SS9055B: Generalized Linear Models

Section 9: Multinomial Regression Models I

Objectives

By the end of the lecture you should be able to:

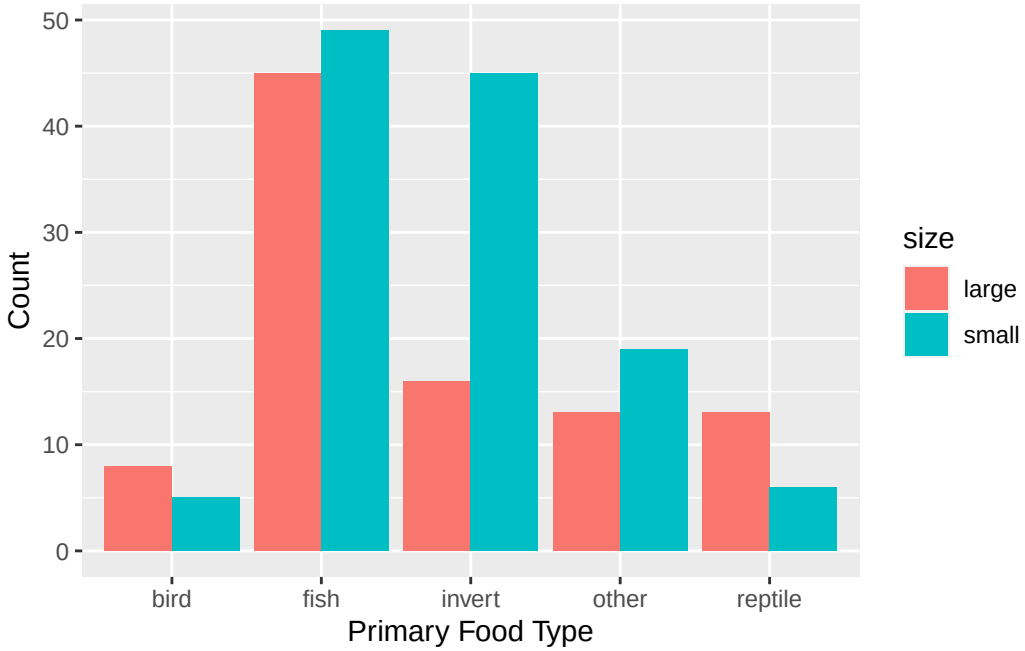
- describe the mathematical structure of a baseline category, multinomial regression model,
- fit multinomial logistic regression models in `R`, and
- interpret the results.

Example

The example we're going to consider is one of my favourite data sets – and I didn't make it up. The data come from a *real* study of the food preferences of alligators living in four lakes in Florida and were collected by Dr. M.F. Delaney and Dr. C.T. Moore. Agresti (2013, 294) describes the experiment as follows:

> "The study captured 219 alligators in four Florida lakes. The nominal response variable is the primary food type, in volume, found in the alligator's stomach. This had five categories: fish, invertebrate, reptile, bird, other."

Agresti goes on to explain that the stomach contents of one alligator contained the "tags of 23 baby alligators released in the lake in the previous year" and that the other category includes "amphibian, mammal, plant material, stones or other debris, or no food or dominant type". The simple answer to the question "What does an alligator eat?" is also the puchline of the joke "What do you feed a 1000 pound gorilla for lunch?" Anything it wants! However, the goal of the analysis is to find out what factors affect the alligators' food preferences, and there are some interesting patterns. In particular, we will model the food choices as a function of each alligator's size, gender, and the lake it was found in. The data are contained in the file `alligator\_data.csv`.



Introduction

The challenge in modelling the alligator data is that the response variable – the type of food that an alligator eats – has more than two categories. If the alligators ate only birds or fish, for example, then we could model the preference with a binomial model. But, alligators eat anything and everything. Another way to analyze this data would be to create a contingency table of the counts by the sex, size, and food type and fit a Poisson log-linear model. However, this has a drawback because all variables in a Poisson log-linear model are treated equally – there is no sense of a response variable and predictor variables. Instead, the Poisson log-linear models reveals associations between the levels of the different variables. In this case, however, the food choice is clearly the response of interest. It is also not clear how to incorporate a continuous predictor into a contingency table, though in this case all of the predictors are categorical. To address these issues, we will consider generalized linear models with multinomial responses for modelling the relationship between a set of predictors and a response variable that is categorical with more than two levels.

Multinomial Distribution as a Multivariate GLM

To develop a GLM for the multinomial response model we need to go through the same steps we did for the binomial and Poisson models. We need to identify the three components:

- Random component
- Systematic component
- Link function. The challenge is that the response is now a vector rather than a scalar. The result will be what we refer to as a vector generalized linear model (VGLM).

