432 Class 22 Slides

github.com/THOMASELOVE/432-2018

2018-04-05

Preliminaries

```
library(skimr)
library(rms)
library(nnet)
library(MASS)
library(broom)
library(tidyverse)
```

```
gator1 <- read.csv("data/gator1.csv") %>% tbl_df
gator2 <- read.csv("data/gator2.csv") %>% tbl_df
asbestos <- read.csv("data/asbestos.csv") %>% tbl_df
```

Today's Agenda

- Data Visualization: A Graphic Memorial
- Multinomial Logistic Regression: An Introduction
- Ordinal Logistic Regression: An Introduction

Data Visualization: Napoleon's Russian Campaign

Wainer: Chapter 4 of *Visual Revelations*

CHAPTER 4 Three Graphic Memorials

"*Hear, forget; see, remember.*" The wisdom of this ancient Confucian saying is apparent. Memorable memorials are visual. Who can ever forget the tragedy chronicled by the austere black granite wall that is the Vietnam Memorial? It is massive in form and content, built from the space taken by the more than 58,000 names inscribed upon it. As the loss of life increases, so too does the height of the wall, and the emotions it evokes. It is a very personal thing. William A. Atwell, Terry Lee Dillard, Ward K. Patton, Jerry Lee Graves, Edward J. Downs, John E. Rice, Jack M. Strong—these names join with thousands of others to form the wall. The interaction of the monument with those who come to it, whether to seek out a particular name or to picnic, often becomes part of the diverse images we take away with us. The tragedy of Vietnam written in the small becomes large and indelible.

The History

It's 1812.

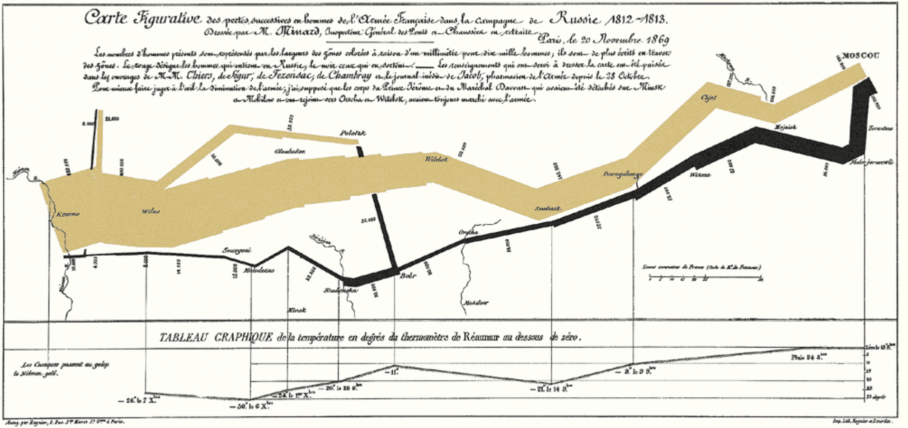
- Napoleon has most of Europe (outside of the United Kingdom) under his control.
- But he cannot break through the defenses of the U.K., so he decides to place an embargo on them.
- The Russian Czar, Alexander, refuses to participate in the embargo.

So Napoleon gathers a massive army of over 400,000 to attack Russia in June 1812.

- Meanwhile, Russia has a plan. As Napoleon's troops advance, the Russian troops burn everything they pass.

Charles Minard's original map

Napoleon's disastrous Russian Campaign of 1812

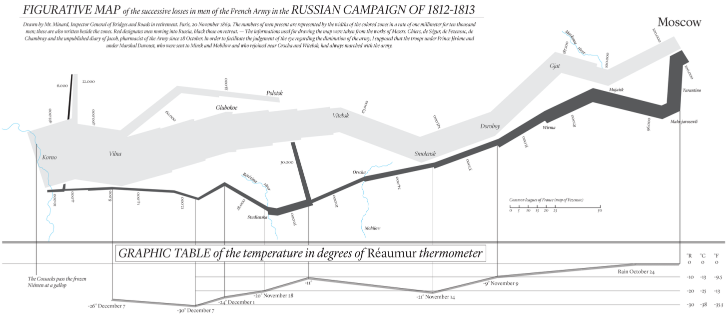


Wainer: Chapter 4 [b]

Napoleon's Russian Campaign

Memorializing that portion of the generation of young French men lost in Napoleon's ill-fated Russian campaign was surely part of Charles Joseph Minard's motivation in the construction of his famous 1869 graphic. Minard's plot, shown in figure 1, depicts the movement of the French army from the time it crossed the Polish-Russian border with 422,000 men in June of 1812. The shrinking size of the army is characterized by the progressive narrowing of the broad band stretching across the map. In the original scale, each millimeter of its width represents 10,000 men. When the army reached Moscow in September, only 100,000 remained. The city was deserted, and the army began its retreat, depicted by the darker line below. It is linked to the temperature scale showing quantitatively the depths of the Russian winter. The banks of the Berezina River were littered with the bodies of the 22,000 men who perished as the November temperature dropped to –20°. When the remainder of the army crossed into Poland as the year ended, only 10,000 men remained.

