Lecture 8: Multinomial Regression (Baseline Category Logits)

This example uses the alligator food choice data in Section 8.1.2 of the Agresti textbook. The study is interested in studying factors that influence primary food source in $n = 219$ alligators. We are interested in modeling the nominal response variable primary food type (fish, invertebrate, reptile, bird, other) as a function of three covariates:

1. Gender: f = female, m = male
2. Size: <2.3 = <= 2.3 meters long, >2.3 = >2.3 meters long
3. Lake: george, hancock, oklawaha, trafford

```
rm(list=ls())
data = read.csv("alligator.csv", header = T)
names(data)
```

```
## [1] "profile"      "Gender"       "Size"         "Lake"
## [5] "Fish"         "Invertebrate" "Reptile"      "Bird"
## [9] "Other"
```

```
dim(data)
```

```
## [1] 16  9
```

```
data
```

```
##    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
```

The data is in grouped table form. The columns labeled *Fish, Invertebrate, Reptile, Bird, Other* are the counts associated with each of the 16 unique combinations of predictors. For example, 3 small female alligators in Lake George (profile 1) eat fish as their primary food source.

1 Model Fit

We will fit the baseline category logit model using *Fish* as the baseline category because it is the largest category. Thus the logit equations are:

$$\log\left(\frac{\pi_j}{\pi_F}\right) = \beta_0 + \beta_1 X_1 + \ldots$$

where π_F is the probability of *Fish* and j = *Invertebrate, Reptile, Bird, Other*. There are 3 categorical predictors, *Gender, Size, Lake*, with 2, 2 and 4 levels, respectively. Thus, we have a total of 5 independent variables in the model plus an intercept, for a total of 24 parameters (4 logits with 6 parameters each).

```
# install.packages("VGAM")
library(VGAM)
```

```
## Warning: package 'VGAM' was built under R version 3.4.4
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
# By default R makes the first level (by alphabetical order) in each categorical predictor
# the baseline category.  It is good practice to use the largest level as the baseline.

n.size = aggregate(data[,5:9],by=list(data$Size),FUN=sum)
cbind(n.size,rowSums(n.size[,2:6]))
```

```
##   Group.1 Fish Invertebrate Reptile Bird Other rowSums(n.size[, 2:6])
## 1    <2.3   49           45       6    5    19                    124
## 2    >2.3   45           16      13    8    13                     95
```

```
n.lake = aggregate(data[,5:9],by=list(data$Lake),FUN=sum)
cbind(n.lake,rowSums(n.lake[,2:6]))
```

```
##    Group.1 Fish Invertebrate Reptile Bird Other rowSums(n.lake[, 2:6])
## 1   george   33           20       1    3     6                     63
## 2  hancock   30            4       3    5    13                     55
## 3 oklawaha   18           19       7    1     3                     48
## 4 trafford   13           18       8    4    10                     53
```

```
n.gender = aggregate(data[,5:9],by=list(data$Gender),FUN=sum)
cbind(n.gender,rowSums(n.gender[,2:6]))
```

```
##   Group.1 Fish Invertebrate Reptile Bird Other rowSums(n.gender[, 2:6])
## 1       f   35           28       6    6    14                      89
## 2       m   59           33      13    7    18                     130
```

```
data$Gender <- relevel(data$Gender, "m")

# By default VGLM will use the last level as the baseline level for creating the logits. #
# To set Fish as the baseline level, specify it last in vglm call below: #

model = vglm(cbind(Invertebrate,Reptile,Bird,Other,Fish)~Lake+Size+Gender, data=data, family=multinomial
summary(model)
```

```
##
## Call:
## vglm(formula = cbind(Invertebrate, Reptile, Bird, Other, Fish) ~
##     Lake + Size + Gender, family = multinomial, data = data)
##
##
## Pearson residuals:
```

```
