**Multinomial Logistic Regression**

Background and Examples, Multinomial Logistic Regression in R, Interpreting Coefficients,

Visualizing and Communicating Results

Prepared by Allison Horst

for ESM 244

Bren School of Environmental Science & Management, UCSB

**Introduction, Background and Examples**

**Multinomial logistic regression** is used to explore relationships between one or more independent variables (continuous or categorical) and a nominal dependent variable (i.e., a variable with multiple outcomes that do not have an order, value or rank). Binary logistic regression is a particular form of multiple logistic regression when there are only two possible outcomes. Logistic regression is particularly useful in determining the *likelihood of a categorical outcome occurring (with respect to a reference outcome) based on the value or category of input variable(s)*.

Examples that may benefit from analysis by multinomial logistic regression:

***Example 1.*** You are trying to explore the relationships between prior work experience (years), marriage status (married, single, divorced) and income ($/yr) on the likelihood that a person will choose to buy a Honda, Toyota, or Ford.

*Independent measurement variables: work experience (continuous), marriage status (categorical), income (continuous). Dependent variable: car type, with three nominal outcomes – Honda, Toyota, Ford.*

***Example 2.*** A researcher is trying to investigate effects of gender (Male/Female), highest education level (HS, some college, college degree), and Age on likelihood to register as Democrat, Republican or Independent.

*Independent categorical variables: gender (binary), education level (categorical), age (continuous). Dichotomous dependent variable: registration choice (three nominal outcomes: Democrat, Republican, Independent).*

Logistic regression explores the *odds* of an event occurring. The *odds ratio* is the probability of an event occurring divided by the probability of the event *not* occurring:

The *log odds* of an event (also called the *logit*) are given by:

where the coefficients (*β*) are determined from the logistic regression model (see approach in *R* below), and the variables (*x1, x2, …xn*) are the values of the independent variables (or dummy variables). If the independent variables are *measured*, the values are input directly. If the independent variables are categorical, the use of ***dummy variables***is required. Dummy variables (typically) assign a value of 0 or 1 to categorical input variables and dichotomous dependent variables, and are automatically created in R.

One different between multinomial logistic regression and other types of regression (binary logistic, ordinal logistic) is that the odds of an outcome occurring are *with respect to the reference outcome*. This means that you need to select the reference outcome for the outcome variable (R will automatically use the first outcome listed in your data frame as the reference variable unless you tell it to do otherwise).

**Multinomial Logistic Regression in *R***

\*\**You can download the .csv file ‘SchoolChoice.csv’ from the website to follow along with this example.*

*\*\*You will need the following packages installed and loaded to follow along: nnet, ggplot2, reshape2*

So now you (hopefully!) have an idea of when multinomial logistic regression can be useful. Let’s first figure out *how* to do it in R, then we’ll interpret, visualize and communicate the results in a way that would be useful for an audience.

Luckily, once you understanding the basics of logistic regression, finding the coefficients in *R* is pretty straightforward. We’ll be using the multinom() function in the ‘nnet’ package, so make sure you have that package loaded in your active RStudio workspace.

To practice multinomial logistic regression, we’ll use a mock dataset called ‘SchoolChoice.csv’ (available on the website), which includes graduate program selection by accepted applicants (Yale, Duke or Bren) along with their undergraduate GPA and undergraduate field of study, identified as a categorical variable as either Eng (Engineering), Sci (Sciences), or LibArts (Liberal Arts). We want to explore relationships between an applicant’s undergraduate GPA and field of study (explanatory variables) and the likelihood that they will choose to attend Yale, Duke or Bren. \*\**Disclaimer: this data was completely fabricated without any intentional trends produced…I think all three of these schools are super awesome.*

The mock dataset looks like this (abbreviated; ntotal = 51):

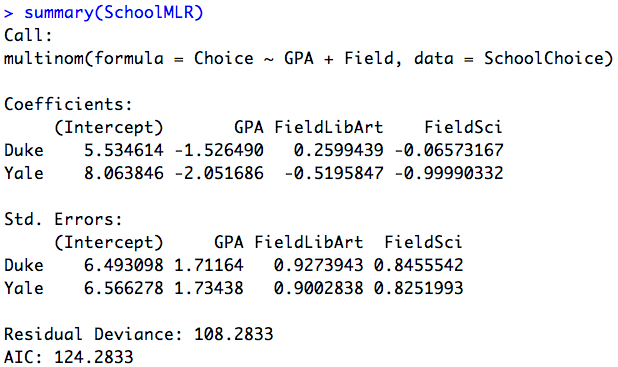


Two important things to notice: (1) the column headers (variable names) are GPA, Field, and Choice, so those are what we’ll input into our regression model, and (2) for the categorical variables (Field and Choice), R will choose the FIRST one that appears as the reference variable (so, “Eng” is the reference variable for the Field variable, and “Bren” is the reference variable for the Choice variable). You can also check the levels within a variable using the levels() function. You’ll see later why you might (in some cases) want to change that around, which is easy to do in R using the relevel() function.

