Multinomial regression

Patrick Breheny

April 18

Introduction

- We have used logistic regression to model binary (yes/no) data
- What if we have multiple categories? For example, different forms of a disease, different types of species, or choices from among several alternatives?
- Today we will discuss the generalization of logistic regression (which involved a binomial outcome) to *multinomial regression*, in which the outcome is multinomial

Alligator food choice data

- To illustrate multinomial regression, we'll analyze a study of factors influencing the primary food choice of alligators
- The study involved 219 alligators captured in four Florida lake
- The outcome variable, `Food`, is the primary food type, and consists of five categories:
  - `bird`
  - `fish`
  - `invert`: snails, crayfish, insects, ...
  - `reptile`: turtles, other alligators, ...
  - `other`: amphibians, mammals, plants, ...

Alligator food choice data

- In addition to the lake in which the alligator was captured, we also have information pertaining to the alligator's
  - `Size`: Either `small` ($\leq$ 2.3 meters long) or `large` ($>$ 2.3 meters long)
  - `Sex`
- The question of interest is the effect that these factors have on the primary food type that an alligator chooses to eat

Notation

We will use the following notation in this lecture and the next to describe multi-class models:

- Let Y be a random variable that can take on one of K discrete values (*i.e.*, fall into one of K classes)
- Number the classes $1, \ldots, K$
- Let $\pi_{i2} = \Pr(Y_i = 2)$ denotes the probability that the ith individual's outcome belongs to the second class
- More generally, $\pi_{ik} = \Pr(Y_i = k)$ denotes the probability that the ith individual's outcome belongs to the kth class

Multinomial distribution

- In case you have not seen it before, the *multinomial distribution* is defined as follows:

$$p(Y = \mathbf{y}) = \frac{n!}{y_1! \cdots y_K!} \pi_1^{y_1} \cdots \pi_K^{y_K},$$

  where $\sum_k y_k = n$ and $\sum_k \pi_k = 1$
- Note that for $K = 2$, this reduces to the binomial distribution
- If the data were iid, we could simply fit the multinomial distribution to our data
- However, the purpose of our analysis is to examine the ways in which factors (which vary from alligator to alligator) change $\boldsymbol{\pi}$; hence the name multinomial regression

The multinomial logistic regression model

- Multinomial logistic regression is essentially equivalent to the following:
  - Let $k = 1$ denote the reference category
  - Fit separate logistic regression models for $k = 2, \ldots, K$, comparing each outcome to the baseline:

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \mathbf{x}_i^T \boldsymbol{\beta}_k$$

- Note that this will result in $K - 1$ vectors of regression coefficients (we don't need to estimate the Kth vector because $\sum_k \pi_k = 1$)
- This is the multinomial regression model, although the estimation procedure is complicated by the constraint that $\sum_k \pi_k = 1$

Probabilities and odds ratios

The fitted class probabilities for an observation with explanatory variable vector $\mathbf{x}$ are therefore

$$\hat{\pi}_1 = \frac{1}{1 + \sum_k \exp(\mathbf{x}^T \widehat{\boldsymbol{\beta}}_k)}$$

$$\hat{\pi}_k = \frac{\exp(\mathbf{x}^T \widehat{\boldsymbol{\beta}}_k)}{1 + \sum_l \exp(\mathbf{x}^T \widehat{\boldsymbol{\beta}}_l)}$$

Probabilities and odds ratios

- Like logistic regression, odds ratios in the multinomial model are easily estimated as exponential functions of the regression coefficients:

$$\begin{aligned}\mathrm{OR}_{kl} &= \frac{\pi_k}{\pi_l} = \frac{\pi_k/\pi_1}{\pi_l/\pi_1} \\ &= \frac{\exp\left((\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta}_k\right)}{\exp\left((\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta}_l\right)} \\ &= \exp\left((\mathbf{x}_2 - \mathbf{x}_1)^T (\boldsymbol{\beta}_k - \boldsymbol{\beta}_l)\right)\end{aligned}$$

- In the simple case of changing x_j by δ_j and comparing k to the reference category,