A Modern Redrawing of Minard's Original Map



Source: By Iñigo Lopez - Own work, CC BY-SA 4.0, at this link

What are we looking at?

- The numbers of Napoleon's troops by location (longitude)
  - Organized by group (at one point they divided into three groups) and direction (advance, then retreat)
- The path that his troops took to Moscow and back again
- The temperature experienced by his troops when winter settled in on the return trip
- Historical context, as shown in the passage of time
- Geography (for example, river crossings)

Wainer: Chapter 4 [c]

The story of the tragedy is clear. We can see the bodies frozen into the snow. Marey told how this graph "brought tears to the eyes of all France."[1] No wonder; there were few families unaffected.

Minard's depiction of Napoleon's Russian campaign has been characterized as perhaps "the best statistical graphic ever drawn."[2] Why? It is not the quality of the pen stroke, although it certainly passes muster in that regard. It is the importance and richness of the data. A single page carries six variables that tell the evocative story of where and how thousands of men died. Its poignancy is heightened through the immediate and graphic answer to the question, Compared to what? Ten thousand men returned. A lot or a few? Opposing the returning trickle against the departing torrent answers the question. The difference between them measures the tragedy. But nowhere does the shrinking distance between two lines depict a more touching tragedy than in my next example.

A Large Version of the Map

As part of the Class 22 materials, the map is here.

Several Useful Sources

- This link at thoughtbot was a major source here
- the work of Edward Tufte, gathered at edwardtufte dot com, as well as his four pivotal books
- the work of Howard Wainer, who has several relevant books, including *Graphic Discovery*, *Picturing the Uncertain World*, and *Visual Revelations*, on which I also drew.

Multinomial Logistic Regression: An Introduction

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.

The `gator1` data: Alligator Food Choice

The data are from a study by the Florida Game and Fresh Water Fish Commission of factors influencing the primary food choice of alligators[1].

The data include the following data for 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,)
  - Other (which includes reptiles, amphibians, mammals, plant material and stones or other debris.)

We'll be trying to predict primary food `choice` using `length`.

[1] My 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.

Alligator Food Choice, Part 1

```
gator1
```

```
# A tibble: 59 x 3
      id length choice
   <int>  <dbl> <fct>
 1     1   1.24 Invertebrates
 2     2   1.30 Invertebrates
 3     3   1.30 Invertebrates
 4     4   1.32 Fish
 5     5   1.32 Fish
 6     6   1.40 Fish
 7     7   1.42 Invertebrates
 8     8   1.42 Fish
 9     9   1.45 Invertebrates
10    10   1.45 Other
# ... with 49 more rows
```

Alligator Food Choice Summaries

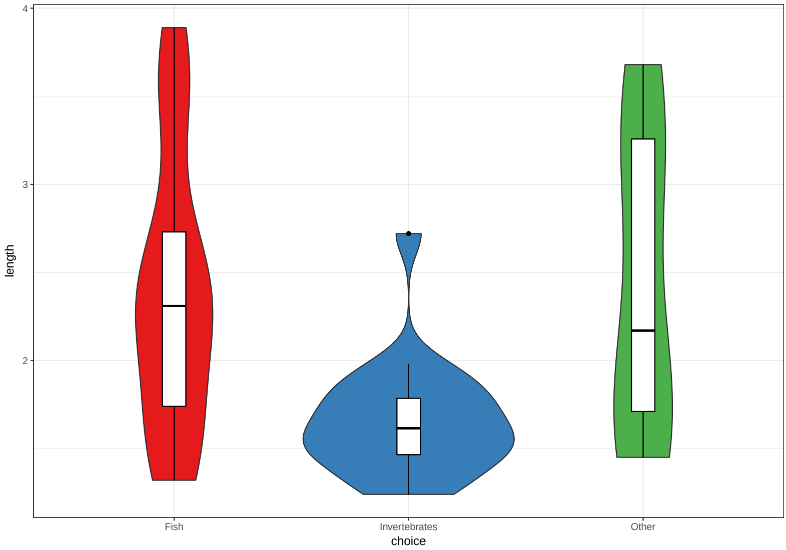
```
> gator1 %>% select(choice, length) %>% skim
Skim summary statistics
 n obs: 59
 n variables: 2

Variable type: factor
 variable missing complete  n n_unique                          top_counts ordered
   choice       0       59 59        3 Fis: 31, Inv: 20, Oth: 8, NA: 0   FALSE

Variable type: numeric
 variable missing complete  n mean   sd   p0  p25 median  p75 p100     hist
   length       0       59 59 2.13 0.74 1.24 1.58   1.85 2.45 3.89
```

```
           choice         length     
 Fish          :31   Min.   :1.240  
 Invertebrates:20   1st Qu.:1.575  
 Other         : 8   Median :1.850  
                     Mean   :2.130  
                     3rd Qu.:2.450  
                     Max.   :3.890  
```

