SS9055B: Generalized Linear Models

Section 10: Multinomial Regression Models 2

Objectives

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

- list different link functions for the multinomial model,
- interpret the coefficients of models fit with the different link functions,
- fit and interpret discrete choice models.

Alternative Link Functions

We saw in the last section of notes that the canonical link for the multinomial model is with J categories the baseline logistic function:

$$g_j(\boldsymbol{\pi}) = \log\left(\frac{\pi_j}{\pi_J}\right) = \eta_j$$

for $j = 1, \ldots, J$. However, there are several different link functions that are commonly used with multinomial data and might be more appropriate depending on how the data is collected the system under study behaves. This is especially true when the response is an ordinal variable meaning that it is categorical but ordered. A common example is a rating on a Likert scale (e.g., rating preferences on a scale of 1 to 5 or 1 to 10). It doesn't really make sense to consider a reference category in this case and, for example, to compare the probability of rating something as a 2, 3, 4, or 5 to the baseline probability of rating something as a 1. It makes more sense to model the probabilities in some ordered way. One possibility is to model the probability that someone assigns a rating of 4 or lower, 3 or lower, 2 or lower, or 1 or lower. The link function that achieves this is called the cumulative link function. Alternatively, you might model the probability that someone assigns a rating of 2 relative to the probability of assigning a rating of 1, then the probability of assigning a rating of 3 relative to the probability of assigning a rating of 2, etc. This is called the adjacent-categories model. A third choice, called the continuation ratio, models the conditional probability of assigning a rating of 2 given

that the rating is 2 or higher, the probability of assigning a 3 given that the rating is three or higher etc.

Cumulative Link Models

The cumulative models the log odds of being in category j or lower for $j = 1, \ldots, J-1$ and is probably the most common alternative to the baseline category model[1]. The link function is:

$$g_j(\boldsymbol{\pi}) = \text{logit}P(Y \leq j) = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \log\left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right)$$

for $j = 1, \ldots, J-1$. As in the baseline category model, we will allow a different set of coefficients for each $j = 1, \ldots, J-1$ creating different linear predictors so that

$$g_j(\boldsymbol{\pi}) = \eta_j = \boldsymbol{x}'\boldsymbol{\beta}_j$$

Solving for $P(Y \leq j)$ we get

$$P(Y \leq j) = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)}$$

from which it follows that

$$\begin{aligned}
\pi_j =& P(Y = j) \\
=& P(Y \leq j) - P(Y \leq j-1) \\
=& \frac{\exp(\eta_j)}{1 + \exp(\eta_j)} - \frac{\exp(\eta_{j-1})}{1 + \exp(\eta_{j-1})}.
\end{aligned}$$

The result is that the individual cell probabilities are defined through the differences of the cumulative probabilities, which depend on separate linear predictors. The probability π_J is excluded from these calculations because $P(Y \leq J) = 1$ by definition. Once again, we compute π_J by subtraction, $\pi_J = 1 - \sum_{j=1}^{J} \pi_j$.

Models that use this link function can be interpreted in relation to a series of cumulative Bernoulli trials. The first linear predictor models the probability that an observation falls into the first cell:

$$g_1(\boldsymbol{\pi}) = \log\left(\frac{P(Y \leq 1)}{1 - P(Y \leq 1)}\right) = \eta_1 = \boldsymbol{x}'\beta_1.$$

The second linear predictor models the probability that the observation falls into the first or second cell

$$g_2(\boldsymbol{\pi}) = \log\left(\frac{P(Y \leq 2)}{1 - P(Y \leq 2)}\right) = \eta_2 = \boldsymbol{x}'\beta_2,$$

and so on.

[1] You can model the probability of being in category j or higher in exactly the same way.

As with the baseline category models, there are many regression coefficients in this model. One common approach to simplify these models is to assume that the effects of the predictors are the same for all j and only allow the intercept to change. The linear predictor becomes:

$$\eta_j = \beta_{0j} + \beta_1 x_1 + \ldots + \beta_p x_p, \quad j = 1, \ldots, J-1.$$

This new model has only $J-1+p$ regression coefficients, instead of $(J-1)(p+1) = J-1+Jp$, and is referred to as the parallel model or the proportional odds model.