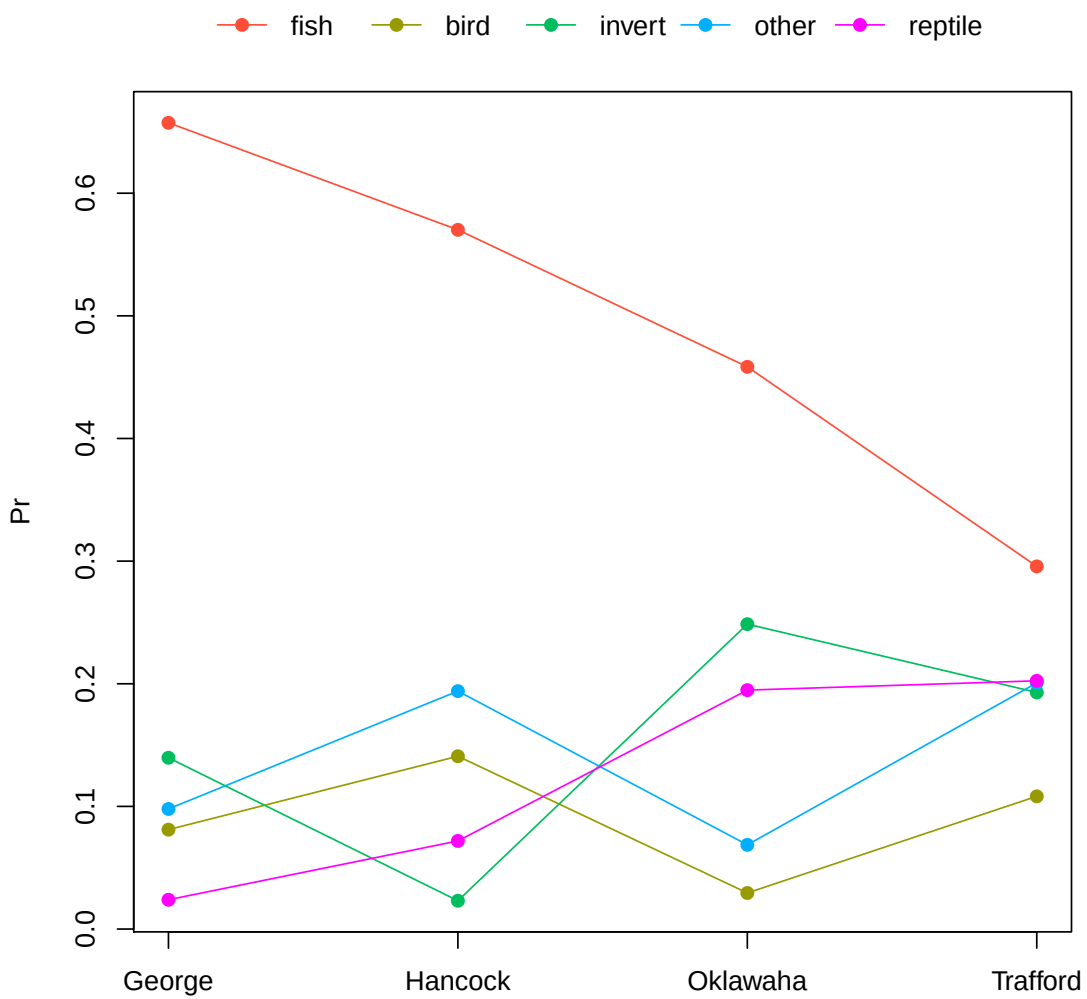
$$\mathrm{OR}_{kl} = \exp(\delta_j \beta_{kj})$$

Some model selection

Model	AIC
Null	612
Size	605
Size + Lake	580
Size + Lake + Sex	586
Size × Lake	587

