**Section 3.4.2 – Contingency tables**

A proportional odds regression model cannot be used to perform the exact same type of test for independence that we saw earlier in this chapter. As mentioned already in this section, the alternative hypothesis with this model specifies one form of the dependence by taking into account the ordinal response. This can be advantageous because a smaller alternative hypothesis leads to a more powerful test.

Suppose we have a categorical variable X that is represented in our model by I – 1 indicator variables. Also, suppose we have an ordinal categorical response Y with category probabilities of π1, …, πJ. A model under independence between X and Y is simply



for j = 1, …, J – 1. A model allowing for dependence is



where x2, …, xI are used as indicator variables for X (subscript matches level of X). A test of independence involves the hypotheses:

H0:  for j=1,…,J–1

Ha:  for j=1,…,J–1

Notice the alternative hypothesis is not as general as what we had when using a multinomial regression model:

H0: log(πj/π1) = βj0 for j = 2, …, J

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

which did not say what type of dependence existed.

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

We previously found that there was *marginal* evidence of dependence between bloating severity and fiber source. Because bloating severity is measured in an ordinal manner (none < low < medium < high), a proportional odds model allows us to account for it and perhaps reach stronger conclusions about the problem of interest.

The alternative hypothesis model here is



where j corresponds to levels 1 (none), 2 (low), 3 (medium), and 4 (high) of bloating severity.

Below is how I estimate the model:

> library(package = MASS)

> levels(diet$bloat)

[1] "none" "low" "medium" "high"

> mod.fit.ord <- polr(formula = bloat ~ fiber, weights = count, data = diet2, method = "logistic")

> summary(mod.fit.ord)

Re-fitting to get Hessian

Call: polr(formula = bloat ~ fiber, data = diet2, weights = count, method = "logistic")

Coefficients:

 Value Std. Error t value

fiberbran -0.3859 0.7813 -0.494

fibergum 2.4426 0.8433 2.896

fiberboth 1.4235 0.7687 1.852

Intercepts:

 Value Std. Error t value

none|low 0.0218 0.5522 0.0395

low|medium 1.6573 0.6138 2.7002

medium|high 3.0113 0.7249 4.1539

Residual Deviance: 112.2242

AIC: 124.2242

The estimated model is



where . A LRT for the fiber source variable gives:

> library(package = car)

> Anova(mod.fit.ord)

Analysis of Deviance Table (Type II tests)

Response: bloat

 LR Chisq Df Pr(>Chisq)

fiber 15.048 3 0.001776 \*\*

Because  = 15.048 is large relative to a distribution (p-value = 0.0018), there is strong evidence that a dependence exists in the form of a trend among the log-odds for the cumulative probabilities. Remember again that the LRT for independence with a more general alternative hypothesis only concluded marginal evidence of dependence and did not specify what type of dependence.

This dependence can be further investigated through examining the odds ratios. Please see the book for a discussion. In the end, there is sufficient evidence to indicate a trend among bloating severity for the gum fiber source.