432 Class 21

https://thomaselove.github.io/432-2025

2025-04-01

## Today’s Topic

**Multinomial Logistic Regression: An Introduction**

* Two Examples about Alligators
* Fitting Multinomial Logistic Regression Models

Chapter 28 of the Course Notes describes an additional example discussing this material.

## Today’s R Setup

knitr::opts\_chunk$set(comment=NA)  
  
library(janitor); library(naniar)  
library(here); library(conflicted)  
library(broom)  
library(mosaic)  
library(gt)  
library(nnet)  
library(rms)  
library(easystats)  
library(tidyverse)  
  
conflicts\_prefer(base::mean, base::max, dplyr::filter,   
 dplyr::select, dplyr::summarize, janitor::clean\_names)  
  
theme\_set(theme\_bw())

## Regression on Multi-categorical Outcomes

Suppose we have a nominal, multi-categorical outcome of interest. Multinomial (also called multicategory or polychotomous) logistic regression models describe the odds of response in one category instead of another.

* Such models pair each outcome category with a baseline category, the choice of which is arbitrary.
* The model consists of J-1 logit equations (for an outcome with J categories) with separate parameters for each.

# A small example: gator1 on Alligator Food Choices

## Today’s Data

Today’s data relates to alligator food choices. We’ll actually work with two different data sets.

In each case, we’ll read in the data, and set some key variables to be factors and, if needed, actively select the baseline category.

## gator1: Alligator Food Choice

The gator1 data are from a study by the Florida Game and Fresh Water Fish Commission of factors influencing the primary food choice of alligators[[1]](#footnote-24).

We’ll be trying to predict primary food choice using the alligator’s length.

## gator1 has data on 59 alligators

* length (in meters)
* choice = primary food type, in volume, found in the alligator’s stomach, specifically…
  + Fish,
  + Invertebrates (mostly apple snails, aquatic insects and crayfish, and I’ll abbreviate this category as Inverts in what follows)
  + Other (which includes reptiles, amphibians, mammals, plant material and stones or other debris.)

## Ingesting the gator1 data

gator1 <- read\_csv(here("c21/data/gator1.csv"),   
 show\_col\_types = FALSE) |>  
 mutate(choice = fct\_relevel(factor(choice), "Other"),  
 choice = fct\_recode(choice,   
 "Inverts" = "Invertebrates"))  
  
gator1

# A tibble: 59 × 3  
 id length choice   
 <dbl> <dbl> <fct>   
 1 1 1.24 Inverts  
 2 2 1.3 Inverts  
 3 3 1.3 Inverts  
 4 4 1.32 Fish   
 5 5 1.32 Fish   
 6 6 1.4 Fish   
 7 7 1.42 Inverts  
 8 8 1.42 Fish   
 9 9 1.45 Inverts  
10 10 1.45 Other   
# ℹ 49 more rows

## Summarizing length by choice

Note that I’ve set up choice so that “Other” is the baseline (first) category, while ingesting the data.

favstats(length ~ choice, data = gator1) |>  
 gt() |> fmt\_number(min:sd, decimals = 2) |>  
 tab\_options(table.font.size = 20)

| choice | min | Q1 | median | Q3 | max | mean | sd | n | missing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Other | 1.45 | 1.71 | 2.17 | 3.26 | 3.68 | 2.42 | 0.88 | 8 | 0 |
| Fish | 1.32 | 1.74 | 2.31 | 2.73 | 3.89 | 2.36 | 0.76 | 31 | 0 |
| Inverts | 1.24 | 1.47 | 1.61 | 1.78 | 2.72 | 1.66 | 0.33 | 20 | 0 |

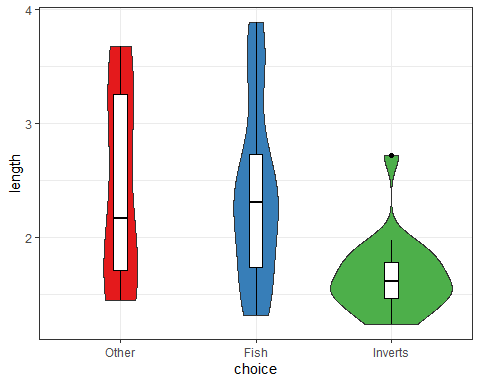
There is no missing data in gator1.

n\_miss(gator1)

[1] 0

## Length and Primary Food Choice

ggplot(gator1, aes(x = choice, y = length, fill = choice)) +  
 geom\_violin() +  
 geom\_boxplot(fill = "white", col = "black", width = 0.1) +  
 scale\_fill\_brewer(palette = "Set1") + guides(fill = "none")



## Re-scaling Length

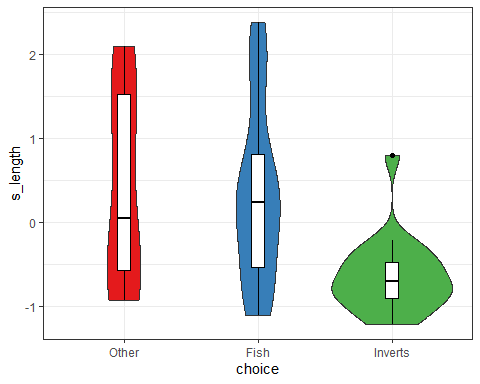
Multinomial models work best if the quantitative predictors are rescaled to have mean 0 and standard deviation 1. Let’s do that for our length variable.

gator1 <- gator1 |>   
 mutate(s\_length = scale(length, center = TRUE, scale = TRUE))  
  
df\_stats(~ s\_length + length, data = gator1) |>  
 gt() |> fmt\_number(columns = min:sd, decimals = 2) |>   
 tab\_options(table.font.size = 20)

| response | min | Q1 | median | Q3 | max | mean | sd | n | missing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| s\_length | -1.21 | -0.75 | -0.38 | 0.43 | 2.39 | 0.00 | 1.00 | 59 | 0 |
| length | 1.24 | 1.58 | 1.85 | 2.45 | 3.89 | 2.13 | 0.74 | 59 | 0 |

## Scaled Length and Primary Food Choice

ggplot(gator1, aes(x = choice, y = s\_length, fill = choice)) +  
 geom\_violin() +  
 geom\_boxplot(fill = "white", col = "black", width = 0.1) +  
 scale\_fill\_brewer(palette = "Set1") + guides(fill = "none")



# Fitting a Multinomial Logistic Regression

## Multinomial Logistic Regression Fit

* “Other” is the first (reference) level for choice

gator1 |> tabyl(choice)

choice n percent  
 Other 8 0.1355932  
 Fish 31 0.5254237  
 Inverts 20 0.3389831