Plotting Length by Primary Food Choice



Plotting Length by Primary Food Choice (code)

```
ggplot(gator1, aes(x = choice, y = length, fill = choice)) +
    geom_violin(trim = TRUE) +
    geom_boxplot(fill = "white", col = "black",
                 width = 0.1) +
    scale_fill_brewer(palette = "Set1") +
    theme_bw() +
    guides(fill = FALSE)
```

Fitting a Multinomial Logistic Regression

- We'll start by setting "Other" as the first (reference) level for the choice outcome

```
gator1 <- gator1 %>%
    mutate(choice = fct_relevel(choice, "Other"))
```

For our first try, we'll use the `multinom` function from the `nnet` package...

```
try1 <- multinom(choice ~ length, data=gator1)
```

```
# weights:  9 (4 variable)
initial  value 64.818125
iter  10 value 49.170785
final  value 49.170622
converged
```

Looking over the first try

```
try1
```

```
Call:
multinom(formula = choice ~ length, data = gator1)

Coefficients:
              (Intercept)     length
Fish             1.617952 -0.1101836
Invertebrates    5.697543 -2.4654695

Residual Deviance: 98.34124
AIC: 106.3412
```

Our R output suggests the following models:

- log odds of Fish rather than Other = 1.62 - 0.110 Length
- log odds of Invertebrates rather than Other = 5.70 - 2.465 Length

Estimating Response Probabilities from our First Try

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

$$\pi_j = \frac{exp(\beta_{0j} + \beta_{1j}x)}{\Sigma_j exp(\beta_{0j} + \beta_{1j}x)}$$

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

- log odds of Fish rather than Other = 1.62 - 0.110 Length
- log odds of Invertebrates rather than Other = 5.70 - 2.465 Length
- For the reference category we use $\beta_{0j} = 0$ and $\beta_{1j} = 0$ so that $exp(\beta_{0j} + \beta_{1j}x) = 1$ for that category.

Estimated Response Probabilities

- log odds of Fish rather than Other = 1.62 - 0.110 Length
- log odds of Invertebrates rather than Other = 5.70 - 2.465 Length

and so our estimates (which will sum to 1) are:

$$Pr(Fish|Length = L) = \frac{exp(1.62 - 0.110L)}{1 + exp(1.62 - 0.110L) + exp(5.70 - 2.465L)}$$

$$Pr(Invert.|Length = L) = \frac{exp(5.70 - 2.465L)}{1 + exp(1.62 - 0.110L) + exp(5.70 - 2.465L)}$$

$$Pr(Other|Length = L) = \frac{1}{1 + exp(1.62 - 0.110L) + exp(5.70 - 2.465L)}$$

Making a Prediction

For an alligator of 3.9 meters, for instance, the estimated probability that primary food choice is "other" equals:

$$\hat{\pi}(Other) = \frac{1}{1 + exp(1.62 - 0.110[3.9]) + exp(5.70 - 2.465[3.9])} = 0.232$$

Storing Predicted Probabilities from `try1`

```
try1_fits <-
    predict(try1, newdata = gator1, type = "probs")

gator1_try1 <- cbind(gator1, try1_fits)

head(gator1_try1, 3)
```

```
  id length         choice      Other      Fish
1  1   1.24 Invertebrates 0.05150117 0.2265417
2  2   1.30 Invertebrates 0.05727232 0.2502677
3  3   1.30 Invertebrates 0.05727232 0.2502677
  Invertebrates
1     0.7219571
2     0.6924600
3     0.6924600
```

Tabulating Response Probabilities

```
gator1_try1 %>% group_by(choice) %>%
    summarize(mean(Other), mean(Fish), mean(Invertebrates))
```

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

Turn Wide Data into Long

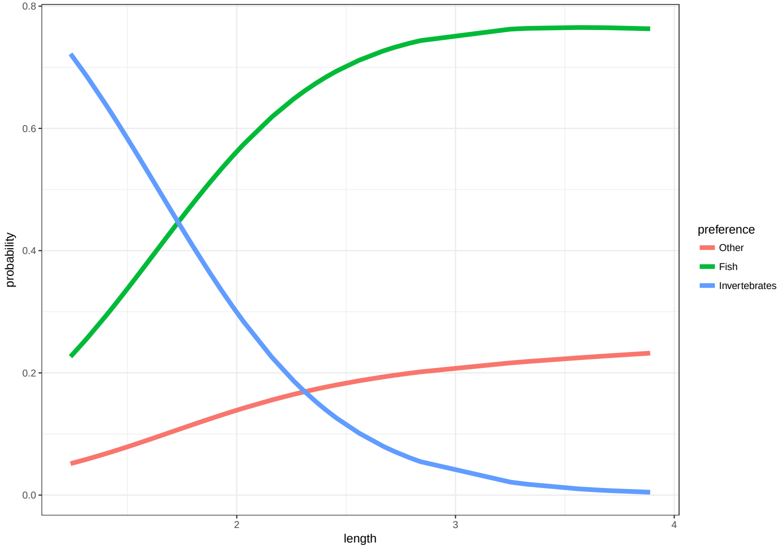
```r
gator1_try1long <-
    gather(gator1_try1, key = preference,
           value = probability,
           Other:Invertebrates, factor_key = TRUE)

head(gator1_try1long,3)
```

```
  id length        choice preference probability
1  1   1.24 Invertebrates      Other  0.05150117
2  2   1.30 Invertebrates      Other  0.05727232
3  3   1.30 Invertebrates      Other  0.05727232
```

See this link at cookbook-r.com.