##                         Min      1Q   Median     3Q   Max
## log(mu[,1]/mu[,5]) -1.3218 -0.4611  0.01054 0.3810 1.866
## log(mu[,2]/mu[,5]) -0.7033 -0.5751 -0.35511 0.2610 2.064
## log(mu[,3]/mu[,5]) -1.1985 -0.5478 -0.22421 0.3678 3.478
## log(mu[,4]/mu[,5]) -1.6945 -0.2893 -0.10807 1.1236 1.367
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -0.29394    0.35527  -0.827 0.408030
## (Intercept):2  -4.04360    1.18395  -3.415 0.000637 ***
## (Intercept):3  -3.03854    0.83192  -3.652 0.000260 ***
## (Intercept):4  -1.68330    0.52097  -3.231 0.001233 **
## Lakehancock:1  -1.78051    0.62321  -2.857 0.004277 **
## Lakehancock:2   1.12946    1.19280   0.947 0.343687
## Lakehancock:3   0.57527    0.79522   0.723 0.469429
## Lakehancock:4   0.76658    0.56855   1.348 0.177563
## Lakeoklawaha:1  0.91318    0.47612   1.918 0.055114 .
## Lakeoklawaha:2  2.53026    1.12211   2.255 0.024139 *
## Lakeoklawaha:3 -0.55035    1.20980  -0.455 0.649174
## Lakeoklawaha:4  0.02606    0.77776   0.034 0.973273
## Laketrafford:1  1.15582    0.49279   2.345 0.019002 *
## Laketrafford:2  3.06105    1.12972   2.710 0.006737 **
## Laketrafford:3  1.23699    0.86610   1.428 0.153225
## Laketrafford:4  1.55776    0.62567   2.490 0.012784 *
## Size>2.3:1     -1.33626    0.41119  -3.250 0.001155 **
## Size>2.3:2      0.55704    0.64661   0.861 0.388977
## Size>2.3:3      0.73024    0.65228   1.120 0.262918
## Size>2.3:4     -0.29058    0.45993  -0.632 0.527515
## Genderf:1       0.46296    0.39552   1.171 0.241796
## Genderf:2       0.62756    0.68528   0.916 0.359785
## Genderf:3       0.60643    0.68884   0.880 0.378666
## Genderf:4       0.25257    0.46635   0.542 0.588100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])
##
## Residual deviance: 50.2637 on 40 degrees of freedom
##
## Log-likelihood: -73.3221 on 40 degrees of freedom
##
## Number of iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', '(Intercept):3'
##
## Reference group is level  5  of the response
```

```
exp(coefficients(model))
```

```
##  (Intercept):1  (Intercept):2  (Intercept):3  (Intercept):4  Lakehancock:1
##     0.74532203     0.01753422     0.04790473     0.18575973     0.16855177
```

```
##  Lakehancock:2  Lakehancock:3  Lakehancock:4 Lakeoklawaha:1 Lakeoklawaha:2
##     3.09399475     1.77760394     2.15238205     2.49223994    12.55676642
## Lakeoklawaha:3 Lakeoklawaha:4 Laketrafford:1 Laketrafford:2 Laketrafford:3
##     0.57674745     1.02640013     3.17663308    21.34995379     3.44522908
## Laketrafford:4     Size>2.3:1     Size>2.3:2     Size>2.3:3     Size>2.3:4
##     4.74818559     0.26282653     1.74549121     2.07557742     0.74782762
##      Genderf:1      Genderf:2      Genderf:3      Genderf:4
##     1.58877439     1.87303229     1.83387028     1.28732894
```

2 Inference: Goodness-of-Fit (GOF)

```
deviance(model)
```

```
## [1] 50.26369
```

```
qchisq(.05,16*4 - 6*4,lower.tail=FALSE)
```

```
## [1] 55.75848
```

```
1-pchisq(deviance(model),16*4 - 6*4)
```

```
## [1] 0.128185
```

The deviance GOF test concludes we cannot reject the null/reduced model in favor of the saturated model, indicating the model fit with main effects for all the predictors provides a reasonable fit. Note that the degrees of freedom for the χ^2 is the difference between the number of parameters in the saturated model (a parameter for each unique combination of predictors for each of the four logit models) and the number of parameters in the reduced model.

3 Inference: Predictor Effects

The four sets of coefficients predict the log odds of and describe the effect of each predictor for:

- Invertebrate versus Fish (1 vs 5),
- Reptile versus Fish (2 vs 5),
- Bird versus Fish (3 vs 5), and
- Other versus Fish (4 vs 5).

Referring to the model output, we find that the Wald hypothesis tests indicate none of the coefficients for *Gender* are statistically significant. Some of the coefficients for *Lake* and *Size* are statistically significant. For example, *Size* has a statistically significant effect when comparing *Fish* to *Invertebrate*.

