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

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

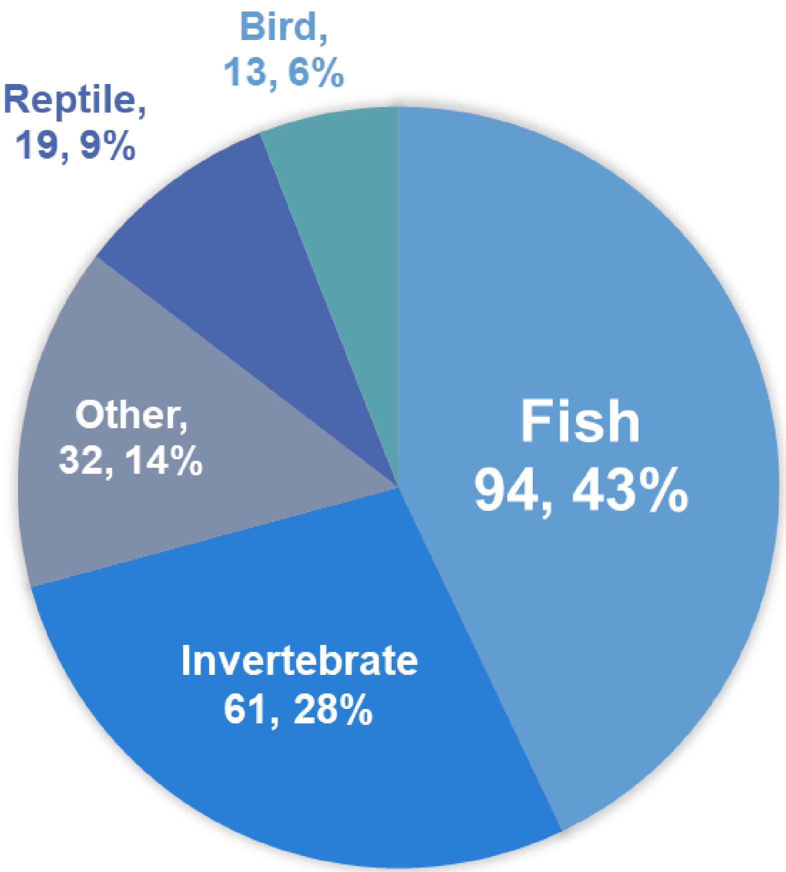
- Multinomial (probability that observation *i* has event *j*, and $j = 1, \ldots, m$):

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

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

Obs	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 $(p_1, p_2, \dots, p_m)$, then the generalized logits are the following:

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

- 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 – SAS

```
proc logistic data=Logistic.Gator plot(only) =
                                  oddsratio(range=clip);
   freq count;
   class lake(param=ref ref='George')
         size(param=ref ref='<= 2.3 meters')
         gender(param=ref ref='Male');
   model food(ref='Fish') = lake size gender /
                            link=glogit clodds=pl;
   title 'Model on Alligator Food Choice';
run;
quit;
```

Multinomial Logistic Regression – SAS

Model on Alligator Food Choice
The LOGISTIC Procedure

Model Information	
Data Set	LOGISTIC.GATOR
Response Variable	food
Number of Response Levels	5
Frequency Variable	count
Model	generalized logit
Optimization Technique	Newton-Raphson

Number of Observations Read	56
Number of Observations Used	56
Sum of Frequencies Read	219
Sum of Frequencies Used	219

Multinomial Logistic Regression – SAS

Response Profile		
Ordered Value	**food**	**Total Frequency**
1	Bird	13
2	Fish	94
3	Invertebrate	61
4	Other	32
5	Reptile	19

Logits modeled use food='Fish' as the reference category.

Multinomial Logistic Regression – SAS

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	612.363	585.865
SC	625.919	667.203
-2 Log L	604.363	537.865