Adjacent-Categories Model

The adjacent categories models the probability that a response falls into one of two adjacent cateogries, i.e., the probability that a response falls into category j given that it is in either category j or category $j+1$. The link function is:

$$g_j(\boldsymbol{\pi}) = \text{logit} P(Y = j | Y = j \text{ or } Y = j+1) = \log\left(\frac{P(Y=j)}{P(Y=j+1)}\right) = \log\left(\frac{\pi_j}{\pi_{j+1}}\right) = \eta_j$$

for $j = 1, \ldots, J-1$. The fitted probabilities are given by:

$$\pi_j \propto \exp\left(\sum_{k=j}^{J-1} \eta_k\right)$$

where the empty sum is taken to be equal to 1 when $j = J$. This implies:

$$\begin{aligned}
\pi_1 &= \frac{\exp(\sum_{k=1}^{J-1}\eta_k)}{\exp(\sum_{k=1}^{J-1}\eta_k) + \exp(\sum_{k=2}^{J-1}\eta_k) + \cdots + 1} \\
\pi_2 &= \frac{\exp(\sum_{k=2}^{J-1}\eta_k)}{\exp(\sum_{k=1}^{J-1}\eta_k) + \exp(\sum_{k=2}^{J-1}\eta_k) + \cdots + 1} \\
&\vdots \\
\pi_J &= \frac{1}{\exp(\sum_{k=1}^{J-1}\eta_k) + \exp(\sum_{k=2}^{J-1}\eta_k) + \cdots + 1}.
\end{aligned}$$

It turns out (though we won't show this) that there is a correspondance between the adjacent-category and the baseline category model. The parameters in an adjacent-category model can be written as linear combinations of the parameters in a baseline category model. However, the adjacent-category model accounts for the ordering of the response and makes more sense when fitting some restricted models. This includes the parallel model obtained by setting

$$\eta_j = \beta_{0j} + \beta_1 x_1 + \ldots + \beta_p x_p, \quad j = 1, \ldots, J-1.$$

as in the previous section. With the adjacent categories link this model implies that increasing the value of a covariate has the same effect on the relative values of π_j and π_{j+1}.

Continuation-Ratio

The final choice we will consider is the continuation ratio which models the probability that a response falls into category j given that it is in category j or higher. The link function is:

$$g_j(\boldsymbol{\pi}) = \text{logit}\left[P(Y = j|Y \geq j)\right] = \log\left(\frac{\pi_j}{\pi_j + \cdots + \pi_J}\right) = \eta_j$$

for $j = 1, \ldots, J-1$. The fitted probabilities are given by:

$$\begin{aligned}
\pi_1 &= \exp(\eta_1) \\
\pi_2 &= \exp(\eta_2)(1 - \exp(\eta_1)) \\
\pi_3 &= \exp(\eta_3)(1 - \exp(\eta_2)(1 - \exp(\eta_1)) - \exp(\eta_1)) \\
&\vdots
\end{aligned}$$

The continuation ratio is often applied to model progression of individuals through some sort of sequence of states. For example, the continuation-ratio would make sense to model survival of people at different ages.

I Scream, You Scream

Here's another fun example. The data come from a study of people's preferences for the amount of fat in ice cream conducted at Pennsylvania State University. The researchers tested ice cream with 8 different fat levels and participants rated the ice cream on a scale from 1 (Yuck!) to 9 (Yum! Yum!). The researchers expected that there is an optimal level of fat somewher in the middle of the range, so we will model the outcome using a quadratic function of the fat concentration.

The data is provided in the file `ice\_cream.csv`. The columns record the number of people (Count) who provided each rating (Rating) for each fat level (Fat). Figure 1 shows the observed proportion of people rating the ice cream at the lowest, highest, and middle levels for each fat content. It appears that more people rate the ice cream lower when the fat level is very low or very high, and higher when the fat level is somewhere in the middle.