Random component

The multinomial distribution is a generalization of the binomial distribution, so will start with a review the binomial case. Suppose that $nY \sim \text{Binomial}(\pi)$ where Y records the proportion of successes in n identical trials. The density of Y is:

$$\begin{aligned} f(y|\pi) &= \binom{ny}{y}\pi^{ny}(1-\pi)^{n(1-y)} \\ &= \exp\left[ny\log\pi + n(1-y)\log(1-\pi) + \log\binom{ny}{y}\right] \\ &= \exp\left[\frac{y\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)}{1/n} + \log\binom{ny}{y}\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{\phi/w} + c(y,\phi)\right] \end{aligned}$$

where $\theta = \text{logit}(\pi)$ is the natural parameter, $b(\theta) = -\log(1-\pi)$ is the cumulant generating function, $\phi = 1$, and $w = n$.

Now consider the multinomial distribution with J categories. In the binomial case we need only 1 random variable to model outcomes in 2 categories (successes and failures). By subtraction, the proportion of failures is $1 - Y$. More generally, we will say that $n\boldsymbol{Y} \sim \text{Multinomial}_J(n, \boldsymbol{\pi})$ if π_j is the probability that a single trial results in outcome j and Y_j records the proportion of times that outcome j occurs on n identical trials of the same experiment. Although $\boldsymbol{Y}$ and $\boldsymbol{\pi}$ are both vectors of length J, we can define the final element of each vector by subtraction

$$Y_J = 1 - \sum_{j=1}^{J-1} Y_j$$

and

$$\pi_J = 1 - \sum_{j=1}^{J-1} \pi_j.$$

This means that we only need to model the counts in the first $J-1$ categories.

If $\boldsymbol{Y}$ is multinomial then its density is:

$$\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\pi}) &= \frac{n!}{ny_1!\cdots ny_J!}\prod_{j=1}^{J}\pi_j^{ny_j} \\
&= \frac{n!}{ny_1!\cdots ny_J!}\prod_{j=1}^{J-1}\pi_j^{ny_j}\cdot\pi_J^{n(1-\sum_{j=1}^{J-1}y_j)} \\
&= \exp\left[n\sum_{j=1}^{J-1}y_j\log\pi_j + n(1-\sum_{j=1}^{J-1}y_j)\log(\pi_J) + \frac{n!}{ny_1!\cdots ny_J!}\right] \\
&= \exp\left[\frac{\sum_{j=1}^{J-1}y_j\log\left(\frac{\pi_j}{\pi_J}\right)+\log(\pi_J)}{1/n} + \frac{n!}{ny_1!\cdots ny_J!}\right] \\
&= \exp\left[\frac{\boldsymbol{y}'\boldsymbol{\theta}-b(\boldsymbol{\theta})}{\phi}+c(y,\phi)\right].
\end{aligned}$$

This looks similar to the exponential family form we had before, but it is now in the form of a multivariate exponential family where $\boldsymbol{y}$ and $\boldsymbol{\theta}$ are both vectors of length $J-1$. The natural parameter is the $(J-1)$-vector

$$\boldsymbol{\theta} = \left(\log\frac{\pi_1}{\pi_J},\ldots,\log\frac{\pi_{J-1}}{\pi_J}\right)',$$

the cumulant generating function is

$$b(\boldsymbol{\theta}) = -\log(\pi_J).$$

which is really

$$b(\boldsymbol{\theta}) = \log\left[1+\exp\left(\sum_{j=1}^{J-1}\theta_j\right)\right]$$

because

$$\pi_J = \frac{1}{1+\exp(\sum_{j=1}^{J-1}\theta_j)}.$$

Finally, $\phi = 1$ and $w = n$.

Link Function

As with the binomial model, we can obtain the canonical link function. In the case of the binomial model $\mu = \pi$ and

$$\theta = \log\left(\frac{\pi}{1-\pi}\right) \text{ so that } \pi = \frac{\exp(\theta)}{1+\exp(\theta)}.$$

To obtain the canonical link we equate the natural parameter and linear predictor, which yields

$$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \text{ or } g^{-1}(\mu) = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