Graphing the Model's Response Probabilities



Graphing the Response Probabilities (code)

```
ggplot(gator1_try1long, aes(x = length, y = probability,
                            col = preference)) +
    geom_line(size = 2) +
    scale_fill_brewer(palette = "Set1") +
    theme_bw()
```

summary of try1

```
Call:
multinom(formula = choice ~ length, data = gator1)

Coefficients:
              (Intercept)     length
Fish             1.617952 -0.1101836
Invertebrates    5.697543 -2.4654695

Std. Errors:
              (Intercept)    length
Fish             1.307291 0.5170838
Invertebrates    1.793820 0.8996485

Residual Deviance: 98.34124
AIC: 106.3412
```

Assess the try1 model as a whole with a drop in deviance test

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

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

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

ANOVA to compare try0 to try1

```
anova(try0, try1)
```

```
Likelihood ratio tests of Multinomial Models

Response: choice
   Model Resid. df Resid. Dev   Test    Df LR stat.
1      1       116  115.14186
2 length       114   98.34124 1 vs 2     2 16.80061
       Pr(Chi)
1
2 0.0002247985
```

Does the inclusion of `length` produce a significantly better fit to the data than simply fitting an intercept?

Wald Z tests for individual predictors

```
z <- summary(try1)$coefficients /
  summary(try1)$standard.errors  ## Wald Z tests
p <- (1 - pnorm(abs(z), 0, 1)) * 2 ## 2-sided p values
z
```

```
              (Intercept)     length
Fish             1.237637 -0.2130865
Invertebrates    3.176206 -2.7404808
```

```
p
```

```
              (Intercept)      length
Fish          0.215850665 0.831259475
Invertebrates 0.001492149 0.006134937
```

A Larger Alligator Food Choice Example

The `gator2.csv` data[2] 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)

The 219 alligators are also categorized by sex, and by length (< 2.3 and ≥ 2.3 meters) and by which of four lakes they were captured in (Hancock, Oklawaha, Trafford or George.)

[2] Source: https://onlinecourses.science.psu.edu/stat504/node/226

Table of gator2 data

Lake	Sex	Size	Primary Food Choice: Fish	Inv.	Rept.	Bird	Other
Hancock	M	small	7	1	0	0	5
		large	4	0	0	1	2
	F	small	16	3	2	2	3
		large	3	0	1	2	3
Oklawaha	M	small	2	2	0	0	1
		large	13	7	6	0	0
	F	small	3	9	1	0	2
		large	0	1	0	1	0
Trafford	M	small	3	7	1	0	1
		large	8	6	6	3	5
	F	small	2	4	1	1	4
		large	0	1	0	0	0
George	M	small	13	10	0	2	2
		large	9	0	0	1	2
	F	small	3	9	1	0	1
		large	8	1	0	0	1

Model Setup

$$\pi_1 = Pr(Fish), \pi_2 = Pr(Invert.), \pi_3 = Pr(Reptiles),$$
$$\pi_4 = Pr(Birds), \pi_5 = Pr(Other)$$

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

$$log(\frac{\pi_j}{\pi_i}) = \beta_0 + \beta_1[Lake = Hancock] + \beta_2[Lake = Oklawaha] +$$
$$\beta_3[Lake = Trafford] + \beta_4[Length \geq 2.3] + \beta_5[Sex = Female]$$

for $j = 2, 3, 4, 5$.

- We have six coefficients to estimate in each of four logit equations (one each for $j = 2, 3, 4, 5$) so there are 24 parameters to estimate.

Rearranging the gator2 data

We re-order the levels of the factors to get our reference category as first in each list.

```
gator2$food   <- factor(gator2$food,
                        levels = c("fish", "invert",
                            "rep", "bird", "other"))
gator2$size   <- factor(gator2$size,
                        levels = c(">=2.3","<2.3"))
gator2$gender <- factor(gator2$gender,
                        levels=c("m","f"))
gator2$lake   <- factor(gator2$lake,
                        levels=c("george", "hancock",
                              "oklawaha","trafford"))
```

gator2 summary

```
summary(gator2)
```

```
       id              food         size       gender
 Min.   :  1.0   fish   :94   >=2.3: 95   m:130
 1st Qu.: 55.5   invert:61   <2.3 :124   f: 89
 Median :110.0   rep    :19
 Mean   :110.0   bird   :13
 3rd Qu.:164.5   other  :32
 Max.   :219.0
       lake
 george  :63
 hancock :55
 oklawaha:48
 trafford:53
```

Complete Set of Models We Will Fit

- Response: Category of Primary Food Choice
- Predictors: L = lake, G = gender, S = size