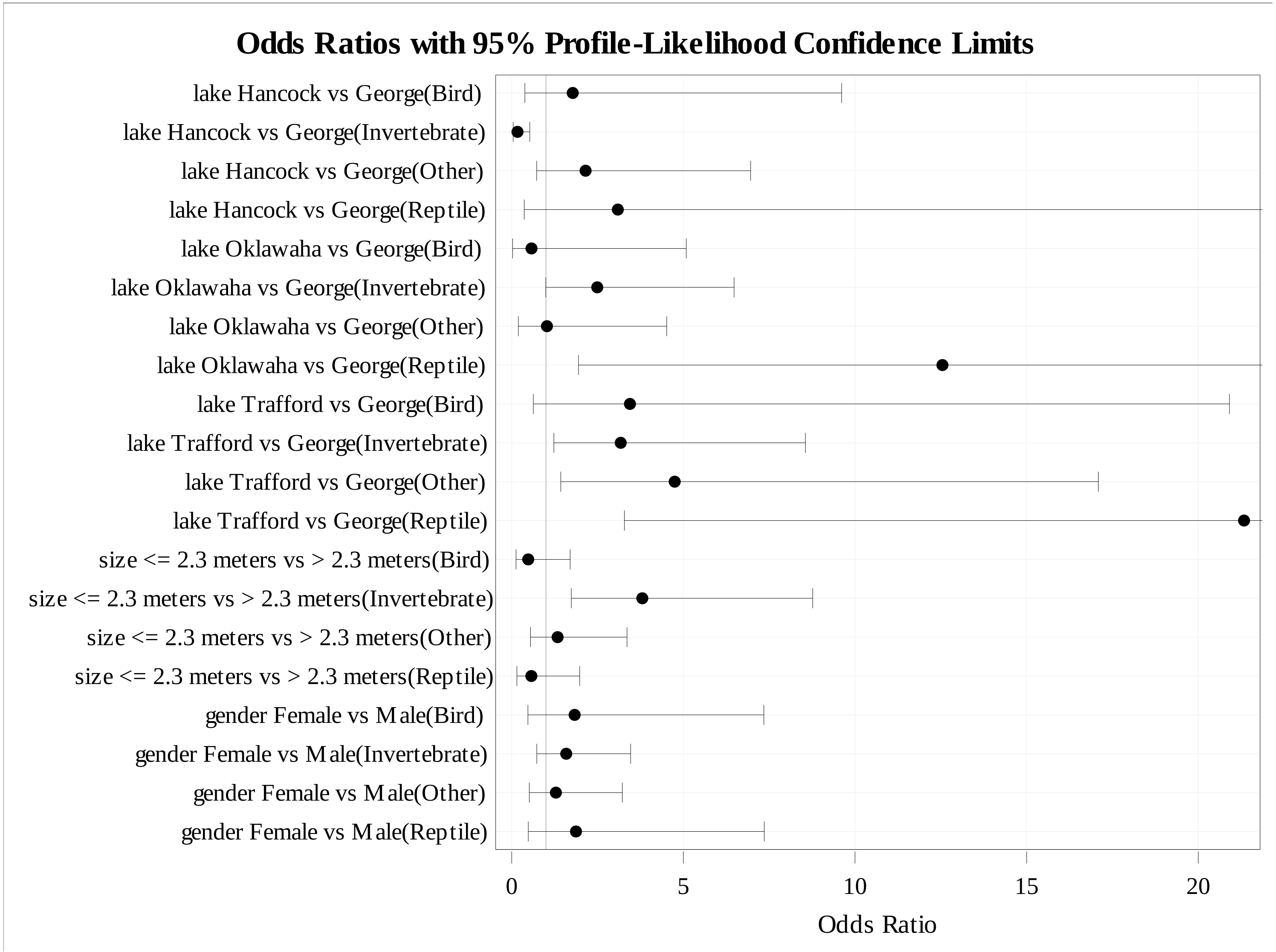
Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	66.4974	20	<.0001
Score	59.4616	20	<.0001
Wald	51.2336	20	0.0001

Multinomial Logistic Regression – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
lake	12	36.2293	0.0003
size	4	15.8873	0.0032
gender	4	2.1850	0.7018