In the case of the multinomial model the mean of $\boldsymbol{Y}$ is

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{J-1})' = (\pi_1, \ldots, \pi_{J-1})' = \boldsymbol{\pi}$$

and the natural parameter is

$$\boldsymbol{\theta} = \left(\log\frac{\pi_1}{\pi_J}, \ldots, \log\frac{\pi_{J-1}}{\pi_J}\right)' = \left(\log\frac{\mu_1}{\mu_J}, \ldots, \log\frac{\mu_{J-1}}{\mu_J}\right)'.$$

where $\mu_J = 1 - \sum_{j=1}^{J-1}\mu_j$, as you probably guessed. To obtain the canonical link we set $\boldsymbol{\theta}$ equal to the linear predictor, which must now be a vector of length $J-1$ itself, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{J-1})'$. This yields

$$\boldsymbol{\eta} = \boldsymbol{\theta} = \left(\log\frac{\mu_1}{\mu_J}, \ldots, \log\frac{\mu_{J-1}}{\mu_J}\right)'$$

so that the canonical link function is

$$g(\boldsymbol{\mu}) = \left(\log\frac{\mu_1}{\mu_J}, \ldots, \log\frac{\mu_j}{\mu_J}\right)'.$$

Note that the link function maps $\Re^{J-1}$ to $\Re^{J-1}$ (i.e., there are $J-1$ elements in the mean and $J-1$ elements in the linear predictor).

Systematic Component

Finally, we must define the systematic component (linear predictor). In the logistic regression model, the linear predictor is

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \boldsymbol{x}'\boldsymbol{\beta}.$$

However, for the multinomial model we require the linear predictor to be a vector of length $J-1$. Given a vector of predictors, $\boldsymbol{x}$, it makes sense to allow each of the means to depend on these predictors but with different values of the regression coefficients so that

$$\eta_j = \boldsymbol{x}'\boldsymbol{\beta}_j, \quad j = 1, \ldots, J-1.$$

We can write this in matrix form as:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{B}$$

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}' & & & \\ & \boldsymbol{x}' & & \\ & & \ddots & \\ & & & \boldsymbol{x}' \end{pmatrix}$$

and

$$\boldsymbol{B} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_{J-1}')'.$$

We can also write $\boldsymbol{X}$ in short-form notation as $(I_{J-1} \otimes \boldsymbol{x}')$, which is called the Kronecker product.

Complete Model

The complete model is formed by allowing for observations from N different individuals with separate values of the covariates. Specifically, we will assume that we have N independent observations, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ where each $\boldsymbol{Y}_i$ represents a single multinomial trial with probabilities $\boldsymbol{\pi}_i$ dependent on a vector of covariates $\boldsymbol{x}_i$. Mathematically:

$$\boldsymbol{Y}_i \sim \text{Multinomial}_J(1, \boldsymbol{\pi}_i)$$

where

$$\log \frac{\pi_{ij}}{\pi_J} = \eta_{ij} = \boldsymbol{x}_i' \boldsymbol{\beta}_j, \quad j = 1, \ldots, J-1.$$

Alternatively, we can write this in matrix form as

$$\boldsymbol{\eta} = g(\boldsymbol{\pi}_i) = \left(\log \frac{\pi_{i1}}{\pi_{iJ}}, \ldots, \log \frac{\pi_{iJ-1}}{\pi_{iJ}} \right)'$$

and

$$\boldsymbol{\eta} = \boldsymbol{X}_i \boldsymbol{B}$$

with

$$\boldsymbol{X}_i = (I_{J-1} \otimes \boldsymbol{x}_i') \text{ and } \boldsymbol{B} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_{J-1}')'.$$

Note here that the design matrix, $\boldsymbol{X}_i$, varies by individual depending on the value of the covariates, but the vector of regression coefficients, $\boldsymbol{B}$, remains constant. If the linear predictor includes the intercept and p covariates then the first $p+1$ elements of $\boldsymbol{B}$ model the relative values of π_{i1} and π_{iJ}, the next $p+1$ elements model the relative values of π_{i2} and π_{iJ}, etc.

Baseline Category

This model is called a baseline or reference category model because it treats the last category as a reference and compares the probabilities of the first $J-1$ outcomes, $\pi_1, \ldots, \pi_{J-1}$, to the probability of the last outcome, π_J. The parameter β_{jk} models the increase in the log of the relative probabilities of outcomes j and J when x_k increases by one unit. Equivalently, $\exp(\beta_j)$ represents the multiplicative increase in the odds when x_k increases by one unit.