Specifically, we'll fit (using the `multinom` function in the `nnet` package)

- 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

Our Models (Code)

```
options(contrasts=c("contr.treatment", "contr.poly"))
fitS <- multinom(food ~ lake*size*gender, data=gator2)
        # saturated
fit0<-multinom(food~1,data=gator2)                    # null
fit1<-multinom(food~gender,data=gator2)               # G
fit2<-multinom(food~size,data=gator2)                 # S
fit3<-multinom(food~lake,data=gator2)                 # L
fit4<-multinom(food~size+lake,data=gator2)            # L + S
fit5<-multinom(food~size+lake+gender,data=gator2) # L + S + G
```

What You'll See When Fitting the models

```
options(contrasts=c("contr.treatment", "contr.poly"))
fitS <- 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
```

```
fit0<-multinom(food~1,data=gator2)         # null
```

```
# weights:  10 (4 variable)
initial  value 352.466903
```

Summarizing the Models: Intercept only

```
> summary(fit0)
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
```

Summarizing the Models: Lake only

```
> summary(fit3)
Call:
multinom(formula = food ~ lake, data = gator2)

Coefficients:
       (Intercept) lakehancock lakeoklawaha laketrafford
invert  -0.5008393  -1.5137909   0.55488981    0.8263598
rep     -3.4962205   1.1937161   2.55175319    3.0107928
bird    -2.3982809   0.6065505  -0.49188808    1.2197770
other   -1.7048477   0.8686390  -0.08689071    1.4425750

Std. Errors:
       (Intercept) lakehancock lakeoklawaha laketrafford
invert   0.2833774   0.6029744    0.4341550    0.4612870
rep      1.0148749   1.1817886    1.1083250    1.1099095
bird     0.6031161   0.7727128    1.1912598    0.8310587
other    0.4438217   0.5542879    0.7654142    0.6114775

Residual Deviance: 561.1677
AIC: 593.1677
```

Summarizing the Models: Saturated Model

```
> summary(fit5)
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 lakehancock:genderf
invert            6.0160287             -21.85028           -21.3342850           -3.946342
rep             -15.0175978             -17.43950             1.3387310           24.889170
bird            -22.8201143             -25.18859           -25.8829682           18.248790
other             0.7242536              26.40938            -0.2614093            1.268734
       lakeoklawaha:genderf laketrafford:genderf size<2.3:genderf lakehancock:size<2.3:genderf
invert             4.226498            25.465169      -19.2913107                     2.857688
rep              -13.585689           -18.078274       31.5836415                   -15.396895
bird              62.485154            16.978562        0.6638064                    20.157737
other             -1.758853            -7.586589        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

Std. Errors:
       (Intercept) lakehancock lakeoklawaha laketrafford  size<2.3   genderf lakehancock:size<2.3
invert   0.4573145  275.959121    0.5527936    0.5997448 0.4567653 0.8383382          275.9591215
rep      0.4081802    0.466826    0.5083853    0.5349234 0.6888344 0.5417143            0.5485564
bird     1.0538859    1.536392    0.7186034    1.2526096 1.2990933 0.6429296            0.5479089
other    0.7817035    1.166654    0.7205701    0.9675017 1.0898855 1.3176402            1.5102076
       lakeoklawaha:size<2.3 laketrafford:size<2.3 lakehancock:genderf lakeoklawaha:genderf
invert          1.012159e+00             0.8492097         275.9591112            0.7210548
rep             4.773431e-01             0.7899858           0.4668260            0.4773431
bird            4.038695e-08             0.4466887           0.9324221            0.7186034
other           7.205701e-01             1.6871371           1.7756218            1.0300313
       laketrafford:genderf size<2.3:genderf lakehancock:size<2.3:genderf
invert            0.6796177        1.0109439                  275.9598724
rep               0.6341858        0.6032116                    0.5485564
bird              0.4466887        0.7486254                    0.5479089
other             0.9992618        1.9096101                    2.4087288
       lakeoklawaha:size<2.3:genderf laketrafford:size<2.3:genderf
invert                  8.781523e-01                     0.6796177
rep                     4.773431e-01                     0.6341858
bird                    4.111562e-08                     0.4466887
other                   1.030031e+00                     0.9992618

Residual Deviance: 487.6018
AIC: 615.6018
```

Building a Model Comparison Table

For a model `fitX`, we find the:

- Deviance with `deviance(fitX)` or by listing or summarizing the model
- AIC with `AIC(fitX)` or by listing or summarizing the model
- Effective degrees of freedom with `fitX$edf`

Label	Model	Deviance	Effective df
`fitS`	`L*S*G` (saturated)	487.6	64

Likelihood Ratio Tests

```
anova(fit0, fit1, fit2, fit3, fit4, fit5, fitS)
```

```
Likelihood ratio tests of Multinomial Models

Response: food
                   Model Resid. df Resid. Dev   Test    Df
1                      1       872   604.3629
2                 gender       868   602.2589 1 vs 2     4
3                   size       868   589.2134 2 vs 3     0
4                   lake       860   561.1677 3 vs 4     8
5            size + lake       856   540.0803 4 vs 5     4
6 size + lake + gender       852   537.8655 5 vs 6     4
7 lake * size * gender       812   487.6018 6 vs 7    40
   LR stat.      Pr(Chi)
1
2  2.104069 0.7166248128
3 13.045500 0.0000000000
```

