# 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 $j = 1, \ldots, m$):

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

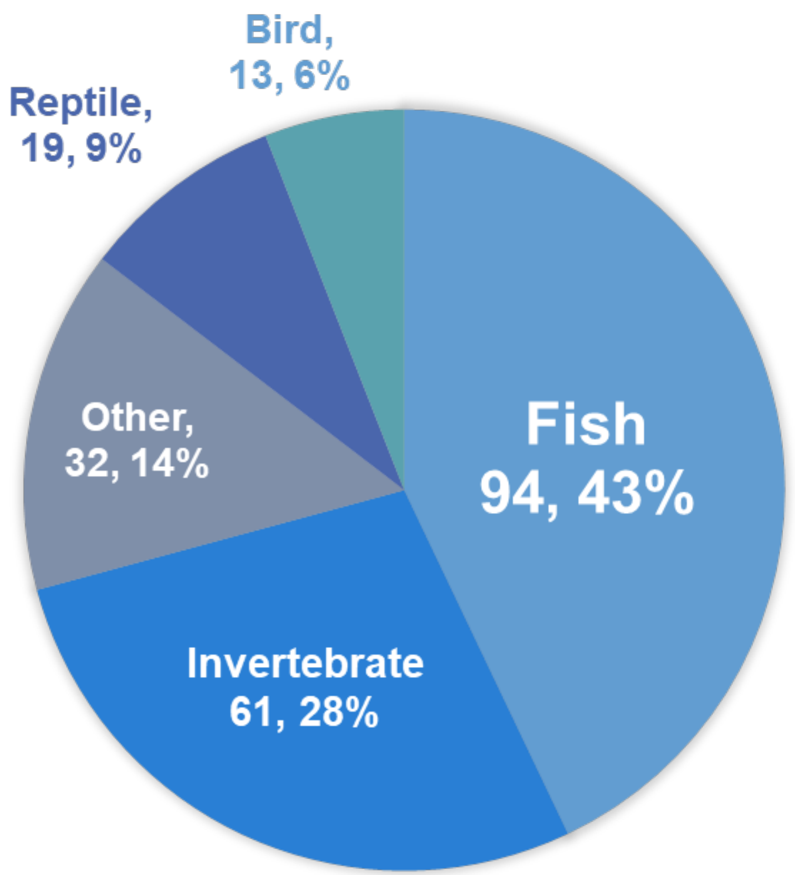
- Multinomial (probability that observation *i* has event *j*, and $j = 1, \ldots, m$):

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

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 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:

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

$$\vdots$$

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

# Multinomial Logistic Regression

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

glogit.model <- multinom(food ~ size + lake + gender,
                         weight = count, data = gator)

summary(glogit.model)
```

# Multinomial Logistic Regression

```
Coefficients:
             (Intercept) size> 2.3 meters lakeHancock lakeOklawaha lakeTrafford  genderMale
Bird          -2.4321397        0.7300740   0.5754699  -0.55020075     1.237216  -0.6064035
Invertebrate   0.1690702       -1.3361658  -1.7805555   0.91304120     1.155722  -0.4629388
Other         -1.4309095       -0.2905697   0.7667093   0.02603021     1.557820  -0.2524299
Reptile       -3.4161432        0.5571846   1.1296426   2.53024945     3.061087  -0.6276217

Std. Errors:
             (Intercept) size> 2.3 meters lakeHancock lakeOklawaha lakeTrafford  genderMale
Bird           0.7706720        0.6522657   0.7952303    1.2098680    0.8661052   0.6888385
Invertebrate   0.3787475        0.4111827   0.6232075    0.4761068    0.4927795   0.3955162
Other          0.5381162        0.4599317   0.5685673    0.7777958    0.6256868   0.4663546
Reptile        1.0851582        0.6466092   1.1928075    1.1221413    1.1297557   0.6852750

Residual Deviance: 537.8655
AIC: 585.8655
```

# INTERPRETATION

# Interpreting Coefficients

- Calculation remains the same:

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

- **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:

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

# 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 probabilites 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) size> 2.3 meters lakeHancock lakeOklawaha lakeTrafford  genderMale
Bird          0.08784866        2.0752341   1.7779659     0.576834     3.446005   0.5453086
Invertebrate  1.18420329        0.2628516   0.1685445     2.491889     3.176316   0.6294311
Other         0.23909136        0.7478374   2.1526708     1.026372     4.748458   0.7769106
Reptile       0.03283884        1.7457506   3.0945502    12.556638    21.350755   0.5338600
```

# 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.
  - Confidence intervals need profile likelihoods.
  - 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")

print(pred_probs)
```

```
        Fish        Bird Invertebrate     Other     Reptile
1  0.6006304 0.051157366   0.07545645 0.24017062 0.032585176
2  0.6006304 0.051157366   0.07545645 0.24017062 0.032585176
3  0.6006304 0.051157366   0.07545645 0.24017062 0.032585176
4  0.6236286 0.110228530   0.02059329 0.18648582 0.059063749
5  0.6236286 0.110228530   0.02059329 0.18648582 0.059063749
6  0.6236286 0.110228530   0.02059329 0.18648582 0.059063749
7  0.5070764 0.079201241   0.10120786 0.26098463 0.051529843
8  0.5070764 0.079201241   0.10120786 0.26098463 0.051529843
9  0.5070764 0.079201241   0.10120786 0.26098463 0.051529843
10 0.5070764 0.079201241   0.10120786 0.26098463 0.051529843
```