The intercepts give the estimated log odds for the reference group: *Lake* = *george*, *Size* = *<2.3*, *Gender* = *m*. For example, the estimated log odds of birds versus fish in this group is -3.04; the estimated log odds of invertebrates versus fish is -0.29; and so on.

The lake effect is characterized by three dummy coefficients in each of the four logit equations. The estimated coefficient for the Lake Hancock dummy in the invertebrate-versus-fish equation is -1.78. This means that alligators in Lake Hancock are less likely to choose invertebrates over fish than the alligators in Lake George are. The estimated odds ratio of $exp(-1.78) = 0.169$ is the same for alligators of all sex and sizes, because this is a model with main effects but no interactions.

4 Predictions

The estimated prediction equations are functions of the predictors and the estimated coefficients. For example, the estimated prediction equation for the log odds of *Bird* relative to *Fish* is:

$$\log\left(\frac{\hat{\pi}_{Bird}}{\hat{\pi}_{Fish}}\right) = -3.03854 + 0.57527 * Hancock - 0.55035 * Oklawaha + 1.23699 * Trafford$$
$$+0.73024 * Size > 2.3 + 0.60643 * female$$

5 Model Comparison

We can use deviances to compare nested models. For example, should Gender be excluded from the model?

```
model2 = vglm(cbind(Invertebrate,Reptile,Bird,Other,Fish)~Lake+Size, data=data, family=multinomial)
summary(model2)
```

```
##
## Call:
## vglm(formula = cbind(Invertebrate, Reptile, Bird, Other, Fish) ~
##     Lake + Size, family = multinomial, data = data)
##
##
## Pearson residuals:
##                         Min      1Q  Median     3Q   Max
## log(mu[,1]/mu[,5]) -1.3716 -0.4379 -0.0248 0.2436 1.995
## log(mu[,2]/mu[,5]) -0.8298 -0.5850 -0.2309 0.2225 2.237
## log(mu[,3]/mu[,5]) -0.9873 -0.5082 -0.1144 0.2373 3.994
## log(mu[,4]/mu[,5]) -1.5873 -0.3189 -0.0159 1.0330 1.413
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -0.090814   0.308039  -0.295 0.768136
## (Intercept):2  -3.665796   1.058958  -3.462 0.000537 ***
## (Intercept):3  -2.723736   0.710393  -3.834 0.000126 ***
## (Intercept):4  -1.572721   0.474822  -3.312 0.000926 ***
## Lakehancock:1  -1.658359   0.612877  -2.706 0.006813 **
## Lakehancock:2   1.242777   1.185418   1.048 0.294461
## Lakehancock:3   0.695118   0.781263   0.890 0.373608
## Lakehancock:4   0.826196   0.557541   1.482 0.138378
## Lakeoklawaha:1  0.937219   0.471905   1.986 0.047030 *
## Lakeoklawaha:2  2.458872   1.118113   2.199 0.027869 *
## Lakeoklawaha:3 -0.653208   1.201917  -0.543 0.586805
## Lakeoklawaha:4  0.005653   0.776571   0.007 0.994192
## Laketrafford:1  1.121985   0.490513   2.287 0.022174 *
## Laketrafford:2  2.935253   1.116395   2.629 0.008558 **
## Laketrafford:3  1.087767   0.841669   1.292 0.196221
## Laketrafford:4  1.516369   0.621435   2.440 0.014683 *
## Size>2.3:1     -1.458205   0.395945  -3.683 0.000231 ***
## Size>2.3:2      0.351263   0.580032   0.606 0.544785
## Size>2.3:3      0.630660   0.642473   0.982 0.326291
## Size>2.3:4     -0.331550   0.448252  -0.740 0.459511
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])
##
## Residual deviance: 52.4785 on 44 degrees of freedom
##
## Log-likelihood: -74.4295 on 44 degrees of freedom
##
## Number of iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', '(Intercept):3'
##
## Reference group is level  5  of the response
```

```
deviance(model2) - deviance(model)
```

```
## [1] 2.214799
```

```
qchisq(.05,6*4 - 5*4,lower.tail=FALSE)
```

```
## [1] 9.487729
```

```
1-pchisq(deviance(model2) - deviance(model),6*4 - 5*4)
```

```
## [1] 0.6963208
```

The difference in deviance test concludes we cannot reject the null/reduced model (excluding *Gender*) in favor of the full model (including *Gender*), indicating the the model excluding *Gender* provides a reasonable fit.