As I mentioned, there are several packages in `R` that can be used to fit multinomial models, each with strengths and weaknesses. I'll continue to use the `VGAM` package because it includes the ability to fit cumulative logit models. The model will have 8 separate sets of coefficients because there are 9 different categories in the rating. I have treated fat content as a continuous predictor and fit two models including both linear and quadratic terms. The first is a parallel model which has 10 parameters including separate intercepts for each of the linear predictors plus common linear and quadratic effects of fat content. The second is an unrestricted model which has 24 parameters including separate intercepts, linear, and quadratic effects for each

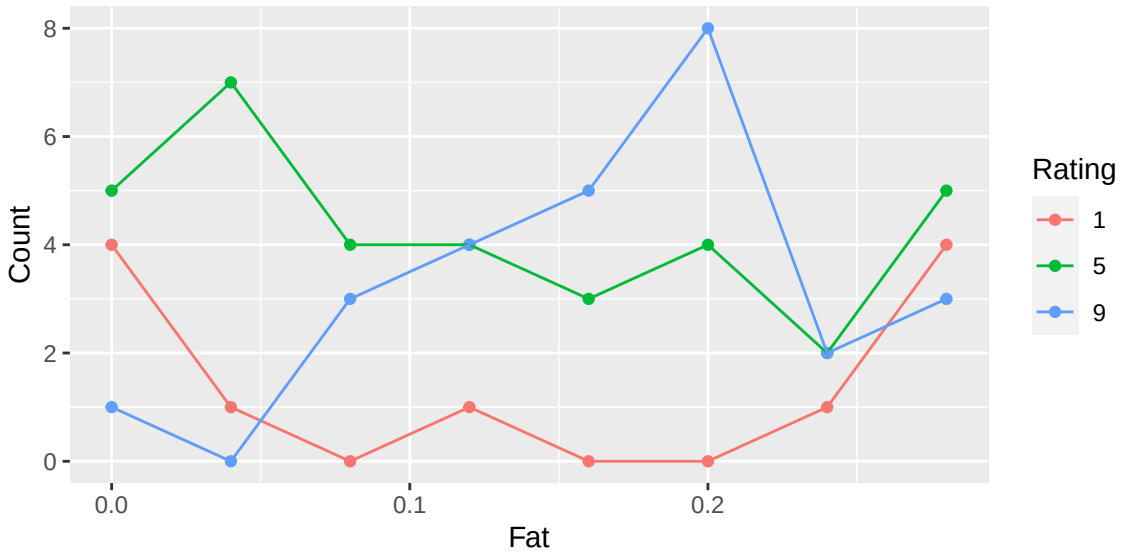

Figure 1: Proportion of ratings of 1 (Yuck!), 5, or 9 (Yum! Yum!) for each fat level.

of the linear predictors. Thankfully the LRT indicates that there is no significant difference between the models (p=.50), and so the parallel model is preferred.

Figure 2 shows the fitted probabilities from the parallel model for the lowest, highest, and middle ratings. The results indicate that the lowest ratings are more likely at the lowest and highest fat levels while the highest ratings occur at the middle fat levels. This clearly supports the hypothesis that people prefer ice cream with moderate fat content.

Discrete Choice Models

One restriction of the multinomial models we have fit so far is that they assume a single set of explanatory variables for each individual. Regardless of what link function we chose, the j^{th} element of the linear predictor for individual i was defined as

$$\eta_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta}_j$$

so that covariates vary only by the individual, i, and the coefficients vary by the category of the response, j.