So once we have our data loaded into R we can go ahead and perform multinomial logistic regression using the multinom() function, just like we would for any other regression:



With the following outcome:



Which looks pretty similar to other regression outputs we’ve found, but what do these log-odds model parameters actually tell us, and how can we think/write/talk about them in a way that makes a little more sense? Now it’s time for…

**Interpreting the Dang Coefficients**

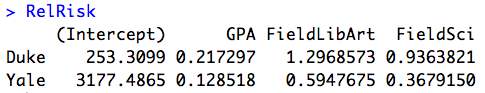
We know that, in general, when we perform logistic regression our coefficients tell us something about the effect of a change in the explanatory variable on the *log odds* of a certain outcome. To get it into a more meaningful format, we can exponentiate the coefficients to find the *odds*, or – in the case of multinomial logistic regression – the *relative risk*. In multinomial logistic regression it is the *relative risk* because it is the odds of one outcome occurring with respect to the odds of the reference outcome occurring.

You’re fine.

OK, so if we *exponentiate* all of the coefficients, using code like this:



which returns the following coefficients for the relative risk:



How do we interpret these ‘relative risk’ coefficients? The main thing to remember is that it is *how a 1 unite change in the explanatory variable changes the odds of a certain outcome occurring versus (or with respect to) the reference outcome.*

Notice that ‘Bren’ and ‘Eng’ don’t show up in the output – that’s because those are the reference variables. We’ll be interpreting the coefficients with respect to those levels.

Here are several examples:

*What does the GPA coefficient of 0.217 for ‘Duke’ outcome mean?* As GPA increases by 1 (assuming all other variables are held constant), the odds of a student choosing Duke versus choosing Bren (the reference outcome) decreases by 78.3% (remember, odds < 1 indicate a decrease in likelihood).

*What does the FieldLibArt coefficient of 1.30 for the ‘Duke’ outcome mean?* As a student’s undergraduate degree changes from Engineering (the reference level) to Liberal Arts, the odds that they will choose Duke versus Bren increases by 0.30 (or, the likelihood of a liberal arts student selecting Duke versus Bren is 30% greater than the odds of an engineering student selecting Duke versus Bren).

*What does the FieldSci coefficient of 0.368 for the ‘Yale’ outcome mean?* As a student’s undergraduate degree changes from Engineering to Science, the relative risk that they will choose Yale versus Bren decreases by 0.632 (i.e., they are less likely to choose Yale v. Bren if they are a Science versus an Engineering major).

Now you try interpreting the FieldSci/Duke and GPA/Yale coefficients using the same approach!

\*\*Note: you can also find the standard error of the relative risk coefficients:

RelRiskSE <- exp(summary(SchoolMLR)$standard.errors)

*and* you could find the *p*-values for each of the coefficients by calculating the z-scores, then the p-values:





The corresponding p-values are stored in ‘p1’ variable.

So, if that’s how I can interpret the coefficients, how can I effectively communicate the results to an audience?

**Visualizing and Communicating Results of Multinomial Logistic Regression**

Saying to yourself (or your best friends, duh) what each of the coefficients means is one thing, but figuring out effective ways to express the results of your regression to an audience is another. Here, we’ll visualize the *probabilities* of a student selecting Bren, Duke or Yale for a range of GPAs and across the different undergraduate fields, and present the results of the regression in a table.

To visualize the *predicted probabilities* of a student selecting Bren, Duke or Yale based on their undergraduate field and GPA, we’ll create a data frame containing series of GPA values and fields, then feed those into our model to get the probability predictions.

Step 1: Create a data frame containing the data you’ll give your model to make predictions with.