Analysis of Maximum Likelihood Estimates							
Parameter		**food**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		**Bird**	1	-2.3083	0.7206	10.2623	0.0014
Intercept		**Invertebrate**	1	-1.6302	0.4262	14.6307	0.0001
Intercept		**Other**	1	-1.9739	0.5393	13.3966	0.0003
Intercept		**Reptile**	1	-3.4858	1.0699	10.6150	0.0011
lake	**Hancock**	**Bird**	1	0.5753	0.7952	0.5233	0.4694
lake	**Hancock**	**Invertebrate**	1	-1.7805	0.6232	8.1623	0.0043
lake	**Hancock**	**Other**	1	0.7666	0.5686	1.8179	0.1776
lake	**Hancock**	**Reptile**	1	1.1287	1.1925	0.8959	0.3439
lake	**Oklawaha**	**Bird**	1	-0.5504	1.2099	0.2069	0.6492
lake	**Oklawaha**	**Invertebrate**	1	0.9132	0.4761	3.6786	0.0551
lake	**Oklawaha**	**Other**	1	0.0261	0.7778	0.0011	0.9733
lake	**Oklawaha**	**Reptile**	1	2.5295	1.1218	5.0845	0.0241
lake	**Trafford**	**Bird**	1	1.2370	0.8661	2.0398	0.1532
lake	**Trafford**	**Invertebrate**	1	1.1558	0.4928	5.5013	0.0190
lake	**Trafford**	**Other**	1	1.5578	0.6257	6.1987	0.0128
lake	**Trafford**	**Reptile**	1	3.0603	1.1294	7.3423	0.0067
size	**<= 2.3 meters**	**Bird**	1	-0.7302	0.6523	1.2533	0.2629
size	**<= 2.3 meters**	**Invertebrate**	1	1.3363	0.4112	10.5606	0.0012
size	**<= 2.3 meters**	**Other**	1	0.2906	0.4599	0.3992	0.5275
size	**<= 2.3 meters**	**Reptile**	1	-0.5570	0.6466	0.7421	0.3890
gender	**Female**	**Bird**	1	0.6064	0.6888	0.7750	0.3787
gender	**Female**	**Invertebrate**	1	0.4630	0.3955	1.3701	0.2418
gender	**Female**	**Other**	1	0.2526	0.4663	0.2933	0.5881
gender	**Female**	**Reptile**	1	0.6275	0.6852	0.8387	0.3598

Multinomial Logistic Regression – SAS



Multinomial Logistic Regression – R

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

summary(glogit.model)
```

```
## Coefficients:
##             (Intercept) size> 2.3 meters lakeHancock lakeOklawaha
## Bird         -2.4321397        0.7300740   0.5754699  -0.55020075
## Invertebrate  0.1690702       -1.3361658  -1.7805555   0.91304120
## Other        -1.4309095       -0.2905697   0.7667093   0.02603021
## Reptile      -3.4161432        0.5571846   1.1296426   2.53024945
##             lakeTrafford genderMale
## Bird            1.237216 -0.6064035
## Invertebrate    1.155722 -0.4629388
## Other           1.557820 -0.2524299
## Reptile         3.061087 -0.6276217
##
## Std. Errors:
##             (Intercept) size> 2.3 meters lakeHancock lakeOklawaha
## Bird          0.7706720        0.6522657   0.7952303    1.2098680
## Invertebrate  0.3787475        0.4111827   0.6232075    0.4761068
## Other         0.5381162        0.4599317   0.5685673    0.7777958
## Reptile       1.0851582        0.6466092   1.1928075    1.1221413
##             lakeTrafford genderMale
## Bird           0.8661052  0.6888385
## Invertebrate   0.4927795  0.3955162
## Other          0.6256868  0.4663546
## Reptile        1.1297557  0.6852750
##
## Residual Deviance: 537.8655
## AIC: 585.8655
```

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:

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

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 – SAS

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals					
Effect	**food**	**Unit**	**Estimate**	**95% Confidence Limits**	
lake Hancock vs George	Bird	1.0000	1.778	0.384	9.612
lake Hancock vs George	Invertebrate	1.0000	0.169	0.044	0.528
lake Hancock vs George	Other	1.0000	2.152	0.727	6.960
lake Hancock vs George	Reptile	1.0000	3.092	0.364	65.177
lake Oklawaha vs George	Bird	1.0000	0.577	0.027	5.084
lake Oklawaha vs George	Invertebrate	1.0000	2.492	0.993	6.479
lake Oklawaha vs George	Other	1.0000	1.026	0.194	4.516
lake Oklawaha vs George	Reptile	1.0000	12.547	1.945	248.047
lake Trafford vs George	Bird	1.0000	3.445	0.631	20.908
lake Trafford vs George	Invertebrate	1.0000	3.177	1.228	8.557
lake Trafford vs George	Other	1.0000	4.748	1.431	17.088
lake Trafford vs George	Reptile	1.0000	21.334	3.282	426.076
size > 2.3 meters vs <= 2.3 meters	Bird	1.0000	2.076	0.588	7.943
size > 2.3 meters vs <= 2.3 meters	Invertebrate	1.0000	0.263	0.114	0.576
size > 2.3 meters vs <= 2.3 meters	Other	1.0000	0.748	0.298	1.827
size > 2.3 meters vs <= 2.3 meters	Reptile	1.0000	1.745	0.505	6.565
gender Female vs Male	Bird	1.0000	1.834	0.472	7.345
gender Female vs Male	Invertebrate	1.0000	1.589	0.731	3.464
gender Female vs Male	Other	1.0000	1.287	0.512	3.222
gender Female vs Male	Reptile	1.0000	1.873	0.483	7.358

Interpretation – R

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

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