Let’s try using multinom() from the nnet package

try1 <- multinom(choice ~ s\_length, data=gator1)

# weights: 9 (4 variable)  
initial value 64.818125   
iter 10 value 49.170630  
final value 49.170622   
converged

## Looking over the first try

try1

Call:  
multinom(formula = choice ~ s\_length, data = gator1)  
  
Coefficients:  
 (Intercept) s\_length  
Fish 1.3831224 -0.08123412  
Inverts 0.4450626 -1.81819507  
  
Residual Deviance: 98.34124   
AIC: 106.3412

Our R output suggests the following models:

* log odds of Fish rather than Other = 1.38 - 0.081 s\_length
* log odds of Inverts. rather than Other = 0.45 - 1.818 s\_length

## Estimating Response Probabilities

We can express the multinomial logistic regression model directly in terms of outcome probabilities:

Our models contrast “Fish” and “Invertebrates” to “Other” as the reference category.

## Estimating Response Probabilities

In our try1, we have:

* log odds of Fish rather than Other = 1.38 - 0.081 s\_length
* log odds of Inverts. rather than Other = 0.45 - 1.818 s\_length
* For the reference category we use and so that for that category (here, Other.)

## try1 Response Probabilities

Our estimates (which will sum to 1) are:

## Making a Prediction with try1

For an alligator of average length (s\_length = 0), for instance, we have:

$$
denominator = {1 + exp(0.45 - 1.818 (0)) + exp(1.38 - 0.081 (0))} = 6.543 \\
Pr(Fish) = \frac{exp(1.38 - 0.081 (0))}{6.543} = \frac{3.975}{6.543} \approx 0.61\\
Pr(Invert.) = \frac{exp(0.45 - 1.818 (0)}{6.543} = \frac{1.568}{6.543} \approx 0.24\\
Pr(Other) = \frac{1}{6.543} \approx 0.15
$$

## Model Parameters (raw)

model\_parameters(try1, pretty\_names = FALSE, ci = 0.90)

# Response level: fish  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
-------------------------------------------------------------  
(Intercept) | 1.38 | 0.42 | [ 0.69, 2.08] | 3.28 | 0.001  
s\_length | -0.08 | 0.38 | [-0.71, 0.55] | -0.21 | 0.831  
  
# Response level: inverts  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
--------------------------------------------------------------  
(Intercept) | 0.45 | 0.54 | [-0.45, 1.34] | 0.82 | 0.413  
s\_length | -1.82 | 0.66 | [-2.91, -0.73] | -2.74 | 0.006

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald normal distribution approximation.

The model has a log- or logit-link. Consider using `exponentiate =  
 TRUE` to interpret coefficients as ratios.

## Interpreting the raw coefficients

# Response level: fish  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
-------------------------------------------------------------  
(Intercept) | 1.38 | 0.42 | [ 0.69, 2.08] | 3.28 | 0.001  
s\_length | -0.08 | 0.38 | [-0.71, 0.55] | -0.21 | 0.831

The log odds of a gator’s choice being fish rather than Other would be 1.38 - 0.08 s\_length, or, for a gator of average length, the log odds would be 1.38, with 90% CI (0.69, 2.08).

## Model Parameters (exponentiated)

model\_parameters(try1, exponentiate = TRUE, pretty\_names = FALSE, ci = 0.90)

# Response level: fish  
  
Parameter | Odds Ratio | SE | 90% CI | z | p  
--------------------------------------------------------------  
(Intercept) | 3.99 | 1.68 | [1.99, 7.98] | 3.28 | 0.001  
s\_length | 0.92 | 0.35 | [0.49, 1.73] | -0.21 | 0.831  
  
# Response level: inverts  
  
Parameter | Odds Ratio | SE | 90% CI | z | p  
--------------------------------------------------------------  
(Intercept) | 1.56 | 0.85 | [0.64, 3.82] | 0.82 | 0.413  
s\_length | 0.16 | 0.11 | [0.05, 0.48] | -2.74 | 0.006

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald normal distribution approximation.

## Interpreting the odds ratios

# Response level: inverts  
  
Parameter | Odds Ratio | SE | 90% CI | z | p  
--------------------------------------------------------------  
(Intercept) | 1.56 | 0.85 | [0.64, 3.82] | 0.82 | 0.413  
s\_length | 0.16 | 0.11 | [0.05, 0.48] | -2.74 | 0.006

The odds ratio for a gator’s choice being invertebrates rather than Other would be 1.56 - 0.16 s\_length, or, for a gator of length one standard deviation greater than the mean (so s\_length = 1), the odds ratio would be 1.56 + 0.16, or 1.72.

## Building a Classification Table

table(predict(try1), gator1$choice, dnn = c("predicted", "actual"))

actual  
predicted Other Fish Inverts  
 Other 0 0 0  
 Fish 5 23 7  
 Inverts 3 8 13

* Our try1 model predicts no choices of “Other”.
* try1 makes a correct prediction 23 + 13 = 36 times out of 59 alligators, or 61% of the time.

## Predicted Probabilities from try1

try1\_fits <-   
 predict(try1, newdata = gator1, type = "probs")  
  
gator1\_try1 <- cbind(gator1, try1\_fits)  
  
head(gator1\_try1); tail(gator1\_try1)

id length choice s\_length Other Fish Inverts  
1 1 1.24 Inverts -1.2073918 0.05150350 0.2265246 0.7219719  
2 2 1.30 Inverts -1.1260256 0.05727547 0.2502516 0.6924730  
3 3 1.30 Inverts -1.1260256 0.05727547 0.2502516 0.6924730  
4 4 1.32 Fish -1.0989035 0.05928343 0.2584548 0.6822618  
5 5 1.32 Fish -1.0989035 0.05928343 0.2584548 0.6822618  
6 6 1.40 Fish -0.9904153 0.06770639 0.2925860 0.6397076

id length choice s\_length Other Fish Inverts  
54 54 3.56 Fish 1.938768 0.2246048 0.7650725 0.010322706  
55 55 3.58 Fish 1.965890 0.2250956 0.7650569 0.009847480  
56 56 3.66 Fish 2.074379 0.2270197 0.7648265 0.008153722  
57 57 3.68 Other 2.101501 0.2274919 0.7647306 0.007777531  
58 58 3.71 Fish 2.142184 0.2281942 0.7645605 0.007245290  
59 59 3.89 Fish 2.386282 0.2322843 0.7629839 0.004731762

