**Section 3.3.2 – Contingency tables**

The multinomial regression model provides a convenient way to perform the same test for independence as earlier in this chapter. We can treat the row variable X as a categorical variable by constructing I – 1 indicator variables. Using Y as the response variable with category probabilities of π1, …, πJ, we have the model

log(πj/π1) = βj0 + βj2x2 + … + βjIxI for j = 2, …, J

where x2, …, xI are used as indicator variables for X (subscript matches level of X). This is a model under dependence.

A model under independence between X and Y is simply

log(πj/π1) = βj0 for j = 2, …, J

Notice that each category of Y can have a different πj, but they do not change as a function of X.

A test for independence involves the hypotheses of

H0: βj2 = = βjI = 0 for j = 2, …, J

Ha: Not all equal for some j

Equivalently, we can state these hypotheses in terms of models:

H0: log(πj/π1) = βj0 for j = 2, …, J

Ha: log(πj/π1) = βj0 + βj2x2 + … + βjIxI for j = 2, …, J

We can use a LRT to test these hypotheses.

Example: Fiber enriched crackers (Fiber.R, Fiber.csv)

Using bloating severity as the response variable and fiber source as the explanatory variable, a multinomial regression model is



where bran, gum, and both in the model represent corresponding indicator variables and the j subscript represents categories low, medium, and high. I could have represented the explanatory variable as fiberbran, fibergum, and fiberboth to match R, but I dropped fiber from the names because there is only one explanatory variable. We can estimate this model using multinom():

> library(package = nnet)

> mod.fit.nom <- multinom(formula = bloat ~ fiber, weights = count, data = diet)

# weights: 20 (12 variable)

initial value 66.542129

iter 10 value 54.519963

iter 20 value 54.197000

final value 54.195737

converged

> summary(mod.fit.nom)

Call: multinom(formula = bloat ~ fiber, data = diet, weights = count)

Coefficients:

 (Intercept) fiberbran fibergum fiberboth

low -0.4057626 -0.1538545 0.4055575 1.322135

medium -1.0980713 -0.8481379 1.5032639 1.503764

high -12.4401085 -4.1103893 13.3561038 12.440403

Std. Errors:

 (Intercept) fiberbran fibergum fiberboth

low 0.6455526 0.8997698 1.190217 1.056797

medium 0.8163281 1.3451836 1.224593 1.224649

high 205.2385583 1497.8087307 205.240263 205.240994

Residual Deviance: 108.3915

AIC: 132.3915

The weights = count argument in multinom() is used because each row of diet represents contingency table counts rather than individual observations.

To perform a LRT for independence, we use the Anova() function from the car package:

> library(package = car)

> Anova(mod.fit.nom)

# weights: 8 (3 variable)

initial value 66.542129

final value 63.635876

converged

Analysis of Deviance Table (Type II tests)

Response: bloat

 LR Chisq Df Pr(>Chisq)

fiber 18.9 9 0.026 \*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Note that we could have also used the anova() function in the appropriate manner.

We have -2log(Λ) = 18.9 and p-value of 0.026. These values match what was found earlier using the assocstats() function!

Comments:

* To examine the potential dependence further, we can examine odds ratios in a similar manner to what we did in the wheat example. Please see my book for further details.
* The 0 cell counts are causing the large standard errors for high bloating severity. In fact, a more stringent convergence criteria (use a different value for reltol – see help for the function), will lead to changes in the regression parameter estimates and standard errors. Therefore, we have non-convergence! Fortunately, the only part of the model affected by the non-convergence corresponds to the high bloating severity. Also, the LRT is not affected. Please see my book for a further discussion and an ad-hoc solution to the problem.

When additional categorical explanatory variables are available, we can examine the data in higher dimensional contingency tables through a multinomial regression model. For example, a model for three categorical variables X, Y, and Z can be written as



In this setting, we can examine if X is independent of Y and/or Z is independent of Y in a very similar manner as before. We can also examine if there is a need for an interaction between X and Z.