Discrete Response Model
Lecture 4

datascience@berkeley

Contingency Table

Constructing a Contingency Table to Test Independence

- The multinomial regression model provides a convenient way to perform the same test for independence as in Section 3.2 (of the text).
- We can treat the row variable X as a qualitative variable by constructing **I – 1 indicator variables.**
- Using Y as the response variable with category probabilities of $\pi_1, \ldots, \pi_J$, we have the model

$$\log(\pi_j/\pi_1) = \beta_{j0} + \beta_{j2}x_2 + \ldots + \beta_{jI}x_I \text{ for } j = 2, \ldots, J$$

where $x_2, \ldots, x_I$ are used as indicator variables for X (subscript matches level of X). This is a model under dependence.

Test for Independence

A model under independence between X and Y is simply

$$\log(\pi_j/\pi_1) = \beta_{j0} \text{ for } j = 2, \ldots, J$$

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

A test for independence involves the hypotheses of

H_0: $\beta_{j2} = \cdots = \beta_{jI} = 0$ for $j = 2, \ldots, J$
H_a: Not all equal for some j

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

H_0: $\log(\pi_j/\pi_1) = \beta_{j0}$ for $j = 2, \ldots, J$
H_a: $\log(\pi_j/\pi_1) = \beta_{j0} + \beta_{j2}x_2 + \ldots + \beta_{jI}x_I$ for $j = 2, \ldots, J$

Example

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

$$\log(\pi_j / \pi_{None}) = \beta_{j0} + \beta_{j1}\text{bran} + \beta_{j2}\text{gum} + \beta_{j3}\text{both}$$

where bran, gum, and both in the model represent corresponding indicator variables and the **j** subscript represents categories low, medium, and high. We can estimate this model using **multinom()**:

```
library(package = nnet)
mod.fit.nom<-multinom(formula = bloat ~ fiber, weights =
  count, data = diet2)

# weights:  20 (12 variable)
initial  value 66.542129
iter  10 value 54.519963
iter  20 value 54.197000
final  value 54.195737
converged
```

Example

```
> summary(mod.fit.nom)
Call: multinom(formula = bloat ~ fiber, data = diet2, 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
```

Example

The weights = count argument in **multinom()** is used because each row of diet2 represents contingency table counts rather than individual observations.

To perform a LRT for independence, we can 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
```

Berkeley SCHOOL OF INFORMATION