## Tabulating Response Probabilities

gator1\_try1 |> group\_by(choice) |>  
 summarise(mean(Other), mean(Fish), mean(Inverts))

# A tibble: 3 × 4  
 choice `mean(Other)` `mean(Fish)` `mean(Inverts)`  
 <fct> <dbl> <dbl> <dbl>  
1 Other 0.155 0.580 0.265  
2 Fish 0.155 0.590 0.255  
3 Inverts 0.0973 0.404 0.499

## Pivot the Wide data to make it longer

We need to have this data organized differently in order to build the plot I want to build.

gator1\_try1long <-   
 pivot\_longer(gator1\_try1,   
 cols = c("Other", "Fish", "Inverts"),  
 names\_to = "preference",  
 values\_to = "probability") |>  
 mutate(preference = factor(preference))

## What does this pivoting accomplish?

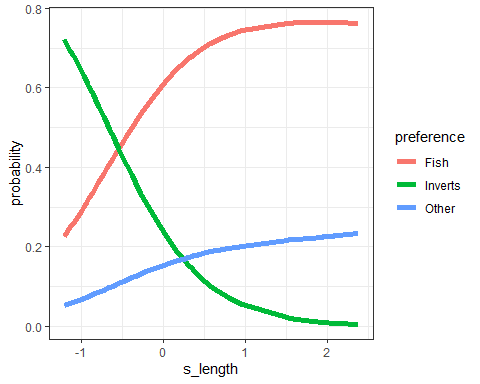
gator1\_try1long

# A tibble: 177 × 6  
 id length choice s\_length[,1] preference probability  
 <dbl> <dbl> <fct> <dbl> <fct> <dbl>  
 1 1 1.24 Inverts -1.21 Other 0.0515  
 2 1 1.24 Inverts -1.21 Fish 0.227   
 3 1 1.24 Inverts -1.21 Inverts 0.722   
 4 2 1.3 Inverts -1.13 Other 0.0573  
 5 2 1.3 Inverts -1.13 Fish 0.250   
 6 2 1.3 Inverts -1.13 Inverts 0.692   
 7 3 1.3 Inverts -1.13 Other 0.0573  
 8 3 1.3 Inverts -1.13 Fish 0.250   
 9 3 1.3 Inverts -1.13 Inverts 0.692   
10 4 1.32 Fish -1.10 Other 0.0593  
# ℹ 167 more rows

## try1 Response Probabilities

ggplot(gator1\_try1long, aes(x = s\_length, y = probability,   
 col = preference)) +  
 geom\_line(size = 2) + scale\_fill\_brewer(palette = "Set1")

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.



## try1 model\_performance() and glance()

model\_performance(try1)

Can't calculate log-loss.

Can't calculate proper scoring rules for ordinal, multinomial or  
 cumulative link models.

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
---------------------------------------------------------------  
106.341 | 107.082 | 114.651 | 0.146 | 0.129 | 0.408 | 1.337

glance(try1)

# A tibble: 1 × 4  
 edf deviance AIC nobs  
 <dbl> <dbl> <dbl> <int>  
1 4 98.3 106. 59

## try1 model summary

summary(try1)

Call:  
multinom(formula = choice ~ s\_length, data = gator1)  
  
Coefficients:  
 (Intercept) s\_length  
Fish 1.3831224 -0.08123412  
Inverts 0.4450626 -1.81819507  
  
Std. Errors:  
 (Intercept) s\_length  
Fish 0.4221186 0.3812914  
Inverts 0.5439490 0.6634395  
  
Residual Deviance: 98.34124   
AIC: 106.3412

## Compare to intercept-only model

Compare try1 to the null model with only an intercept (try0)

try0 <- multinom(choice ~ 1, data=gator1)

# weights: 6 (2 variable)  
initial value 64.818125   
final value 57.570928   
converged

AIC(try0, try1)

df AIC  
try0 2 119.1419  
try1 4 106.3412

BIC(try0, try1)

df BIC  
try0 2 123.2969  
try1 4 114.6514

Does the inclusion of s\_length produce a meaningfully better fit to the data than simply fitting an intercept?

## ANOVA for try0 vs. try1

* If you’d prefer a hypothesis testing approach, use anova…

anova(try0, try1)

Likelihood ratio tests of Multinomial Models  
  
Response: choice  
 Model Resid. df Resid. Dev Test Df LR stat. Pr(Chi)  
1 1 116 115.14186   
2 s\_length 114 98.34124 1 vs 2 2 16.80061 0.0002247985

Does the inclusion of s\_length produce a meaningfully better fit to the data than simply fitting an intercept?

## Wald Z tests for individual predictors

By default, tidy exponentiates multinomial coefficients…

tidy(try1) |>   
 gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| y.level | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| Fish | (Intercept) | 1.383 | 0.422 | 3.277 | 0.001 |
| Fish | s\_length | -0.081 | 0.381 | -0.213 | 0.831 |
| Inverts | (Intercept) | 0.445 | 0.544 | 0.818 | 0.413 |
| Inverts | s\_length | -1.818 | 0.663 | -2.741 | 0.006 |

## Compare the models (parameters)

compare\_parameters(try1, try0, pretty\_names = FALSE, ci = 0.90)

Parameter | try1 | try0  
-------------------------------------------------------  
(Intercept) | 1.38 ( 0.68, 2.09) | 1.35 (0.69, 2.02)  
(Intercept) | 1.38 ( 0.68, 2.09) | 0.92 (0.22, 1.62)  
(Intercept) | 0.45 (-0.46, 1.36) | 1.35 (0.69, 2.02)  
(Intercept) | 0.45 (-0.46, 1.36) | 0.92 (0.22, 1.62)  
s length | -0.08 (-0.72, 0.56) |   
s\_length | -1.82 (-2.93, -0.71) |   
-------------------------------------------------------  
Observations | 59 | 59

* Notice that this isn’t very helpful.

## Using model\_parameters() on try0

model\_parameters(try0, pretty\_names = FALSE, ci = 0.90)

# Response level: fish  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
------------------------------------------------------------  
(Intercept) | 1.35 | 0.40 | [0.70, 2.01] | 3.42 | < .001  
  
# Response level: inverts  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
-----------------------------------------------------------  
(Intercept) | 0.92 | 0.42 | [0.23, 1.60] | 2.19 | 0.028

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald normal distribution approximation.

