MULTINOMIAL LOGISTIC REGRESSION

Dr. Aric LaBarr

Institute for Advanced Analytics

INTRODUCTION

Multiple (Unordered) Outcomes

- Up to this point, we only considered ordinal response variables with binary being a popular special case.
- Easy to generalize the binary case to the ordinal case – many binary models!
- Need to change the underlying model and math slightly to extend to **nominal** response variables.

Logistic Models

- Binary (probability that observation *i* has the event):

$$= \beta_0 + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

- Ordinal (probability that observation *i* has **at most** event *j*, and $j = 1, \ldots, m$):

$$= \beta_{0,j} + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

- Multinomial (probability that observation *i* has event *j*, and $j = 1, \ldots, m$):

$$= \beta_{0,j} + \beta_{1,j} x_{1,i} + \cdots \beta_{k,j} x_{k,i}$$

Logistic Models

- Binary (probability that observation *i* has the event):

$$= \beta_0 + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

- Ordinal (probability that observation *i* has **at most** event *j*, and $j = 1, \ldots, m$):

$$= \beta_{0,j} + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

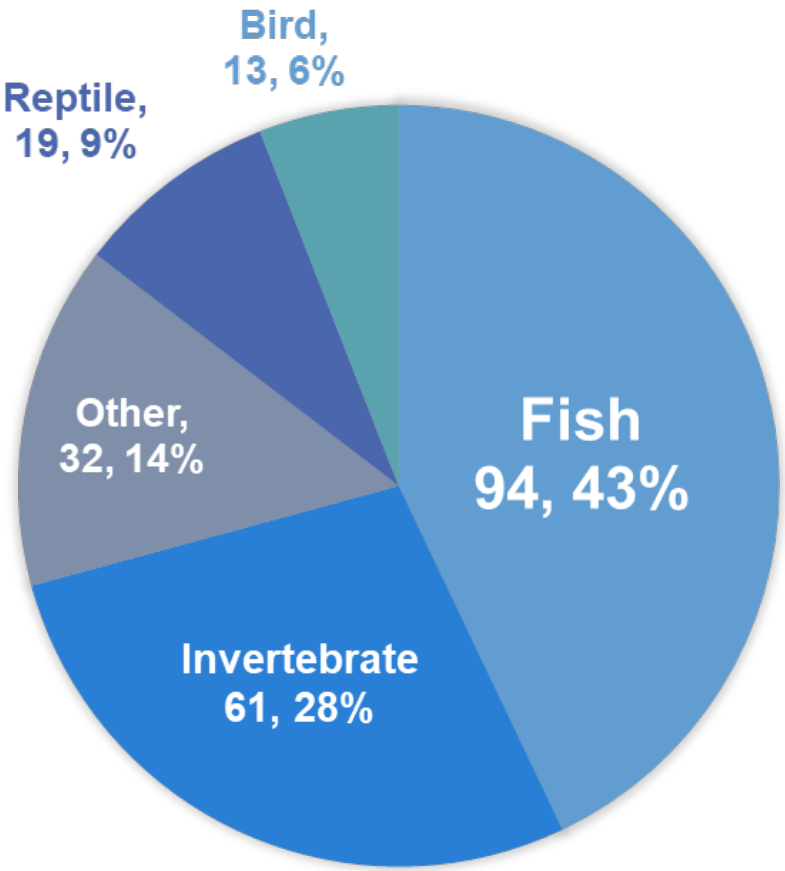
- Multinomial (probability that observation *i* has event *j*, and $j = 1, \ldots, m$):

$$= \beta_{0,j} + \beta_{1,j} x_{1,i} + \cdots \beta_{k,j} x_{k,i}$$

Both intercept and slope changes!

Alligator Food Preference Data Set

- Model the association between various factors and alligator food choices.
- 219 observations in the data set.



