A Brief Orientation, or Where Log-Linear Models Fit in to the big picture.

Michael J. Rosenfeld © 2002

A) Intro:

Let's say we have a hypothetical dataset with 3 social variables of interest (Var1, Var2, Var3), and each variable has 5 categories. There are 5*5*5=125 possible combinations of the 5 categories, and let's say our dataset has 1000 persons, or cases in it.

A simple log-linear model might look like this:

1) Ln(W)= Constant+ Var1+Var2+Var3+Error

Where *W* is the predicted counts of the model, and *Ln* means 'natural logarithm' or the logarithm with base of *e*, where *e* is just a number that has some nice properties (e≈2.7). The *Error*, or residual, is simply the difference between what the model predicts and the actual count of persons, *W*, for each combination of *Var1, Var2, and Var3*. If the *Errors* are big the model fits poorly, and if the *Errors* are small the model fits well (we'll talk a good deal about how to measure this exactly)

The first thing to notice is that we have a log on the left side of the equation, and a linear combination of things on the right side of the equation, and that's where the terminology 'Log-Linear' models comes from. The STATA function for Log-Linear models is **poisson**, and we'll talk later in the course about the Poisson distribution and what it means.

The second thing to notice is that the left hand side, the dependent variable, is ALWAYS the count of events, or in this case the count of persons. All of the interesting variables are on the right side of the equation. This is different from the kind of regressions you may be used to, that would put one of the variables on the left hand side of the equation, and treat it as the dependent variable. See the brief discussion, below, comparing Log-Linear models with Multinomial Logistic Regression.

Although the the equation (1) seems very linear and additive, there's a multiplicative model lurking under the surface, and the multiplicative model is actually actually very important to think about, and easier to work with in some ways.

Take equation (1), and exponentiate both sides.

Since $e^{\ln(W)} = W$, equation (1) can be rewritten as

$$(2)\quad W = e^{(\text{Constant}+\text{Var1}+\text{Var2}+\text{Var3}+\text{Error})} = e^{Constant} e^{Var1} e^{Var2} e^{Var3} e^{Error}$$

Here you see that W can be expressed as a product of coefficients, each exponentiated. We will switch back and forth between equation (1) and equation (2).

B) Log Linear Models vs. Multinomial Logistic Models:

There is substantial overlap between Log Linear Models and Multinomial Logistic Models. For the very simplest possible kind of models (such as a dataset with two variables each of which has two categories), the two approaches are equally easy and exactly equivalent (I'll demonstrate this next week). As soon as the model becomes a bit complicated, however, the two approaches diverge.

The Multinomial Logistic version of equation (1) would look something like this:

3) Logit(Var1)= Constant+Var2+Var3+Error

Since *Var1* has 5 categories, equation (3) would produce 4 resulting equations, with 4 categories of *Var1* compated to the one comparison category of *Var1*.

The advantage of logistic regression is:

* it's much easier to add more variables, including continuous variables

The advantage of log linear models is:

* log linear models provide more control over the interaction of the variables. For instance, every term in equation (3) is given in terms of its effects on Var 1. But what if you want to examine the interaction between Var2 and Var3 without regard for Var1?

In practice, the tradeoff between logistic regression and log linear models is a tradeoff between flexibility and control. There has been some literature that has argued, from a theoretical perspective, that individual level variables could be added to log linear models (DiPrete in ASR, 1990), or that multiple multinomial regressions could be combined in ways that would replicate log linear models. In practice, however, the middle ground between log linear models and multinomial logistic models has proved to be just too cumbersome to deal with. So as a practical matter one must make a trade-off and chose one or the other approach.