## Using model\_parameters() on try1

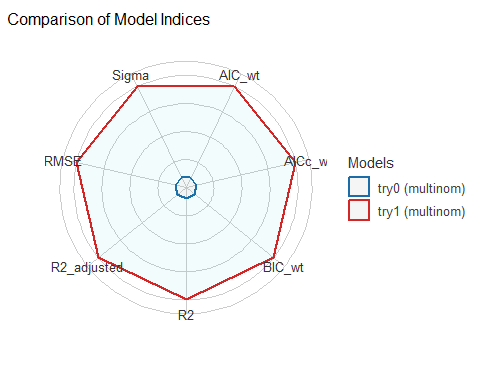
model\_parameters(try1, pretty\_names = FALSE, ci = 0.90)

# Response level: fish  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
-------------------------------------------------------------  
(Intercept) | 1.38 | 0.42 | [ 0.69, 2.08] | 3.28 | 0.001  
s\_length | -0.08 | 0.38 | [-0.71, 0.55] | -0.21 | 0.831  
  
# Response level: inverts  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
--------------------------------------------------------------  
(Intercept) | 0.45 | 0.54 | [-0.45, 1.34] | 0.82 | 0.413  
s\_length | -1.82 | 0.66 | [-2.91, -0.73] | -2.74 | 0.006

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald normal distribution approximation.

## Compare the models (performance)

plot(compare\_performance(try1, try0))



# Working with a larger example: gator2

## gator2 describes 219 alligators

The gator2.csv data[[2]](#footnote-71) considers the stomach contents of 219 alligators, aggregated into 5 categories by primary food choice:

* fish
* invertebrates
* reptiles
* birds
* other (including amphibians, plants, household pets, stones, and debris)

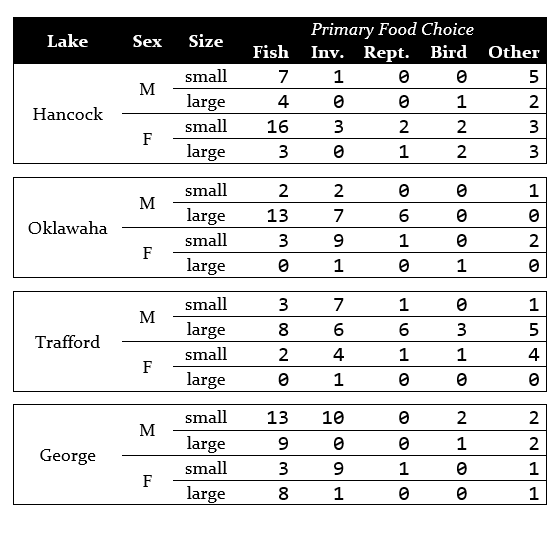
## gator2 has additional predictors

The 219 alligators are also categorized by:

* sex (m or f), and
* by length (< 2.3 and 2.3 meters) and
* by which of four lakes they were captured in (Hancock, Oklawaha, Trafford or George.)

We’ll use as our baseline: fish as a choice, sex as m, s\_length >= 2.3 and lake george.

See the Table on next slide.



## Model Setup

We’ll use Fish as the baseline, so our regression equations take the form

for .

## How many coefficients do we estimate?

* We have six coefficients to estimate in each of four logit equations (one each for ) so there are 24 parameters to estimate.

## Ingesting the gator2 data

gator2 <- read\_csv(here("c21/data/gator2.csv"), show\_col\_types = FALSE) |>  
 mutate(across(where(is\_character), as\_factor),  
 id = as.character(id))

We re-level the factors to put our reference categories first.

gator2 <- gator2 |>  
 mutate(food = fct\_relevel(food, "fish", "invert",   
 "rep", "bird", "other"),  
 size = fct\_relevel(size, ">=2.3"),  
 gender = fct\_relevel(gender, "m"),  
 lake = fct\_relevel(lake, "george"))  
  
summary(gator2 |> select(-id))

food size gender lake   
 fish :94 >=2.3: 95 m:130 george :63   
 invert:61 <2.3 :124 f: 89 hancock :55   
 rep :19 oklawaha:48   
 bird :13 trafford:53   
 other :32

## The Models We Will Fit

We’ll fit (using multinom() from nnet)

* A *saturated* model, including all three predictors and all two-way interactions and the three-way interaction
* A *null* model, with the intercept alone
* Simple logistic regression models for each of the three predictors as a main effect alone
* The model including both L(ake) and S(ize) but nothing else
* The model including all three predictors as main effects, but no interactions

## The Models We Will Fit (code only)

options(contrasts=c("contr.treatment", "contr.poly"))  
fit\_SAT <- multinom(food ~ lake\*size\*gender, data=gator2) # saturated  
fit\_1 <-multinom(food~1,data=gator2) # null  
fit\_G <-multinom(food~gender,data=gator2) # G  
fit\_L <-multinom(food~lake,data=gator2) # L  
fit\_S <-multinom(food~size,data=gator2) # S  
fit\_LS <-multinom(food~lake+size,data=gator2) # L+S  
fit\_GLS <-multinom(food~gender+lake+size,data=gator2) # G+L+S

## What You’ll See When Fitting the models

options(contrasts=c("contr.treatment", "contr.poly"))  
fit\_SAT <- multinom(food ~ lake\*size\*gender, data=gator2)

# weights: 85 (64 variable)  
initial value 352.466903   
iter 10 value 261.200857  
iter 20 value 245.788420  
iter 30 value 244.090612  
iter 40 value 243.812122  
iter 50 value 243.801212  
final value 243.800899   
converged

and we’ll see something similar for each of the other models…

fit\_1<-multinom(food~1,data=gator2) # null

# weights: 10 (4 variable)  
initial value 352.466903   
final value 302.181462   
converged

fit\_G<-multinom(food~gender,data=gator2) # G

# weights: 15 (8 variable)  
initial value 352.466903   
iter 10 value 301.286320  
final value 301.129428   
converged

fit\_L<-multinom(food~lake,data=gator2) # L

# weights: 25 (16 variable)  
initial value 352.466903   
iter 10 value 281.030560  
iter 20 value 280.583926  
final value 280.583844   
converged

fit\_S<-multinom(food~size,data=gator2) # S

# weights: 15 (8 variable)  
initial value 352.466903   
iter 10 value 294.670879  
final value 294.606678   
converged

fit\_LS<-multinom(food~lake+size,data=gator2) # L+S

# weights: 30 (20 variable)  
initial value 352.466903   
iter 10 value 271.607785  
iter 20 value 270.046051  
final value 270.040140   
converged

fit\_GLS<-multinom(food~gender+lake+size,data=gator2) # G+L+S

# weights: 35 (24 variable)  
initial value 352.466903   
iter 10 value 271.128034  
iter 20 value 268.936875  
final value 268.932741   
converged