However, there are many situations in which we might be interested in studying how the probabilities are affected by the characteristics of the categories themselves. Suppose for example that we wanted to model a person's choice of meal in a restaurant. The models we have examined so far would allow us to study how differences between people affect their choices. E.g., we might be interested to know if sex, age, or weight affects a persons probability

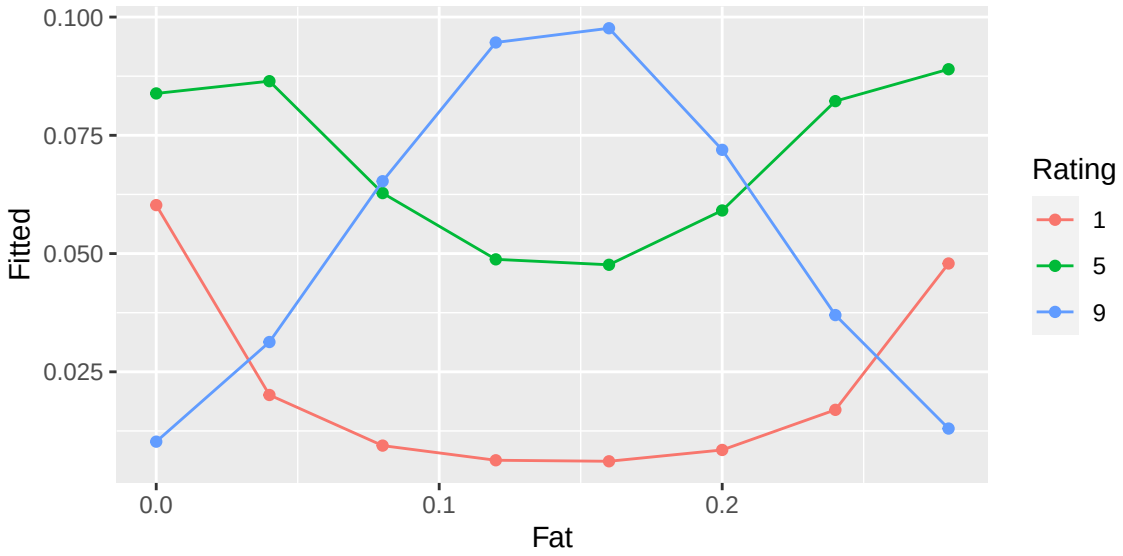

Figure 2: Fitted probabilities for the ratings of 1 (Yuck!), 5, or 9 (Yum! Yum!) for each fat level.

of choosing salad over steak. These explanatory variables are called *characteristics of the chooser*. However, we might also want to study how a subject's choice is affected by the certain features of choices themselves. If the people we are studying are eating in different restaurants then the cost of the steak and the salad might vary, and we might want to know how the cost of a meal affects a subject's choice. Alternatively, we might want to know if choices depend on the fat content of a dish. Such explanatory variables are called *characteristics of the choice*.

The discrete choice model expands on the multinomial logistic regression model by adding a new set of covariates and even more predictor variables. Suppose that we have observations from i individuals who can each choose from J categories. Once again we assume the multinomial model with the baseline category link function so that

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

with

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \eta_{ij}.$$

However, we are now going to add a second set of predictor variables to the linear predictor variables, z_{ij}, to the linear predictor so that

$$\eta_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{z}_{ij}'\boldsymbol{\alpha}$$

and setting

$$\pi_{ij} = \frac{\exp(\eta_{ij})}{\sum_{j=1}^{J}\exp(\eta_{ij})}.$$

The extra predictor variables depend upon both the subject, i, and the category, j, and model the characteristics of the choice. Note that the vector of coefficients associated with these predictors, $\boldsymbol{\alpha}$, must be assumed to be constant in order to keep the number of parameters in check and ensure that the model is identifiable.

In this new model, the log-odds between any two categories j and k in $1, \ldots, J$ is given by

$$\log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \boldsymbol{x}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k) + (\boldsymbol{z}_{ij} - \boldsymbol{z}_{ik})'\boldsymbol{\alpha}$$

or equivalently

$$\frac{\pi_{ij}}{\pi_{ij} + \pi_{ik}} \Bigg/ \frac{\pi_{ik}}{\pi_{ij} + \pi_{ik}} = \exp\left[\boldsymbol{x}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k) + (\boldsymbol{z}_{ij} - \boldsymbol{z}_{ik})'\boldsymbol{\alpha}\right].$$

Given that an individual chooses from one of two categories, labelled j and k, then their final choice depends both on their own characteristics, defined through $\boldsymbol{x}_i$, and on the difference between the characteristics of the choice for the two categories, $(\boldsymbol{z}_{ij} - \boldsymbol{z}_{ik})$. Note that the characteristics of the choices may not be the same for two individuals even if the names of the categories available to these individuals are the same. In the restaurant example, the costs and the fat content of salads and steak will vary between restaurants, but the menu listings will be the same. As another example, we might study how both income (a characteristic of the choosers) and cost (a characteristic of the choices) affects peoples' transportation choices on their commute to work. In this case, the cost of one type of transportation, say taking the bus, will depend on where in the city people live.