Summary Table

#	Model	Test	p	AIC
1	`1`	-	-	612.36
2	`G`	1 vs 2	0.717	618.26
3	`S`	2 vs 3	<0.001	605.21
4	`L`	3 vs 4	<0.001	593.17
5	`L+S`	4 vs 5	<0.001	580.08
6	`G+L+S`	5 vs 6	0.696	585.87
7	`G*L*S`	6 vs 7	0.128	615.6

So, which model appears to fit the data best?

Summary Table

#	Model	Test	p	AIC
1	1	-	-	612.36
2	G	1 vs 2	0.717	618.26
3	S	2 vs 3	<0.001	605.21
4	L	3 vs 4	<0.001	593.17
5	L+S	4 vs 5	<0.001	580.08
6	G+L+S	5 vs 6	0.696	585.87
7	G*L*S	6 vs 7	0.128	615.6

According to AIC and to the direct *p* value comparisons, the best model (of these) is apparently the model which collapses on Gender, and uses only Lake and Size as predictors for Food Choice. A stepwise procedure starting with the `G+L+S` model, i.e. `step(fit5)`, will also land on this same model.

The L+S Model

```
> fit4
Call:
multinom(formula = food ~ size + lake, data = gator2)

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

Residual Deviance: 540.0803
AIC: 580.0803
```

- So, for instance, log odds of invertebrates rather than fish are:

```
-1.54 + 1.46 Small - 1.66 Hancock
      + 0.94 Oklawaha + 1.12 Trafford
```

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

Response Probabilities in the L+S Model

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:

$$\hat{\pi}(Fish) = \frac{1}{1 + exp(-1.54) + exp(-3.31) + exp(-2.09) + exp(-1.90)}$$

$$= \frac{1}{1.524} = 0.66$$

Response Probabilities in the L+S Model

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

$$\hat{\pi}(Inv.) = \frac{exp(-1.54)}{1 + exp(-1.54) + exp(-3.31) + exp(-2.09) + exp(-1.90)}$$

$$= \frac{0.214}{1.524} = 0.14$$

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

- 0.02 for Reptiles, 0.08 for Birds, and 0.10 for Other
- And the five probabilities will sum to 1, at least within rounding error.

Comparing Model Estimates to Observed Counts

For large size alligators in Lake George, we have...

Food Type	Fish	Invertebrates	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, but I'll leave that to the dedicated.

Storing Predicted Probabilities from `fit4`

```r
fit4_fits <-
    predict(fit4, newdata = gator2, type = "probs")

gator2_fit4 <- cbind(gator2, fit4_fits)

head(gator2_fit4, 3)
```

```
  id food size gender    lake      fish     invert
1  1 fish <2.3      m hancock 0.5352844 0.09311221
2  2 fish <2.3      m hancock 0.5352844 0.09311221
3  3 fish <2.3      m hancock 0.5352844 0.09311221
         rep       bird     other
1 0.04745855 0.07040277 0.2537421
2 0.04745855 0.07040277 0.2537421
3 0.04745855 0.07040277 0.2537421
```

Tabulating Response Probabilities

```
gator2_fit4 %>% group_by(food) %>%
    summarize(mean(fish), mean(invert), mean(rep),
              mean(bird), mean(other))
```

```
# A tibble: 5 x 6
  food   `mean(fish)` `mean(invert)` `mean(rep)`
  <fct>         <dbl>          <dbl>       <dbl>
1 fish          0.481          0.230      0.0763
2 invert        0.361          0.393      0.0858
3 rep           0.381          0.258      0.148
4 bird          0.452          0.197      0.0960
5 other         0.426          0.246      0.0791
# ... with 2 more variables: `mean(bird)` <dbl>,
#   `mean(other)` <dbl>
```

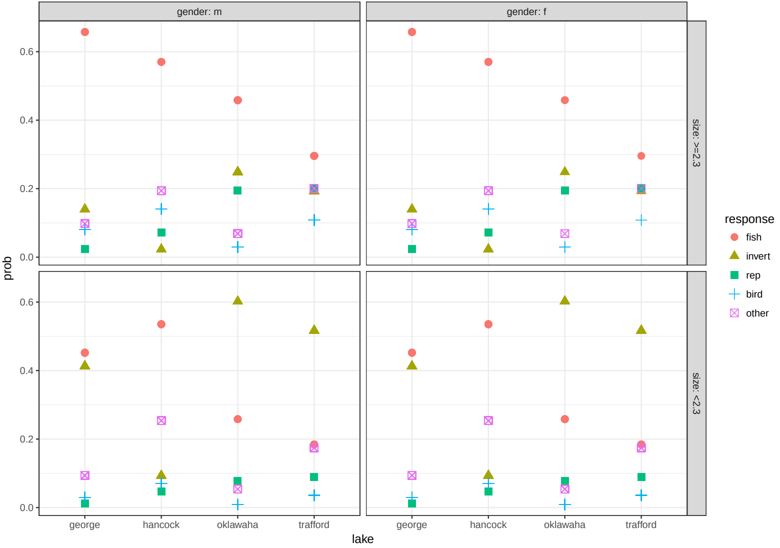
Turn Wide Data into Long

```
gator2_fit4long <-
    gather(gator2_fit4, key = response,
           value = prob,
           fish:other, factor_key = TRUE)

head(gator2_fit4long,3)
```

```
  id food size gender    lake response      prob
1  1 fish <2.3      m hancock     fish 0.5352844
2  2 fish <2.3      m hancock     fish 0.5352844
3  3 fish <2.3      m hancock     fish 0.5352844
```