## The Intercept only Model fit\_1

summary(fit\_1)

Call:  
multinom(formula = food ~ 1, data = gator2)  
  
Coefficients:  
 (Intercept)  
invert -0.4324211  
rep -1.5988558  
bird -1.9783458  
other -1.0775589  
  
Std. Errors:  
 (Intercept)  
invert 0.1644133  
rep 0.2515350  
bird 0.2959078  
other 0.2046663  
  
Residual Deviance: 604.3629   
AIC: 612.3629

## Tidying this summary

tidy(fit\_1, exponentiate = FALSE) |>   
 gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| y.level | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| invert | (Intercept) | -0.432 | 0.164 | -2.630 | 0.009 |
| rep | (Intercept) | -1.599 | 0.252 | -6.356 | 0.000 |
| bird | (Intercept) | -1.978 | 0.296 | -6.686 | 0.000 |
| other | (Intercept) | -1.078 | 0.205 | -5.265 | 0.000 |

glance(fit\_1) |> gt() |>   
 tab\_options(table.font.size = 20)

| edf | deviance | AIC | nobs |
| --- | --- | --- | --- |
| 4 | 604.3629 | 612.3629 | 219 |

## Size only model

tidy(fit\_S, exponentiate = FALSE) |> gt() |>   
 fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| y.level | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| invert | (Intercept) | -1.034 | 0.291 | -3.553 | 0.000 |
| invert | size<2.3 | 0.949 | 0.357 | 2.659 | 0.008 |
| rep | (Intercept) | -1.242 | 0.315 | -3.944 | 0.000 |
| rep | size<2.3 | -0.858 | 0.535 | -1.604 | 0.109 |
| bird | (Intercept) | -1.727 | 0.384 | -4.502 | 0.000 |
| bird | size<2.3 | -0.555 | 0.606 | -0.916 | 0.360 |
| other | (Intercept) | -1.242 | 0.315 | -3.944 | 0.000 |
| other | size<2.3 | 0.294 | 0.415 | 0.709 | 0.478 |

## Size only model

glance(fit\_S) |> gt() |> tab\_options(table.font.size = 20)

| edf | deviance | AIC | nobs |
| --- | --- | --- | --- |
| 8 | 589.2134 | 605.2134 | 219 |

## Gender only model

tidy(fit\_G, exponentiate = FALSE) |> gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| y.level | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| invert | (Intercept) | -0.581 | 0.217 | -2.673 | 0.008 |
| invert | genderf | 0.358 | 0.334 | 1.072 | 0.284 |
| rep | (Intercept) | -1.513 | 0.306 | -4.937 | 0.000 |
| rep | genderf | -0.251 | 0.538 | -0.467 | 0.641 |
| bird | (Intercept) | -2.132 | 0.400 | -5.332 | 0.000 |
| bird | genderf | 0.368 | 0.596 | 0.618 | 0.537 |
| other | (Intercept) | -1.187 | 0.269 | -4.409 | 0.000 |
| other | genderf | 0.271 | 0.415 | 0.652 | 0.514 |

## Gender only model

glance(fit\_G) |> gt() |> tab\_options(table.font.size = 20)

| edf | deviance | AIC | nobs |
| --- | --- | --- | --- |
| 8 | 602.2589 | 618.2589 | 219 |

## Lake only model (part 1 of 2)

tidy(fit\_L, exponentiate = FALSE) |> slice(1:10) |>   
 gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| y.level | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| invert | (Intercept) | -0.501 | 0.283 | -1.767 | 0.077 |
| invert | lakehancock | -1.514 | 0.603 | -2.511 | 0.012 |
| invert | lakeoklawaha | 0.555 | 0.434 | 1.278 | 0.201 |
| invert | laketrafford | 0.826 | 0.461 | 1.791 | 0.073 |
| rep | (Intercept) | -3.496 | 1.015 | -3.445 | 0.001 |
| rep | lakehancock | 1.194 | 1.182 | 1.010 | 0.312 |
| rep | lakeoklawaha | 2.552 | 1.108 | 2.302 | 0.021 |
| rep | laketrafford | 3.011 | 1.110 | 2.713 | 0.007 |
| bird | (Intercept) | -2.398 | 0.603 | -3.976 | 0.000 |
| bird | lakehancock | 0.607 | 0.773 | 0.785 | 0.432 |

## Lake only model (part 2 of 2)

tidy(fit\_L, exponentiate = FALSE) |> slice(11:16) |>   
 gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| y.level | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| bird | lakeoklawaha | -0.492 | 1.191 | -0.413 | 0.680 |
| bird | laketrafford | 1.220 | 0.831 | 1.468 | 0.142 |
| other | (Intercept) | -1.705 | 0.444 | -3.841 | 0.000 |
| other | lakehancock | 0.869 | 0.554 | 1.567 | 0.117 |
| other | lakeoklawaha | -0.087 | 0.765 | -0.114 | 0.910 |
| other | laketrafford | 1.443 | 0.611 | 2.359 | 0.018 |

glance(fit\_L) |> gt() |> tab\_options(table.font.size = 20)

| edf | deviance | AIC | nobs |
| --- | --- | --- | --- |
| 16 | 561.1677 | 593.1677 | 219 |

## The Saturated Model

fit\_SAT

Call:  
multinom(formula = food ~ lake \* size \* gender, data = gator2)  
  
Coefficients:  
 (Intercept) lakehancock lakeoklawaha laketrafford size<2.3 genderf  
invert -22.731435 -7.6997047 22.11245 22.443706 22.4691578 20.6519880  
rep -29.030622 4.5446124 28.25748 28.742943 -2.1497924 -1.5018889  
bird -2.196705 0.8106289 -18.76043 1.215771 0.3248760 -17.2683965  
other -1.503884 0.8107459 -25.23128 1.033839 -0.3675892 -0.5756885  
 lakehancock:size<2.3 lakeoklawaha:size<2.3 laketrafford:size<2.3  
invert 6.0160287 -21.85028 -21.3342850  
rep -15.0175978 -17.43950 1.3387310  
bird -22.8201143 -25.18859 -25.8829682  
other 0.7242536 26.40938 -0.2614093  
 lakehancock:genderf lakeoklawaha:genderf laketrafford:genderf  
invert -3.946342 4.226498 25.465169  
rep 24.889170 -13.585689 -18.078274  
bird 18.248790 62.485154 16.978562  
other 1.268734 -1.758853 -7.586589  
 size<2.3:genderf lakehancock:size<2.3:genderf  
invert -19.2913107 2.857688  
rep 31.5836415 -15.396895  
bird 0.6638064 20.157737  
other 1.3479978 -3.378585  
 lakeoklawaha:size<2.3:genderf laketrafford:size<2.3:genderf  
invert -4.488351 -26.979637  
rep 2.767887 -11.597631  
bird -24.265617 25.472087  
other 1.274620 8.606604  
  