Example: Ketchup

The following example originally comes from Kim et al.~(1995). The researchers conducted a study to determine the effects of price on brand preferences for several products, including tomato ketchup. The data set contains information on the preference of ketchup brands for 1956 US individuals. Each individual was allowed to choose from one of four brands: Heinz, Hunts, Delmonte or the store brand. We are interested in the effect of price on peoples' choice and the price varies both by brand and by individual, presumably because different stores price the same products differently.

In this case, the model has only a single predictor, price, which is a characteristic of the choice. If we let z_{ij} denote the price of the j^{th} brand of ketchup for the i^{th} consumer and assume the baseline category link function then the probability that the i^{th} customer chooses the j^{th} brand of ketchup will be given by

$$\log\left(\frac{\pi_{ij}}{\pi_{i4}}\right) = \beta_{0j} + \alpha z_{ij}.$$

The parameter β_{0j} models the preference for the different ketchups when $z_{ij} = 0$ (i.e., when there is no cost) and α models how the relative preferences vary with cost. Since α is constant,

the model assumes that changes in cost have the same effect on the relative preferences for all individuals and all brands.

Table 1 provides output from fitting a model with price considered in units of $.10. The baseline category represents the first brand alphabetically, Delmonte. Note that this requires a new package, `mlogit`, since characteristics of the choices cannot be included in the `vglm()` function from `VGAM`. The results indicate that if price was no question (i.e., all the Ketchups were the same price) then there is a strong preference for Heinz and Hunts over Delmonte and the store brand. However, price also has a strong, negative effect on peoples choices. The log-odds that someone chooses Heinz, Hunts, or the store brand relative to Delmonte decreases by .45 (95%CI=.41,.49) for each ten cent increase in the price ($p < .001$).

% “‘{r}»= % ## Create new data % test_data <- crossing(heinz=prices, % hunts=prices, % delmonte = prices, % stb=prices) %>% % rowid_to_column(var=“chid”)

% test_data_wide <- test_data %>% % gather(key=“alt”,value=“price”,-chid) %>% % arrange(chid)

% test_predict <- bind_cols(test_data, % as_tibble(predict(Ketchup.mn,newdata = test_data_wide))) % @

References

Kim, Byong-Do, Robert C. Blattberg and Peter E. Rossi (1995) “Modeling the distribution of price sensitivity and implications for optimal retail pricing”, `Journal of Business Economics and Statistics`, 13(3), 291.

Table 1: Summary of discrete choice model for ketchup preferences.

```
Call:
mlogit(formula = Ketchup.choice ~ price | 1, data = Ketchup.wide,
    method = "nr")

Frequencies of alternatives:choice
delmonte    heinz    hunts      stb
0.038855 0.640082 0.187117 0.133947

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 9.55E-07
gradient close to zero

Coefficients :
                   Estimate Std. Error  z-value  Pr(>|z|)
(Intercept):heinz  2.149077   0.125572  17.1143 < 2.2e-16 ***
(Intercept):hunts  1.223044   0.133398   9.1684 < 2.2e-16 ***
(Intercept):stb   -0.879805   0.158420  -5.5536 2.798e-08 ***
price             -0.451678   0.019718 -22.9070 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1632.6
McFadden R^2:  0.16086
Likelihood ratio test : chisq = 625.92 (p.value = < 2.22e-16)

                       2.5 %     97.5 %
(Intercept):heinz  1.9029598  2.3951937
(Intercept):hunts  0.9615894  1.4844993
(Intercept):stb   -1.1903017 -0.5693084
price             -0.4903238 -0.4130312
```