.981726
converged
>
> summary(test)
Call:
multinom(formula = prog2 ~ ses + write, data = ml)

Coefficients:
         (Intercept)  sesmiddle    seshigh      write
general     2.852198 -0.5332810 -1.1628226 -0.0579287
vocation    5.218260  0.2913859 -0.9826649 -0.1136037

Std. Errors:
         (Intercept) sesmiddle   seshigh      write
general     1.166441 0.4437323 0.5142196 0.02141097
vocation    1.163552 0.4763739 0.5955665 0.02221996

Residual Deviance: 359.9635
AIC: 375.9635
```

Examples for Multinomial Logistic Regression

```
> z <- summary(test)$coefficients/summary(test)$standard.errors
> z
         (Intercept)  sesmiddle   seshigh     write
general     2.445214 -1.2018081 -2.261334 -2.705562
vocation    4.484769  0.6116747 -1.649967 -5.112689
>
> p <- (1 - pnorm(abs(z), 0, 1)) * 2
> p
          (Intercept) sesmiddle    seshigh        write
general  0.0144766100 0.2294379 0.02373856 6.818902e-03
vocation 0.0000072993 0.5407530 0.09894976 3.176045e-07
> exp(coef(test))
         (Intercept) sesmiddle   seshigh     write
general     17.32582 0.5866769 0.3126026 0.9437172
vocation   184.61262 1.3382809 0.3743123 0.8926116
```

Example: Career Choices

```
> dses <- data.frame(ses = c("low", "middle", "high"), write = mean(ml$write))
> dses
     ses  write
1    low 52.775
2 middle 52.775
3   high 52.775
> predict(test, newdata = dses, "probs")
   academic   general  vocation
1 0.4396845 0.3581917 0.2021238
2 0.4777488 0.2283353 0.2939159
3 0.7009007 0.1784939 0.1206054
```

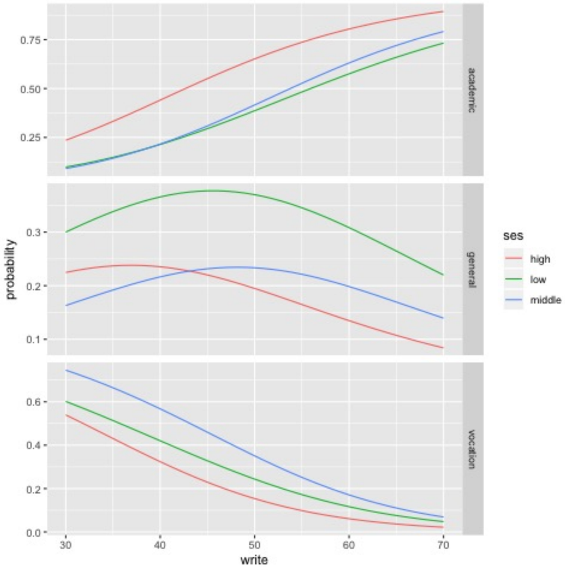
Example

```
> dwrite[1:5,]
  ses write
1 low    30
2 low    31
3 low    32
4 low    33
5 low    34
> pp.write <- cbind(dwrite, predict(test, newdata = dwrite, type = "probs", se = TRUE))
> pp.write[1:5,]
  ses write   academic   general  vocation
1 low    30 0.09843588 0.2999880 0.6015762
2 low    31 0.10716868 0.3082195 0.5846118
3 low    32 0.11650390 0.3162093 0.5672868
4 low    33 0.12645834 0.3239094 0.5496323
5 low    34 0.13704576 0.3312711 0.531683
> by(pp.write[, 3:5], pp.write$ses, colMeans)
pp.write$ses: high
 academic   general  vocation
0.6164315 0.1808037 0.2027648
---------------------------------------------------------------------------------------
pp.write$ses: low
 academic   general  vocation
0.3972977 0.3278174 0.2748849
---------------------------------------------------------------------------------------
pp.write$ses: middle
 academic   general  vocation
0.4256198 0.2010864 0.3732938
```

Example

```
> lpp <- melt(pp.write, id.vars = c("ses", "write"), value.name = "probability")
> head(lpp)
  ses write variable probability
1 low    30 academic  0.09843588
2 low    31 academic  0.10716868
3 low    32 academic  0.11650390
4 low    33 academic  0.12645834
5 low    34 academic  0.13704576
6 low    35 academic  0.14827643
> ggplot(lpp, aes(x = write, y = probability, colour = ses)) + geom_line() + facet_grid(variable ~
+     ., scales = "free")
```

Example: Career Choices



Example: Alligator

TABLE 7.1 Primary Food Choice of Alligators

			Primary Food Choice				
Lake	Gender	Size (m)	Fish	Invertebrate	Reptile	Bird	Other
Hancock	Male	≤ 2.3	7	1	0	0	5
		> 2.3	4	0	0	1	2
	Female	≤ 2.3	16	3	2	2	3
		> 2.3	3	0	1	2	3
Oklawaha	Male	≤ 2.3	2	2	0	0	1
		> 2.3	13	7	6	0	0
	Female	≤ 2.3	3	9	1	0	2
		> 2.3	0	1	0	1	0
Trafford	Male	≤ 2.3	3	7	1	0	1
		> 2.3	8	6	6	3	5
	Female	≤ 2.3	2	4	1	1	4
		> 2.3	0	1	0	0	0
George	Male	≤ 2.3	13	10	0	2	2
		> 2.3	9	0	0	1	2
	Female	≤ 2.3	3	9	1	0	1
		> 2.3	8	1	0	0	1

Source: Data courtesy of Clint Moore, from an unpublished manuscript by M. F. Delaney and C. T. Moore.

Example