Residual Deviance: 487.6018   
AIC: 615.6018

## Building a Model Comparison Table

For a model fitX, we find the:

* Effective degrees of freedom with fitX$edf
* Deviance with deviance(fitX) or by listing or summarizing the model
* AIC and BIC with AIC(fitX) and BIC(fitX)

fit\_SAT$edf; deviance(fit\_SAT); AIC(fit\_SAT); BIC(fit\_SAT)

[1] 64

[1] 487.6018

[1] 615.6018

[1] 832.5024

Note that for these models, AIC = Deviance + 2(edf)

## Results across all models we’ve fit

| fit | Model | edf | Deviance | AIC | BIC |
| --- | --- | --- | --- | --- | --- |
| 1 | Intercept only | 4 | 604.4 | 612.4 | 625.9 |
| G | Gender only | 8 | 602.3 | 618.3 | 645.4 |
| S | Size only | 8 | 589.2 | 605.2 | 632.3 |
| L | Lake only | 16 | 561.2 | 593.2 | 647.4 |
| LS | Lake and Size | 20 | 540.1 | 580.1 | 647.9 |
| GLS | G, L, S main effects | 24 | 537.9 | 585.9 | 667.2 |
| SAT | G\*S\*L (saturated) | 64 | 487.6 | 615.6 | 615.6 |

Which model looks like it fits the data best?

## Drop in deviance tests (example 1)

Compare Model G to intercept-only

anova(fit\_G, fit\_1)

Likelihood ratio tests of Multinomial Models  
  
Response: food  
 Model Resid. df Resid. Dev Test Df LR stat. Pr(Chi)  
1 1 872 604.3629   
2 gender 868 602.2589 1 vs 2 4 2.104069 0.7166248

## Drop in deviance tests (example 2)

Compare Model SAT to Model GLS

anova(fit\_SAT, fit\_GLS)

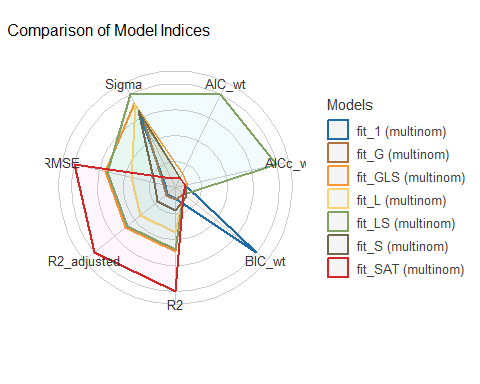
Likelihood ratio tests of Multinomial Models  
  
Response: food  
 Model Resid. df Resid. Dev Test Df LR stat. Pr(Chi)  
1 gender + lake + size 852 537.8655   
2 lake \* size \* gender 812 487.6018 1 vs 2 40 50.26368 0.1281851

## Results of testing

| fit | Model | edf | Deviance | versus | *p* |
| --- | --- | --- | --- | --- | --- |
| 1 | Intercept only | 4 | 604.4 | – | – |
| G | Gender only | 8 | 602.3 | 1 | 0.717 |
| S | Size only | 8 | 589.2 | 1 | 0.004 |
| L | Lake only | 16 | 561.2 | 1 | 0 |
| LS | Lake and Size | 20 | 540.1 | L | 0 |
| GLS | G, L, S main effects | 24 | 537.9 | LS | 0.696 |
| SAT | G\*S\*L (saturated) | 64 | 487.6 | GLS | 0.128 |

## Compare performance

plot(compare\_performance(fit\_SAT, fit\_GLS, fit\_LS, fit\_S, fit\_L, fit\_G, fit\_1),   
 measures = "common")



## Compare performance

compare\_performance(fit\_SAT, fit\_GLS, fit\_LS, fit\_S, fit\_L, fit\_G, fit\_1)

# Comparison of Model Performance Indices  
  
Name | Model | AIC (weights) | AICc (weights) | BIC (weights) | R2  
---------------------------------------------------------------------------  
fit\_SAT | multinom | 615.6 (<.001) | 669.6 (<.001) | 832.5 (<.001) | 0.193  
fit\_GLS | multinom | 585.9 (0.052) | 592.1 (0.020) | 667.2 (<.001) | 0.110  
fit\_LS | multinom | 580.1 (0.946) | 584.3 (0.976) | 647.9 (<.001) | 0.106  
fit\_S | multinom | 605.2 (<.001) | 605.9 (<.001) | 632.3 (0.039) | 0.025  
fit\_L | multinom | 593.2 (0.001) | 595.9 (0.003) | 647.4 (<.001) | 0.071  
fit\_G | multinom | 618.3 (<.001) | 618.9 (<.001) | 645.4 (<.001) | 0.003  
fit\_1 | multinom | 612.4 (<.001) | 612.5 (<.001) | 625.9 (0.961) | 0.000  
  
Name | R2 (adj.) | RMSE | Sigma  
-----------------------------------  
fit\_SAT | 0.190 | 0.347 | 1.774  
fit\_GLS | 0.107 | 0.357 | 1.661  
fit\_LS | 0.103 | 0.358 | 1.647  
fit\_S | 0.022 | 0.372 | 1.671  
fit\_L | 0.068 | 0.365 | 1.663  
fit\_G | 1.722e-04 | 0.375 | 1.689  
fit\_1 | -0.003 | 0.376 | 1.677

## Which model fits the data best?

The model with the lowest AIC is the model which collapses on Gender, and uses only Lake and Size as predictors for Food Choice. (fit\_LS).

* This (fit\_LS) is also the model which has the most evidence in its favor from the drop in deviance testing.