```
proc logistic data=Logistic.Gator plot(only)=oddsratio(range=clip);
   freq count;
   class lake(param=ref ref='George')
         size(param=ref ref='<= 2.3 meters')
         gender(param=ref ref='Male');
   model food(ref='Fish') = lake size gender / link=glogit clodds=pl;
   output out=pred predprobs=I;
run;
quit;

proc print data=pred;
run;

proc freq data=pred;
   weight count;
   tables _FROM_*_INTO_;
run;
```

Predicted Probabilities – SAS

Obs	size	food	lake	gender	count	_FROM_	_INTO_	IP_Bird	IP_Fish	IP_Inv.	IP_Other	IP_Rep.
1	<= 2.3 meters	Fish	Hancock	Male	7	Fish	Fish	0.05115	0.60065	0.07546	0.24016	0.03259
2	<= 2.3 meters	Invertebrate	Hancock	Male	1	Invertebrate	Fish	0.05115	0.60065	0.07546	0.24016	0.03259
3	<= 2.3 meters	Other	Hancock	Male	5	Other	Fish	0.05115	0.60065	0.07546	0.24016	0.03259
4	> 2.3 meters	Fish	Hancock	Male	4	Fish	Fish	0.11023	0.62365	0.02059	0.18647	0.05906
5	> 2.3 meters	Bird	Hancock	Male	1	Bird	Fish	0.11023	0.62365	0.02059	0.18647	0.05906
6	> 2.3 meters	Other	Hancock	Male	2	Other	Fish	0.11023	0.62365	0.02059	0.18647	0.05906
7	<= 2.3 meters	Fish	Hancock	Female	16	Fish	Fish	0.07919	0.50708	0.10121	0.26100	0.05153
8	<= 2.3 meters	Invertebrate	Hancock	Female	3	Invertebrate	Fish	0.07919	0.50708	0.10121	0.26100	0.05153

⋮

Predicted Probabilities – SAS

Obs	size	food	lake	gender	count	_FROM_	_INTO_	IP_Bird	IP_Fish	IP_Inv.	IP_Other	IP_Rep.
1	<= 2.3 meters	Fish	Hancock	Male	7	Fish	Fish	0.05115	0.60065	0.07546	0.24016	0.03259
2	<= 2.3 meters	Invertebrate	Hancock	Male	1	Invertebrate	Fish	0.05115	0.60065	0.07546	0.24016	0.03259
3	<= 2.3 meters	Other	Hancock	Male	5	Other	Fish	0.05115	0.60065	0.07546	0.24016	0.03259
4	> 2.3 meters	Fish	Hancock	Male	4	Fish	Fish	0.11023	0.62365	0.02059	0.18647	0.05906
5	> 2.3 meters	Bird	Hancock	Male	1	Bird	Fish	0.11023	0.62365	0.02059	0.18647	0.05906
6	> 2.3 meters	Other	Hancock	Male	2	Other	Fish	0.11023	0.62365	0.02059	0.18647	0.05906
7	<= 2.3 meters	Fish	Hancock	Female	16	Fish	Fish	0.07919	0.50708	0.10121	0.26100	0.05153
8	<= 2.3 meters	Invertebrate	Hancock	Female	3	Invertebrate	Fish	0.07919	0.50708	0.10121	0.26100	0.05153

⋮

Table of _FROM_ by _INTO_				
FROM	_INTO_			
	Fish	Invertebrate	Reptile	Total
Bird	12 5.48 92.31 7.50	1 0.46 7.69 1.72	0 0.00 0.00 0.00	13 5.94
Fish	81 36.99 86.17 50.63	13 5.94 13.83 22.41	0 0.00 0.00 0.00	94 42.92
Invertebrate	29 13.24 47.54 18.13	31 14.16 50.82 53.45	1 0.46 1.64 100.00	61 27.85
Other	23 10.50 71.88 14.38	9 4.11 28.13 15.52	0 0.00 0.00 0.00	32 14.61
Reptile	15 6.85 78.95 9.38	4 1.83 21.05 6.90	0 0.00 0.00 0.00	19 8.68
Total	160 73.06	58 26.48	1 0.46	219 100.00

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
```

⋮