The choice of baseline category is arbitrary because the log-odds between any two categories can easily be calculated regardless of which category is chosen as the reference. We have written the model so that the last category is the reference category, but we can order the categories any way we choose. Choosing another level of the response as the baseline does not change the results. Mathematically:

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_k(\boldsymbol{x})} = \log \frac{\pi_j(\boldsymbol{x})/\pi_J(\boldsymbol{x})}{\pi_k(\boldsymbol{x})/\pi_J(\boldsymbol{x})} = \log \frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} - \log \frac{\pi_k(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}.$$

In terms of the coefficients, this implies that

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_k(\boldsymbol{x})} = \boldsymbol{x}'\boldsymbol{\beta}_j - \boldsymbol{x}'\boldsymbol{\beta}_k = \boldsymbol{x}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)'.$$

where β_j and β_k refer to the parameters of the model with baseline category J. If we were to use category k as the reference level and model

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_k(\boldsymbol{x})} = \boldsymbol{x}'\boldsymbol{\beta}_j^{*\prime}, \quad j \neq k$$

then we find that

$$\hat{\boldsymbol{\beta}}_j^* = \hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_k.$$

This shows that we can obtain estimates for the parameters in the model with baseline category k by taking differences of the estimates of the parameters from the model with baseline category J and has two implications. First, it doesn't matter which category we choose as the reference. The fitted probabilities will be the same, but we will have to do more work to answer specific questions depending on which category we choose. Second, there is no need to refit the model if you want to answer questions about the relationship between two categories where neither is the baseline.

Example: Alligator Food Preferences

There are several different packages that provide functions for fitting multinomial models in `R`. All of them are intended to do other things and none of them provide all of the output

available from `glm()`, which I find frustrating. The one that I prefer for fitting standard multinomial models is the function `vglm()` from the `VGAM` package. The following code models the alligators' food preferences as a function of their size and sex, ignoring the lake they come from. The command looks very similar to fitting a binomial logistic regression model, except for the change in the name of the function and the family argument

```
## Load package
library(VGAM)

## Fit multinomial regression model (Method 1)
vglm1 <- vglm(food ~ size + sex, data = alligators, family = multinomial())
```

As for binomial models, the same model can be fit by providing a matrix with summary counts of each outcome of the response in the columns with the different combinations of the predictors indexing the rows. Also as for binomial models, you need to consider the summary data in order to compute meaningful residuals and assess the fit of the model. If the data had included a continuous covariate, say the exact weight of the alligator, then we would have had to fit the model with individual data first to estimate the parameters and then fit the model again after binning the age to assess the model's fit. Here is the alternative method of fitting the same model

```
## Fit multinomial regression model (Method 2)
alligators1 <- group_by(alligators,food,lake,sex,size) |>
  summarize(Count=n()) |>
  spread(key=food,value=Count,fill=0)

vglm2 <- vglm(cbind(bird,fish,invert,other,reptile) ~ size + sex,
              data = alligators1, family = multinomial())
```

Output from this model is presented in Table 1 with confidence intervals in Table 2. The first thing that you need to know is which level is the reference. By default, `vglm()` will use the last level in alphabetical order as the reference. In this case, the reference category is `reptile`. The other categories are then numbered in alphabetical order so that category 1 is `bird`, category 2 is `fish` etc. You can change the reference level with the argument `refLevel` in the `multinomial()` function.

The first thing that we should do is to test the goodness-of-fit of the model, and we can do this again with the deviance goodness-of-fit test. The test is appropriate in this case because all of the covariates are fixed and the number of categories does not depend on the sample size. The residual deviance is 100.58 on 52 degrees of freedom which results in a p-value which is very small ($< .0001$). Based on this test, we must reject the fit of the model. We likely need to include the interaction between size and sex, but this will add 16 terms to the model (eek!) so I'm not going to do that now.

Notice that the output contains three parameters for each of the food choices other than the reference category: an intercept, an effect of size, an effect of sex. Each of these models how the preference for that food choice changes relative to the preference for eating reptiles when the covariate changes. For example, based on this model we would estimate that the log-odds that a large, female alligator eats birds (Level 1) relative to reptiles (Level 5) is -.06 with 95%CI (-1.48,1.35) and p-value .932 for the Wald-type test that this coefficeint is zero. The conclusion is that there is no evidence that large, female alligators prefer eating birds to eating reptiles. Similarly, we estimate that the difference between the log of the probability of eating birds and the log of the probability of eating reptiles is .11 with 95%CI (-1.45,1.67) and p-value .89. Equivalently, we can say that we estimate that the probability of eating birds relative to the probability of eating reptiles is $e^{.11} = 1.12$ times higher for small alligators than for large alligators with 95% CI (.23,5.31). Since the first confidence interval is very wide and easily covers 0 (equivalently, the second covers 1) we have no evidence that the preference for birds in comparison to reptiles depends on size. Similarly, there is no evidence of an effect of sex on the preference for birds over reptiles (p=.460). Based on these results, there is no reason to believe that alligators in any of the groups prefer to eat birds over fish or vice versa.