The model with the lowest BIC is the model which collapses on both Gender and Lake, and uses only Size as a predictor for Food Choice. (fit\_S)

## fit\_LS coefficients

coef(fit\_LS)

(Intercept) lakehancock lakeoklawaha laketrafford size<2.3  
invert -1.549021 -1.6581178 0.937237973 1.122002 1.4581457  
rep -3.314512 1.2428408 2.458913302 2.935262 -0.3512702  
bird -2.093358 0.6954256 -0.652622721 1.088098 -0.6306329  
other -1.904343 0.8263115 0.005792737 1.516461 0.3315514

## fit\_LS parameters

model\_parameters(fit\_LS, pretty\_names = FALSE, ci = 0.90)

# Response level: invert  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
----------------------------------------------------------------  
(Intercept) | -1.55 | 0.42 | [-2.25, -0.85] | -3.65 | < .001  
lakehancock | -1.66 | 0.61 | [-2.67, -0.65] | -2.71 | 0.007   
lakeoklawaha | 0.94 | 0.47 | [ 0.16, 1.71] | 1.99 | 0.047   
laketrafford | 1.12 | 0.49 | [ 0.32, 1.93] | 2.29 | 0.022   
size<2.3 | 1.46 | 0.40 | [ 0.81, 2.11] | 3.68 | < .001  
  
# Response level: rep  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
---------------------------------------------------------------  
(Intercept) | -3.31 | 1.05 | [-5.05, -1.58] | -3.15 | 0.002  
lakehancock | 1.24 | 1.19 | [-0.71, 3.19] | 1.05 | 0.294  
lakeoklawaha | 2.46 | 1.12 | [ 0.62, 4.30] | 2.20 | 0.028  
laketrafford | 2.94 | 1.12 | [ 1.10, 4.77] | 2.63 | 0.009  
size<2.3 | -0.35 | 0.58 | [-1.31, 0.60] | -0.61 | 0.545  
  
# Response level: bird  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
---------------------------------------------------------------  
(Intercept) | -2.09 | 0.66 | [-3.18, -1.00] | -3.16 | 0.002  
lakehancock | 0.70 | 0.78 | [-0.59, 1.98] | 0.89 | 0.373  
lakeoklawaha | -0.65 | 1.20 | [-2.63, 1.32] | -0.54 | 0.587  
laketrafford | 1.09 | 0.84 | [-0.30, 2.47] | 1.29 | 0.196  
size<2.3 | -0.63 | 0.64 | [-1.69, 0.43] | -0.98 | 0.326  
  
# Response level: other  
  
Parameter | Log-Odds | SE | 90% CI | z | p  
-------------------------------------------------------------------  
(Intercept) | -1.90 | 0.53 | [-2.77, -1.04] | -3.62 | < .001  
lakehancock | 0.83 | 0.56 | [-0.09, 1.74] | 1.48 | 0.138   
lakeoklawaha | 5.79e-03 | 0.78 | [-1.27, 1.28] | 7.46e-03 | 0.994   
laketrafford | 1.52 | 0.62 | [ 0.49, 2.54] | 2.44 | 0.015   
size<2.3 | 0.33 | 0.45 | [-0.41, 1.07] | 0.74 | 0.460

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald normal distribution approximation.

## fit\_LS parameters (exponentiated)

model\_parameters(fit\_LS, pretty\_names = FALSE, exponentiate = TRUE, ci = 0.90)

# Response level: invert  
  
Parameter | Odds Ratio | SE | 90% CI | z | p  
----------------------------------------------------------------  
(Intercept) | 0.21 | 0.09 | [0.11, 0.43] | -3.65 | < .001  
lakehancock | 0.19 | 0.12 | [0.07, 0.52] | -2.71 | 0.007   
lakeoklawaha | 2.55 | 1.20 | [1.17, 5.55] | 1.99 | 0.047   
laketrafford | 3.07 | 1.51 | [1.37, 6.88] | 2.29 | 0.022   
size<2.3 | 4.30 | 1.70 | [2.24, 8.24] | 3.68 | < .001  
  
# Response level: rep  
  
Parameter | Odds Ratio | SE | 90% CI | z | p  
------------------------------------------------------------------  
(Intercept) | 0.04 | 0.04 | [0.01, 0.21] | -3.15 | 0.002  
lakehancock | 3.47 | 4.11 | [0.49, 24.35] | 1.05 | 0.294  
lakeoklawaha | 11.69 | 13.07 | [1.86, 73.55] | 2.20 | 0.028  
laketrafford | 18.83 | 21.02 | [3.00, 118.10] | 2.63 | 0.009  
size<2.3 | 0.70 | 0.41 | [0.27, 1.83] | -0.61 | 0.545  
  
# Response level: bird  
  
Parameter | Odds Ratio | SE | 90% CI | z | p  
----------------------------------------------------------------  
(Intercept) | 0.12 | 0.08 | [0.04, 0.37] | -3.16 | 0.002  
lakehancock | 2.00 | 1.57 | [0.55, 7.25] | 0.89 | 0.373  
lakeoklawaha | 0.52 | 0.63 | [0.07, 3.76] | -0.54 | 0.587  
laketrafford | 2.97 | 2.50 | [0.74, 11.85] | 1.29 | 0.196  
size<2.3 | 0.53 | 0.34 | [0.18, 1.53] | -0.98 | 0.326  
  
# Response level: other  
  
Parameter | Odds Ratio | SE | 90% CI | z | p  
--------------------------------------------------------------------  
(Intercept) | 0.15 | 0.08 | [0.06, 0.35] | -3.62 | < .001  
lakehancock | 2.28 | 1.27 | [0.91, 5.72] | 1.48 | 0.138   
lakeoklawaha | 1.01 | 0.78 | [0.28, 3.61] | 7.46e-03 | 0.994   
laketrafford | 4.56 | 2.83 | [1.64, 12.66] | 2.44 | 0.015   
size<2.3 | 1.39 | 0.62 | [0.67, 2.91] | 0.74 | 0.460

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald normal distribution approximation.

## The start of the L+S Model

tidy(fit\_LS, exponentiate = FALSE) |>   
 slice(1:5) |> gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| y.level | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| invert | (Intercept) | -1.549 | 0.425 | -3.645 | 0.000 |
| invert | lakehancock | -1.658 | 0.613 | -2.706 | 0.007 |
| invert | lakeoklawaha | 0.937 | 0.472 | 1.986 | 0.047 |
| invert | laketrafford | 1.122 | 0.491 | 2.287 | 0.022 |
| invert | size<2.3 | 1.458 | 0.396 | 3.683 | 0.000 |

* log odds of invertebrates rather than fish are:

-1.549 - 1.658 (Hancock) + 0.937 (Oklahawa) + 1.122 (Trafford) + 1.458 (size < 2.3)

For baseline category, log odds of fish = 0, so exp(log odds) = 1.

## Response Probabilities in fit\_LS

To keep things relatively simple, we’ll look at the class of Large size alligators (so the small size indicator is 0), in Lake George, (so the three Lake indicators are all 0, also).