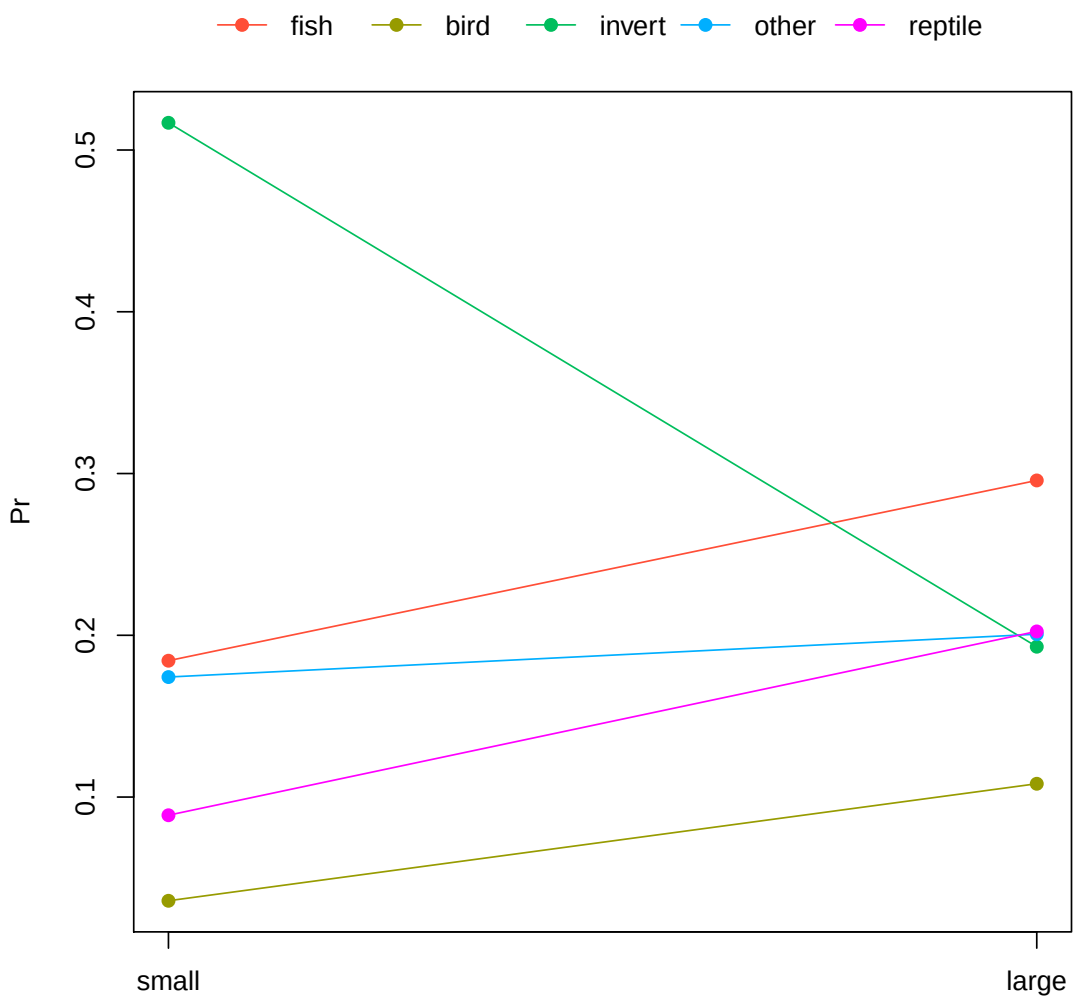
It would seem, therefore, that Size and Lake influence eating preferences, although there is no evidence of an interaction between the two, or any meaningful differences in the eating preferences of male and female gators

Lake


fish
bird
invert
other
reptile
Pr
0.0
0.1
0.2
0.3
0.4
0.5
0.6
George
Hancock
Oklawaha
Trafford

$p < 0.0001$

Size



$$p = 0.0003$$

ORs: Size

Odds ratios comparing large vs. small alligators, with invertebrates as reference group

	OR	95% CI		p
		Lower	Upper	
Bird	8.1	2.1	31.5	0.003
Reptile	6.1	1.9	19.9	0.003
Fish	4.3	2.0	9.3	0.0002
Other	3.1	1.1	8.3	0.03

ORs: Lake

Odds ratios comparing Lake Trafford vs. Lake George, with fish as reference group

		95% CI		
	OR	Lower	Upper	p
Reptile	18.8	2.1	167.9	0.009
Other	4.6	1.3	15.4	0.01
Invert	3.1	1.2	8.0	0.02
Bird	3.0	0.6	15.4	0.2