Log-linear Models

The classical log-linear models are tools for analyzing relationships among two or more categorical variables. They are based on multi-dimensional joint frequency tables. In the sample, each cell in such a table contains the number of cases with a particular combination of values of the variables. In the population, each cell in our multi-dimensional table contains a probability -- the probability of selecting a case with that combination of values. This table is exactly the joint probability distribution of the variables in the analysis.

If you multiply the probabilities by the size of the sample, you get expected frequencies. A "log-linear model" is a statistical model for the natural logarithm (ln) of the expected frequency. It looks like a multiple regression model with effect coding, in which the interaction terms correspond to associations among variables. If the variables are unrelated (conditionally upon the values of other variables in the model), the interaction terms are missing. The terms corresponding to main effects represent departures from equal marginal frequencies.

The model with terms corresponding to all possible main effects and interactions is called the *saturated model*. The saturated model always fits the data as well as any model possibly can. That is, it is equivalent to the very simple and unrestricted model in which all the cell probabilities in our multi-way table are estimated by the corresponding sample proportions. The -2 Log Likelihood quantity is the same for the saturated model and the simple cell proportions model.

You can't test the saturated model, but you can estimate a non-saturated model with software, and test the difference between that one and the saturated model. Such tests are often called "goodness of fit" tests, because they tell you whether the model in question is significantly worse than the saturated (perfect) model. There are several good ways to conduct goodness of fit tests, but we will confine ourselves to *likelihood ratio tests*. We really are testing the difference between a reduced model and a full model. In this case, the full model is the saturated model.

If the reduced model is true (that's the null hypothesis), the likelihood ratio statistic (minus two times the natural log of the likelihood function evaluated at the Maximum Likelihood Estimate) has a distribution that approaches a chi-square distribution as the sample size increases. The degrees of freedom are the number of terms present in the saturated model but missing from the reduced model.

Often, the tests that are really interesting can be expressed as the difference between the saturated model and a carefully chosen reduced model. You are testing (simultaneously) all the associations among variables that are *absent* in the reduced model. Sometimes, especially when four or more variables are involved, what you are interested in may correspond to the difference between a reduced model and an even more reduced one. In this case, the *difference* between -2 log(likelihood) of the two models will have a chi-square distribution with df equal to the difference in degrees of freedom -- **provided** that the terms in the more reduced model are a subset of the terms in the less reduced one. That is, the models have to be *nested* in order for the large-sample likelihood ratio tests to be valid. If you do this, a good practice is to be sure that your less restricted model (that's your new "full" model that's not the saturated model) fits the data fairly well. At the very least, it should not fit significantly worse than the saturated model.

In order to avoid ambiguities and tricky problems interpreting results, we will confine ourselves to *hierarchical* log-linear models. Hierarchical means that if an effect is present in a model, then all the lower-order effects that make it up must also be in the model. For example, if a model contains a three-way A by B by C association, then it must also contain A, B, C, A by B, A by C and B by C.

For hierarchical models, there is a very convenient bracket notation for expressing association and lack of association among variables, especially if the variables can all be represented with single symbols like letters or numbers. Just enclose sets of variables that are associated within the same set of brackets. A variable that does not appear at all has equal marginal frequencies. For example,

- For three variables numbered one through three, the model [1] [2] [3] allows each variable to have unequal marginal frequencies, but it contains no relationships among variables. It is a model of complete independence. If the test for goodness of fit is significant, the conclusion is that there is *some* relationship among variables. This is not a bad place to start in any analysis.

- The model [1,2] [1,3] [2,3] allows for lack of independence in each of the three two-way marginal tables. Because the model is hierarchical, the three single-variable terms are implicitly present. The only term that is missing from this model is the three-factor relationship 1*2*3, so the test for goodness of fit is a test for whether, equivalently,

- * The relationship between 1 and 2 is the same for all values of variable 3
- * The relationship between 1 and 3 is the same for all values of variable 2
- * The relationship between 2 and 3 is the same for all values of variable 1

That is, it is analogous to a two-factor interaction in the normal linear model.

° The model [1,2] [1,3] says that the only thing going on is (possibly) a relationship between Variables 1 and 2, and a relationship between 1 and 3. Any apparent relationship between 2 and 3 arises from the fact that they are both related to 1. This is a model of conditional independence. That is, conditionally on (controlling for) the value of Variable 1, Variables 2 and 3 are unrelated.

You can get the test statistic for this model another way. Produce separate two-way tables of Variable 2 by Variable 3 -- one for each value of Variable 1. This is the subdivision approach to controlling for Variable 1. Add the chisquare values for testing independence in the sub-tables. Under the null hypothesis that Variables 2 and 3 are independent for each fixed value of Variable 1, the sum of chisquares has a chisquare distribution, with degrees of freedom equal to the sum of degrees of freedom from the sub-tests. If you add likelihood ratio chi-squares, you get the standard test of conditional independence for loglinear models. But adding Pearson chi-squares is valid too.

By the way, suppose the test just described is significant. To see where the effect comes from, try looking one-at-a-time at the chi-square statistics you just added up. It would be best to apply a Bonferroni correction.

This is a good way to slice up a test for conditional independence, but it is not the only good way. The model [1,2] [1,3] lacks two terms that are present in the saturated model. They are 2*3 and 1*2*3. If we added just the first one to the model [1,2] [1,3], we would get [1,2] [1,3] [2,3]. This says that Variables 2 and 3 may be related, but if so they are related *in the same way* for all values of Variable 1. To test for this limited form of dependence of 2 and 3, use [1,2] [1,3] [2,3] as the full model and [1,2] [1,3] as the reduced model. Again, the test statistic is the difference in -2 times the log likelihood for the two models, distributed as chisquare with df equal to the difference in degrees of freedom. The numbers are quite easy to locate on the printout from SAS proc catmod. By the way, this all works out because the two models are *nested*. The terms in [1,2] [1,3] are a subset of the terms in [1,2] [1,3] [2,3].

Next, you can test for the other piece of departure from conditional independence, by testing the goodness of fit of [1,2] [1,3] [2,3]. This tests [1,2] [1,3] [2,3] against the saturated model, equivalently testing for 1*2*3. Again the models are nested.

° Suppose you want to test association between two *sets* of variables. For example, suppose Variables 1 through 4 represent employment history, and Variables 5 through 8 represent employment history of the parent (of the same sex). The model to test against the saturated model is [1234] [5678]. This model says that the employment history variables may be related any way at all, and parental employment history variables may be related any way at all, but the employment history variables are completely independent of the parental employment history variables.

In the theory of log-linear models, there is no distinction between independent variables and dependent variables. But there are some situations where you want to make the distinction. Say, two or more of the variables are randomly assigned. Logistic regression with more than two categories in the dependent variable (using generalized logits) covers this situation, but it turns out that as long as both the full and the reduced model contain all possible associations among *independent* variables (even if they are set up to be unrelated), classical loglinear models yield exactly the same test statistics. This leads to the following simple rule. **If you want to make a distinction between independent and dependent variables, just include all possible associations among independent variables in the model.**

Examples will be given in lecture.

If you want a more detailed exposition of log-linear models, a good and somewhat readable book is Stephen Feinberg's *Analysis of cross-classified categorical data.* It's an excellent reference for the subdivision approach (adding up the chisquare values for sub-tables), too. A more comprehensive treatment of log-linear models at a somewhat higher level may be found in Bishop, Feinberg and Holland's *Discrete multivariate analysis* (same Feinberg). This book is affectionately known as the "Green Monster" because of the colour of the cover and the number of pages. It's a good source if you want to cite something authoritative without necessarily reading it.