Alligator Food Preference Data Set

- Model the association between various factors and alligator food choices.
- 4 lakes in Florida.
- Predictors:
  - **size:** alligator's size ($\leq$ 2.3m long = small, $>$ 2.3m long = large)
  - **lake:** lake where alligator was captured (George, Hancock, Oklawaha, Trafford)
  - **gender:** male or female alligator

View Data

	size	food	lake	gender	count
1	<= 2.3 meters	Fish	Hancock	Male	7
2	<= 2.3 meters	Invertebrate	Hancock	Male	1
3	<= 2.3 meters	Other	Hancock	Male	5
4	> 2.3 meters	Fish	Hancock	Male	4
5	> 2.3 meters	Bird	Hancock	Male	1
6	> 2.3 meters	Other	Hancock	Male	2
7	<= 2.3 meters	Fish	Hancock	Female	16
8	<= 2.3 meters	Invertebrate	Hancock	Female	3
9	<= 2.3 meters	Reptile	Hancock	Female	2
10	<= 2.3 meters	Bird	Hancock	Female	2

⋮

GENERALIZED LOGIT MODEL

Generalized Logits

- If the outcome variable had *m* levels (with *m* being the reference category) with proportions $(p_1, p_2, \ldots, p_m)$, then the generalized logits are the following:

$$\log\left(\frac{p_1}{p_m}\right), \log\left(\frac{p_2}{p_m}\right), \ldots, \log\left(\frac{p_{m-1}}{p_m}\right)$$

- Fitting *m-1* models but the denominator in the logit **is not** the complement of the numerator – it is the reference level probability.

Alligator Food Preference Models

- For the alligator data, we have $m = 5$ outcomes, so the models with the fish category as the reference are:

$$\log\left(\frac{p_{i,\text{bird}}}{p_{i,\text{fish}}}\right) = \beta_{0,\text{bird}} + \beta_{1,\text{bird}}\text{lakeH}_i + \beta_{2,\text{bird}}\text{lakeO}_i + \beta_{3,\text{bird}}\text{lakeT}_i + \beta_{4,\text{bird}}\text{size}_i + \beta_{5,\text{bird}}\text{gender}_i$$

$$\vdots$$

$$\log\left(\frac{p_{i,\text{other}}}{p_{i,\text{fish}}}\right) = \beta_{0,\text{other}} + \beta_{1,\text{other}}\text{lakeH}_i + \beta_{2,\text{other}}\text{lakeO}_i + \beta_{3,\text{other}}\text{lakeT}_i + \beta_{4,\text{other}}\text{size}_i + \beta_{5,\text{other}}\text{gender}_i$$

Multinomial Logistic Regression

```
gator$food <- factor(gator$food)
gator$food <- relevel(gator$food, ref = "Fish")

library(VGAM)

glogit.model <- vglm(food ~ size + lake + gender,
                     weight = count, data = gator,
                     family = multinomial(refLevel = "Fish"))

summary(glogit.model)
```

Multinomial Logistic Regression

```
Call: vglm(formula = food ~ size + lake + gender, family = multinomial(refLevel = "Fish"),
data = gator, weights = count)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept):1      -2.43211    0.77066      NA       NA
(Intercept):2       0.16902    0.37875   0.446  0.65541
(Intercept):3      -1.43073    0.53809  -2.659  0.00784 **
(Intercept):4      -3.41604    1.08513      NA       NA
size> 2.3 meters:1  0.73024    0.65228   1.120  0.26292
size> 2.3 meters:2 -1.33626    0.41119  -3.250  0.00116 **
size> 2.3 meters:3 -0.29058    0.45993  -0.632  0.52752
size> 2.3 meters:4  0.55704    0.64661   0.861  0.38898
lakeHancock:1       0.57527    0.79522   0.723  0.46943
lakeHancock:2      -1.78051    0.62321  -2.857  0.00428 **
lakeHancock:3       0.76658    0.56855   1.348  0.17756
lakeHancock:4       1.12946    1.19280   0.947  0.34369
lakeOklawaha:1     -0.55035    1.20987  -0.455  0.64919
```

INTERPRETATION

Interpreting Coefficients

- Calculation remains the same:

$$e^{\widehat{\beta}} = e^{0.7302} = 2.076$$

- **Incorrect** interpretation: The probability of eating birds is 2.076 times as likely for large alligators compared to small alligators.
- **Correct** interpretation: The predicted **relative probability** of eating birds **rather than** fish is 2.076 times as likely for large alligators than for small alligators.
- Sometimes these are called **conditional** interpretations.