* The estimated probability of Fish in Large size alligators in Lake George according to our model is:

## Response Probabilities in fit\_LS

* The estimated probability of Invertebrates in Large size alligators in Lake George according to our model is:

The estimated probabilities for the other categories in Large size Lake George alligators are:

* 0.024 for Reptiles, 0.081 for Birds, and 0.098 for Other
* And the five probabilities will sum to 1, at least within rounding error.

## Model Estimates vs. Observed Counts

For large size alligators in Lake George, we have…

| Food Type | Fish | Inverts | Reptiles | Birds | Other |
| --- | --- | --- | --- | --- | --- |
| Observed # | 17 | 1 | 0 | 1 | 3 |
| Observed Prob. | 0.77 | 0.045 | 0 | 0.045 | 0.14 |
| L+S Model Prob. | 0.66 | 0.14 | 0.02 | 0.08 | 0.10 |

We could perform similar calculations for all other combinations of size and lake.

## Classification Table for fit\_LS

table(predict(fit\_LS), gator2$food, dnn = c("predicted", "actual"))

actual  
predicted fish invert rep bird other  
 fish 84 39 16 12 24  
 invert 10 22 3 1 8  
 rep 0 0 0 0 0  
 bird 0 0 0 0 0  
 other 0 0 0 0 0

* So we only predict “fish” and “invert” with fit\_LS and modal prediction, and fit\_LS makes 84 + 22 = 106 correct predictions out of 219 opportunities (48.4% correct.)

## Classification Table for fit\_SAT

table(predict(fit\_SAT), gator2$food, dnn = c("predicted", "actual"))

actual  
predicted fish invert rep bird other  
 fish 83 30 15 11 24  
 invert 11 30 4 1 8  
 rep 0 0 0 0 0  
 bird 0 1 0 1 0  
 other 0 0 0 0 0

* fit\_SAT only predicts “fish” and “invert” with this model, and make 83 + 30 = 113 correct predictions out of 219 opportunities (51.6% correct.)

## Classification Table for fit\_1

table(predict(fit\_1), gator2$food, dnn = c("predicted", "actual"))

actual  
predicted fish invert rep bird other  
 fish 94 61 19 13 32  
 invert 0 0 0 0 0  
 rep 0 0 0 0 0  
 bird 0 0 0 0 0  
 other 0 0 0 0 0

* fit\_SAT only predicts “fish” and makes 94 correct predictions out of 219 opportunities (42.9% correct.)

## Predicted fit\_LS Probabilities

fitLS\_fits <-   
 predict(fit\_LS, newdata = gator2, type = "probs")  
  
gator2\_fit\_LS <- cbind(gator2, fitLS\_fits)  
  
tail(gator2\_fit\_LS, 3)

id food size gender lake fish invert rep bird  
217 217 fish >=2.3 f george 0.6574619 0.1396817 0.02389991 0.08104695  
218 218 invert >=2.3 f george 0.6574619 0.1396817 0.02389991 0.08104695  
219 219 other >=2.3 f george 0.6574619 0.1396817 0.02389991 0.08104695  
 other  
217 0.09790956  
218 0.09790956  
219 0.09790956

## Tabulating Response Probabilities

gator2\_fit\_LS |> group\_by(food) |>  
 summarize(mean(fish), mean(invert), mean(rep),   
 mean(bird), mean(other)) |>  
 gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 20)

| food | mean(fish) | mean(invert) | mean(rep) | mean(bird) | mean(other) |
| --- | --- | --- | --- | --- | --- |
| fish | 0.481 | 0.230 | 0.076 | 0.063 | 0.149 |
| invert | 0.361 | 0.393 | 0.086 | 0.040 | 0.121 |
| rep | 0.381 | 0.258 | 0.148 | 0.064 | 0.148 |
| bird | 0.452 | 0.197 | 0.096 | 0.084 | 0.171 |
| other | 0.426 | 0.246 | 0.079 | 0.073 | 0.175 |

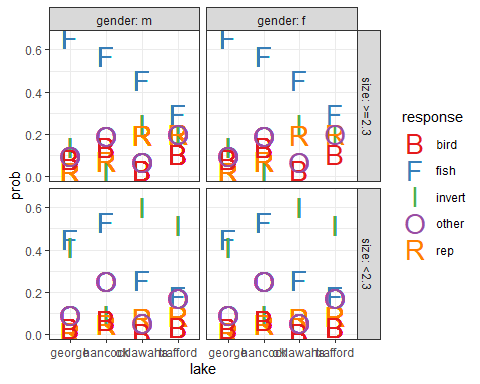
## Turn Wide Data into Long

gator2\_fitLSlong <-  
 pivot\_longer(gator2\_fit\_LS,   
 cols = fish:other,   
 names\_to = "response",  
 values\_to = "prob")  
  
head(gator2\_fitLSlong)

# A tibble: 6 × 7  
 id food size gender lake response prob  
 <chr> <fct> <fct> <fct> <fct> <chr> <dbl>  
1 1 fish <2.3 m hancock fish 0.535   
2 1 fish <2.3 m hancock invert 0.0931  
3 1 fish <2.3 m hancock rep 0.0475  
4 1 fish <2.3 m hancock bird 0.0704  
5 1 fish <2.3 m hancock other 0.254   
6 2 fish <2.3 m hancock fish 0.535

## fit\_LS Response Probabilities

ggplot(gator2\_fitLSlong, aes(x = lake, y = prob,   
 col = response,  
 shape = response)) +  
 geom\_point(size = 7) +  
 scale\_color\_brewer(palette = "Set1") +  
 facet\_grid(size ~ gender, labeller = "label\_both") +  
 scale\_shape\_manual(values = c("B", "F", "I", "O", "R"))



## Other Sources

In addition to the example found in our Course Notes (Chapter 28)…

* A good source of information on fitting these models is <https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/>
* Using the tidymodels structure to fit these models is another good idea. Julia Silge has a very nice example at <https://juliasilge.com/blog/multinomial-volcano-eruptions/>
* More mathematically oriented sources include the following texts:
  + Hosmer DW Lemeshow S Sturdivant RX (2013) Applied Logistic Regression, 3rd Edition, Wiley
  + Agresti A (2007) An Introduction to Categorical Data Analysis, 2nd Edition, Wiley.

1. Source: Agresti’s 1996 first edition of An Introduction to Categorical Data Analysis, Table 8.1. These were provided by Delany MF and Moore CT. [↑](#footnote-ref-24)
2. Source: <https://online.stat.psu.edu/stat504/lesson/8> [↑](#footnote-ref-71)