Graphing the Model's Response Probabilities



Graphing the Model's Response Probabilities (code)

```
ggplot(gator2_fit4long, aes(x = lake, y = prob,
                            col = response,
                            shape = response)) +
    geom_point(size = 3) +
    scale_fill_brewer(palette = "Set1") +
    theme_bw() +
    facet_grid(size ~ gender, labeller = "label_both")
```

Some Sources for Multinomial Logistic Regression

- A good source of information on fitting these models is http://www.ats.ucla.edu/stat/r/dae/mlogit.htm
- 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.
        - There's a related resource for this text that shows R code for doing everything in the book at https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf

Ordinal Logistic Regression: An Extra Example

Asbestos Exposure in the U.S. Navy

These data describe 83 Navy workers, engaged in jobs involving potential asbestos exposure.

- The workers were either removing asbestos tile or asbestos insulation, and we might reasonably expect that those exposures would be different (with more exposure associated with insulation removal).
- The workers either worked with general ventilation (like a fan or naturally occurring wind) or negative pressure (where a pump with a High Efficiency Particulate Air filter is used to draw air (and fibers) from the work area.)
- The duration of a sampling period (in minutes) was recorded, and their asbestos exposure was measured and classified in three categories:
  - low exposure (< 0.05 fibers per cubic centimeter),
  - action level (between 0.05 and 0.1) and
  - above the legal limit (more than 0.1 fibers per cc).

Our Outcome and Modeling Task

We'll predict the ordinal Exposure variable, in an ordinal logistic regression model with a proportional odds assumption, using the three predictors

- Task (Insulation or Tile),
- Ventilation (General or Negative pressure) and
- Duration (in minutes).

Exposure is determined by taking air samples in a circle of diameter 2.5 feet around the worker's mouth and nose.

Summarizing the Asbestos Data

We'll make sure the Exposure factor is ordinal...

```
asbestos$Exposure <- factor(asbestos$Exposure, ordered=T)
summary(asbestos[,2:5])
```

```
         Task                      Ventilation     Duration    
 Insulation:46   General             :34   Min.   : 30.0  
 Tile      :37   Negative pressure:49   1st Qu.: 85.0  
                                        Median :138.0  
                                        Mean   :147.1  
                                        3rd Qu.:212.5  
                                        Max.   :300.0  
                 Exposure
 (1) Low exposure     :45  
 (2) Action level     : 6  
 (3) Above legal limit:32  
```

The Proportional-Odds Cumulative Logit Model

We'll use the `polr` function in the `MASS` library to fit our ordinal logistic regression.

- Clearly, Exposure group (3) Above legal limit, is worst, followed by group (2) Action level, and then group (1) Low exposure.
- We'll have two indicator variables (one for Task and one for Ventilation) and then one continuous variable (for Duration).
- The model will have two logit equations: one comparing group (1) to group (2) and one comparing group (2) to group (3), and three slopes, for a total of five free parameters.

Equations to be Fit

The equations to be fit are:

$$log(\frac{Pr(Exposure \leq 1)}{Pr(Exposure > 1)}) = \beta_{0[1]} + \beta_1 Task + \beta_2 Ventilation + \beta_3 Duration$$

and

$$log(\frac{Pr(Exposure \leq 2)}{Pr(Exposure > 2)}) = \beta_{0[2]} + \beta_1 Task + \beta_2 Ventilation + \beta_3 Duration$$

where the intercept term is the only piece that varies across the two equations.

- A positive coefficient β means that increasing the value of that predictor tends to *lower* the Exposure category, and thus the asbestos exposure.