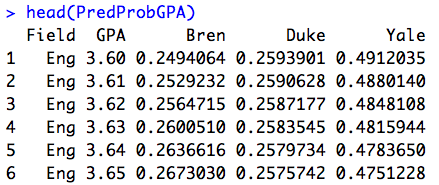
*This code creates a new data frame called ‘GPAFrame,’ which includes a Field column with “Eng”, “LibArt” and “Sci” each repeated 41 times, and a connected GPA column which contains values from 3.60 to 4.00 by increments of 0.01 (3.60, 3.61, 3.62…a total of 41 values to get to 4.00) repeated 3 times. Thus, in each column we will have a length of 41\*3 = 123.*

Make sure that you look at the data frame you’ve just created so that you know what’s in it, and to ensure that it worked!

Step 2: Use the data in the data frame you just created (here, GPAFrame) to find model predictions for probabilities over the range of GPAs for each undergraduate field. Recall that in this example, the multinomial logistic regression model is called ‘SchoolMLR’:



The above code will find the probabilities of each choice for every combination of undergraduate GPA and field included in the ‘GPAFrame’ data frame (and binds it together with the actual data it’s using to make those predictions to make a nice little table). The first 6 lines of the output look like this:



So, based on the ‘SchoolMLR’ model, the probability of an applicant with a GPA of 3.63 and an Engineering background choosing Bren is 0.26, or 26%, versus a 25.8% probability of choosing Duke and a 48.2% probability of choosing Yale.

But, we definitely don’t just want to make a huge table with all of the possible outcomes, so let’s make a graph about it (make sure ggplot2 packages is loaded into RStudio).

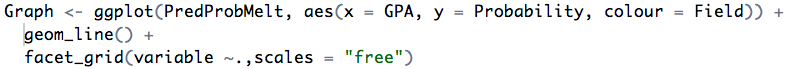
Step 3. Melt the data, then make a graph.

We’ll first “melt” our data frame (with the predicted data) to get it into a format that ggplot() will be a little happier about (it’s kind of like stacking on steroids):



Make sure you check out what that does – notice that now all of the probabilities are in a single column (stacked), and the outcome variable (Bren, Duke, Yale) is now in a column with heading ‘variable.’

Once it’s stacked, make a graph! Here’s some example code using ggplot():



The first two lines should look somewhat familiar. The third line just separates the information into three distinct graphs by the different levels in the ‘variable’ column (so we’ll have a different graph of probabilities for Bren, Duke, and Yale), and doesn’t specify a range of y-values for the probability (scales = “free”).

Run that graphics code and you should end up with something like this:



**Figure 1.** Effect of GPA and Undergraduate Field (Engineering (red); Liberal Arts (green); or Science (blue)) on the predicted probabilities (left y-axis) of an accepted applicant choosing to attend a graduate program at Bren, Duke, or Yale (right y-axis), as determined by multinomial logistic regression (n = 51).

That’s where I’ll leave you with the graphing/visualization, which you can now go on to edit and add layers to and make totally stunning. Now let’s work on a table with the actual results of our logistic regression:

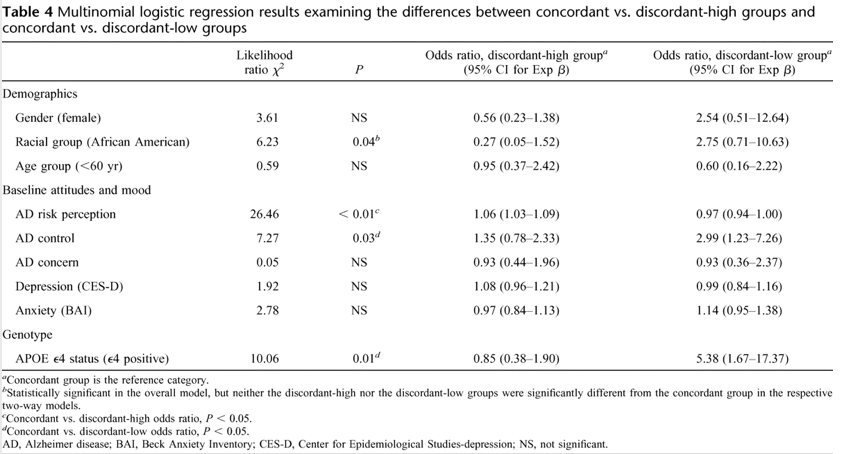
**Table 1.** Multinomial logistic regression coefficients (converted to relative risk versus the reference variable) examining effects of undergraduate field of study and GPA on an applicant’s odds of choosing Duke or Yale versus the reference school, Bren. No coefficients differed significantly from 0 (p > 0.05 for all; n = 51).



Other examples (you might need to check out the mlogit() function, or others that are more advanced, if you need to see the overall model statistics, find CIs, etc.):

*From:*





*From:*

**