Relative Probability?

- Although these are often called odds ratios (or conditional odds ratios) they are **not** mathematically odds ratios.
- The exponentiated coefficients from multinomial logistic regressions are **relative risks**, not odds.

$$\exp\left(\log\left(\frac{p_1}{p_m}\right)\right) = \frac{p_1}{p_m}$$

Odds vs. Probability

- **Odds** is the ratio of events to non-events:

$$Odds = \frac{\#yes}{\#no}$$

- **Probability** is the ratio of event to the total number of outcomes:

$$p = \frac{\#yes}{\#yes + \ \#no}$$

- **Odds** and **Probability** are related:

$$Odds = \frac{p}{1-p} \qquad\qquad p = \frac{Odds}{1+Odds}$$

Relative Risk

- **Relative Risk** indicates how likely (in terms of probability) an event is for one group relative to another:

$$RR = \frac{p_A}{p_B}$$

- Since probabilities are always non-negative, so are relative risks
  - RR > 1 → Event **more likely for A than for B**
  - RR < 1 → Event **more likely for B than for A**
  - RR = 1 → Event **equally likely in each group**

Relative Probability!

- Although these are often called odds ratios (or conditional odds ratios) they are **not** mathematically odds ratios.
- The exponentiated multinomial logistic regressions are relative risks, not odds.

$$\exp\left(\log\left(\frac{p_1}{p_m}\right)\right) = \frac{p_1}{p_m}$$

- Exponentiated **coefficients** from a multinomial logistic regression are **relative risk ratios** (RRR), not odds ratios.

Interpretation – R

```
exp(coef(glogit.model))
```

```
     (Intercept):1      (Intercept):2      (Intercept):3      (Intercept):4
       -91.214894          18.414855          -76.086613         -96.715783

size> 2.3 meters:1 size> 2.3 meters:2 size> 2.3 meters:3 size> 2.3 meters:4
       107.557742         -73.717347         -25.217238          74.549121

     lakeHancock:1      lakeHancock:2      lakeHancock:3      lakeHancock:4
        77.760394         -83.144823         115.238205         209.399475

    lakeOklawaha:1     lakeOklawaha:2     lakeOklawaha:3     lakeOklawaha:4
       -42.325256         149.223994           2.640013        1155.676642

    lakeTrafford:1     lakeTrafford:2     lakeTrafford:3     lakeTrafford:4
       244.522908         217.663308         374.818559        2034.995379

      genderMale:1       genderMale:2       genderMale:3       genderMale:4
       -45.470516         -37.058401         -22.319776         -46.610637
```

PREDICTIONS AND DIAGNOSTICS

Similarities

- Multinomial logistic regression has a lot of the same aspects/issues as a binary logistic regression:
  - Multicollinearity still exists.
  - Non-convergence problems still exist.
  - Concordance, Discordance, Tied pairs still exist – so the c statistic still exists.
  - Generalized R^2 remains the same.

Differences

- Multinomial logistic regression has a few aspects/issues that differ from a binary logistic regression:
  - A lot of the diagnostics for binary regression cannot be calculated easily since there are actually **multiple** models – ROC curves for each model?
  - Diagnostics / Influence plots are not available – residuals for each model?
  - Predicted probabilities are for **each** category.

Predicted Probabilities – R

```
pred_probs <- predict(glogit.model, newdata = gator, type = "probs")

head(pred_probs)
```

```
        Bird        Fish Invertebrate      Other     Reptile
1  0.0511489   0.6006519   0.07545711  0.2401562  0.03258584
2  0.0511489   0.6006519   0.07545711  0.2401562  0.03258584
3  0.0511489   0.6006519   0.07545711  0.2401562  0.03258584
4  0.1102286   0.6236514   0.02059152  0.1864723  0.05905622
5  0.1102286   0.6236514   0.02059152  0.1864723  0.05905622
6  0.1102286   0.6236514   0.02059152  0.1864723  0.05905622
```