Comparisons between the preference for reptiles to the other food choices can be conducted in the same way. The most significant effect is that modelling the effect of size on the probability of eating invertebrates (category 3) relative to the probability of eating reptiles. The model suggests that the relative probability is $e^{1.79} = 5.99$ times higher for small alligators than for large alligators (95%CI = 1.84,19.46). The confidence interval is wide, but does not cover 1 and the message is clear. Small alligators are more likely to eat invertebrates. This makes sense. Invebrates (things without spines) are small and easily eaten by small alligators. Reptiles are big and make a nice lunch for big alligators.

Table 1: Results of fitting the multinomial model with default baseline category to the alligator data.

```
Call:
vglm(formula = cbind(bird, fish, invert, other, reptile) ~ size +
    sex, family = multinomial(), data = alligators1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -0.06189    0.72102  -0.086  0.93160
(Intercept):2  1.21698    0.54553   2.231  0.02569 *
(Intercept):3  0.25318    0.59593   0.425  0.67095
(Intercept):4  0.12640    0.63329   0.200  0.84180
sizesmall:1    0.11450    0.79616   0.144  0.88565
sizesmall:2    0.86801    0.56292   1.542  0.12308
sizesmall:3    1.78909    0.60159   2.974  0.00294 **
sizesmall:4    1.10100    0.64276   1.713  0.08673 .
sexmale:1     -0.58333    0.78409  -0.744  0.45690
sexmale:2      0.03154    0.56938   0.055  0.95582
sexmale:3     -0.05866    0.59665  -0.098  0.92168
sexmale:4     -0.16483    0.64541  -0.255  0.79843
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]),
log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])

Residual deviance: 100.5817 on 52 degrees of freedom

Log-likelihood: -98.4811 on 52 degrees of freedom

Number of Fisher scoring iterations: 4

No Hauck-Donner effect found in any of the estimates


Reference group is level  5  of the response
```

Table 2: Confidence intervals obtained by fitting the multinomial model with default baseline category to the alligator data.

```
                   2.5 %    97.5 %
(Intercept):1 -1.4750608 1.3512906
(Intercept):2  0.1477623 2.2862049
(Intercept):3 -0.9148284 1.4211849
(Intercept):4 -1.1148269 1.3676190
sizesmall:1   -1.4459477 1.6749415
sizesmall:2   -0.2353018 1.9713231
sizesmall:3    0.6099845 2.9681898
sizesmall:4   -0.1587830 2.3607749
sexmale:1     -2.1201306 0.9534653
sexmale:2     -1.0844229 1.1475061
sexmale:3     -1.2280696 1.1107443
sexmale:4     -1.4298014 1.1001458
```

Relationship between Multinomial and Poisson

Although I chose to fit the alligator data with a multinomial regression model instead of a Poisson regression model there is a very close tie between the two. Suppose that $Y_1, \ldots, Y_K$ are independent Poisson random variables such that $E(Y_k) = \mu_k$. Let $n = \sum_{k=1}^K Y_k$. Note that n is also Poisson with mean $\mu = \sum_{k=1}^K \mu_k$. The joint distribution of $Y_1, \ldots, Y_K$ conditional on n has pdf

$$\begin{aligned} f(\boldsymbol{y}|n, \boldsymbol{\mu}) &= \frac{\prod_{k=1}^K e^{\mu_k} \mu_k^{y_k}/y_k!}{e^{\mu}\mu^n/n!} \\ &= \frac{n!}{y_1! \cdots y_K!} \prod_{k=1}^K \frac{\mu_k}{\mu}^{y_k} \end{aligned}$$

which shows that $\boldsymbol{Y}|n \sim \text{Multinomial}(n, \boldsymbol{\pi})$ where $\boldsymbol{\pi} = (\mu_1/\mu, \ldots, \mu_K/\mu)'$.

What this means is that fitting a Poisson model to a contingency table is essentially the same as fitting a multinomial model. The difference is that the Poisson model treats the sample size as random (i.e., not fixed before the experiment) while the multinomial treats the sample size as fixed. In this case we should probably have fit a Poisson model since it is hard to believe that the researchers pre-specified a sample size of 219. It's more likely that this is simply the number they were able to catch. However, if what we are interested in is the association between two variables then it doesn't matter if we fit a multinomial regression or a Poisson log-linear model. Our inferences will be the same.

References

Agresti, Alan. 2013. *Categorical Data Analysis*. 3rd ed. Wiley.