```
> gatorurl<-"http://people.stat.sc.edu/hoyen/Stat705/Data/gator.txt"
> gator<-read.table(file=gatorurl, header=T, sep="")
> gator
   profile Gender Size     Lake Fish Invertebrate Reptile Bird Other
1        1      f <2.3   george    3            9       1    0     1
2        2      m <2.3   george   13           10       0    2     2
3        3      f >2.3   george    8            1       0    0     1
4        4      m >2.3   george    9            0       0    1     2
5        5      f <2.3  hancock   16            3       2    2     3
6        6      m <2.3  hancock    7            1       0    0     5
7        7      f >2.3  hancock    3            0       1    2     3
8        8      m >2.3  hancock    4            0       0    1     2
9        9      f <2.3 oklawaha    3            9       1    0     2
10      10      m <2.3 oklawaha    2            2       0    0     1
11      11      f >2.3 oklawaha    0            1       0    1     0
12      12      m >2.3 oklawaha   13            7       6    0     0
13      13      f <2.3 trafford    2            4       1    1     4
14      14      m <2.3 trafford    3            7       1    0     1
15      15      f >2.3 trafford    0            1       0    0     0
16      16      m >2.3 trafford    8            6       6    3     5
```

Example

```
> gator$Size
 [1] <2.3 <2.3 >2.3 >2.3 <2.3 <2.3 >2.3 >2.3 <2.3 <2.3 >2.3 >2.3 <2.3 <2.3 >2.3 >2.3
Levels: <2.3 >2.3
> gator$Size = factor(gator$Size,levels=levels(gator$Size)[2:1])
> gator$Size
 [1] <2.3 <2.3 >2.3 >2.3 <2.3 <2.3 >2.3 >2.3 <2.3 <2.3 >2.3 >2.3 <2.3 <2.3 >2.3 >2.3
Levels: >2.3 <2.3
> totaln=sum(gator[1:16,5:9]) ## total sample size
> contrasts(gator$Size)=contr.treatment(levels(gator$Size),base=2)
> contrasts(gator$Size)
     >2.3
>2.3    1
<2.3    0
> contrasts(gator$Lake)<-contr.treatment(levels(gator$Lake), base=2)
> contrasts(gator$Lake)
         george oklawaha trafford
george        1        0        0
hancock       0        0        0
oklawaha      0        1        0
trafford      0        0        1
>
> contrasts(gator$Gender)=contr.treatment(levels(gator$Gender),base=2)
> contrasts(gator$Gender)
  f
f 1
m 0
```

Example

```
> fit5=vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Lake+Size+Gender, data=gator, family=multinomial)
> summary(fit5)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept):1   -2.4633     0.7739 -3.18294       NA
(Intercept):2   -2.0745     0.6117  -3.392 0.000695 ***
(Intercept):3   -2.9141     0.8856 -3.29043       NA
(Intercept):4   -0.9167     0.4782  -1.917 0.055217 .
Lakegeorge:1    -0.5753     0.7952  -0.723 0.469429
Lakegeorge:2     1.7805     0.6232   2.857 0.004277 **
Lakegeorge:3    -1.1295     1.1928  -0.947 0.343687
Lakegeorge:4    -0.7666     0.5686  -1.348 0.177563
Lakeoklawaha:1  -1.1256     1.1923  -0.944 0.345132
Lakeoklawaha:2   2.6937     0.6693   4.025 5.70e-05 ***
Lakeoklawaha:3   1.4008     0.8105   1.728 0.083926 .
Lakeoklawaha:4  -0.7405     0.7421  -0.998 0.318372
Laketrafford:1   0.6617     0.8461   0.782 0.434145
Laketrafford:2   2.9363     0.6874   4.272 1.94e-05 ***
Laketrafford:3   1.9316     0.8253   2.340 0.019263 *
Laketrafford:4   0.7912     0.5879   1.346 0.178400
Size>2.3:1       0.7302     0.6523   1.120 0.262918
Size>2.3:2      -1.3363     0.4112  -3.250 0.001155 **
Size>2.3:3       0.5570     0.6466   0.861 0.388977
Size>2.3:4      -0.2906     0.4599  -0.632 0.527515
Genderf:1        0.6064     0.6888   0.880 0.378666
Genderf:2        0.4630     0.3955   1.171 0.241796
Genderf:3        0.6276     0.6853   0.916 0.359785
Genderf:4        0.2526     0.4663   0.542 0.588100
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example

$$\begin{aligned}\log\left[\frac{\pi_{Bird}}{\pi_{Fish}}\right] &= -2.4633 - 1.1256\ Oklawaha + 0.6617\,Trafford \\ &- 0.5753\,George + 0.7302\,large + 0.6064\,female\end{aligned}$$

Example

```
> exp(coefficients(fit5))
 (Intercept):1  (Intercept):2  (Intercept):3  (Intercept):4   Lakegeorge:1   Lakegeorge:2   Lakegeorge:3
    0.08515564     0.12562535     0.05425079     0.39982590     0.56255501     5.93289534     0.32320675
  Lakegeorge:4 Lakeoklawaha:1 Lakeoklawaha:2 Lakeoklawaha:3 Lakeoklawaha:4 Laketrafford:1 Laketrafford:2
    0.46460153     0.32445217    14.78619873     4.05843171     0.47686707     1.93813088    18.84663158
Laketrafford:3 Laketrafford:4     Size>2.3:1     Size>2.3:2     Size>2.3:3     Size>2.3:4      Genderf:1
    6.90044926     2.20601430     2.07557742     0.26282653     1.74549121     0.74782762     1.83387028
     Genderf:2      Genderf:3      Genderf:4
    1.58877439     1.87303229     1.28732894
```

Example: Alligator