Fitting the Model with the polr function in MASS

```
model.A <- polr(Exposure ~ Task + Ventilation + Duration,
                    data=asbestos)
```

Model Summary

```
> summary(model.A)

Re-fitting to get Hessian

Call:
polr(formula = Exposure ~ Task + Ventilation + Duration, data = asbestos)

Coefficients:
                                 Value Std. Error t value
TaskTile                     -2.251333   0.644792 -3.4916
VentilationNegative pressure -2.156979   0.567540 -3.8006
Duration                     -0.000708   0.003799 -0.1864

Intercepts:
                                       Value   Std. Error t value
(1) Low exposure|(2) Action level      -2.0575  0.6611    -3.1123
(2) Action level|(3) Above legal limit -1.5111  0.6344    -2.3820

Residual Deviance: 99.87952
AIC: 109.8795
```

Explaining the Model Summary

The first part of the output provides coefficient estimates for the three predictors.

```
                             Value Std. Error t value
TaskTile                 -2.251333   0.644792 -3.4916
VentilationNegative pressure -2.156979   0.567540 -3.8006
Duration                 -0.000708   0.003799 -0.1864
```

- The estimated slope for Task = Tile is -2.25. This means that Task = Tile provides less exposure than does the other Task (Insulation) so long as the other predictors are held constant.
- Typically, we would express this in terms of an odds ratio.

Odds Ratios and CI for Model A

```r
exp(coef(model.A))
```

```
                   TaskTile VentilationNegative pressure 
                  0.1052589                    0.1156740 
                   Duration 
                  0.9992922 
```

```r
exp(confint(model.A))
```

```
Waiting for profiling to be done...


Re-fitting to get Hessian

                                  2.5 %    97.5 %
TaskTile                     0.02718379 0.3538549
VentilationNegative pressure 0.03641039 0.3427734
Duration                     0.99187230 1.0069533
```

Assessing the Ventilation Coefficient

```
                             Value Std. Error t value
TaskTile                 -2.251333   0.644792 -3.4916
VentilationNegative pressure -2.156979   0.567540 -3.8006
Duration                 -0.000708   0.003799 -0.1864
```

Similarly, the estimated slope for Ventilation = Negative pressure (-2.16) means that Negative pressure provides less exposure than does General Ventilation. We see a relatively modest effect (near zero) associated with Duration.

Summary of Model A: Estimated Intercepts

```
Intercepts:
                                          Value    Std. Error t va
(1) Low exposure|(2) Action level        -2.0575  0.6611     -3.1
(2) Action level|(3) Above legal limit -1.5111  0.6344     -2.3
```

The first parameter (-2.06) is the estimated log odds of falling into category (1) low exposure versus all other categories, when all of the predictor variables (Task, Ventilation and Duration) are zero. So the first estimated logit equation is:

$$log(\frac{Pr(Exposure \leq 1)}{Pr(Exposure > 1)}) =$$

$$-2.06 - 2.25[Task = Tile] - 2.16[Vent = NP] - 0.0007 Duration$$

Summary of Model A: Estimated Intercepts

```
Intercepts:
                                        Value    Std. Error t va
(1) Low exposure|(2) Action level      -2.0575  0.6611     -3.1
(2) Action level|(3) Above legal limit -1.5111  0.6344     -2.3
```

The second parameter (-1.51) is the estimated log odds of category (1) or (2) vs. (3). The estimated logit equation is:

$$log(\frac{Pr(Exposure \leq 2)}{Pr(Exposure > 2)}) =$$

$$-1.51 - 2.25[Task = Tile] - 2.16[Vent = NP] - 0.0007 Duration$$

Comparing Model A to an "Intercept only" Model

```r
model.null <- polr(Exposure ~ 1, data=asbestos)
anova(model.null, model.A)
```

```
Likelihood ratio tests of ordinal regression models

Response: Exposure
                          Model Resid. df Resid. Dev   Test
1                             1        81  147.61971
2 Task + Ventilation + Duration        78   99.87952 1 vs 2
     Df LR stat.      Pr(Chi)
1
2     3 47.74019 2.41857e-10
```

Comparing Model A to Model without Duration

```
model.B <- polr(Exposure ~ Task + Ventilation, data=asbestos)
anova(model.A, model.B)
```

```
Likelihood ratio tests of ordinal regression models

Response: Exposure
                              Model Resid. df Resid. Dev   Test
1              Task + Ventilation        79   99.91421
2 Task + Ventilation + Duration        78   99.87952 1 vs 2
     Df    LR stat.   Pr(Chi)
1
2     1 0.03469471 0.8522368
```

Is a Task*Ventilation Interaction significant?

```
model.C <- polr(Exposure ~ Task * Ventilation, data=asbestos)
anova(model.B, model.C)
```

```
Likelihood ratio tests of ordinal regression models

Response: Exposure
               Model Resid. df Resid. Dev   Test    Df
1 Task + Ventilation        79   99.91421
2 Task * Ventilation        78   99.64326 1 vs 2     1
   LR stat.   Pr(Chi)
1
2 0.2709469 0.6026973
```

Some Sources for Ordinal Logistic Regression

- A good source of information on fitting these models is http://www.ats.ucla.edu/stat/r/dae/ologit.htm
    - Another good source, that I leaned on heavily here, using a simple example, is https://onlinecourses.science.psu.edu/stat504/node/177.
    - Also helpful is https://onlinecourses.science.psu.edu/stat504/node/178 which shows a more complex example nicely.
- The asbestos example I discussed comes from Simonoff JS (2003) *Analyzing Categorical Data*. New York: Springer, Chapter 10.