Loglinear Models

Angela Jeansonne

This page last updated

Brief History

Until the late 1960's, contingency tables - two-way tables formed by cross classifying categorical variables - were typically analyzed by calculating chi-square values testing the hypothesis of independence. When tables consisted of more than two variables, researchers would compute the chi-squares for two-way tables and then again for multiple sub-tables formed from them in order to determine if associations and/or interactions were taking place among the variables. In the 1970's the analysis of cross-classified data changed quite dramatically with the publication of a series of papers on loglinear models by L.A. Goodman. Many other books appeared around that time building on Goodman's work (Bishop, Finberg & Holland 1975; Haberman 1975). Now researchers were introduced to a wide variety of models that could be fitted to cross-classified data. Thus, the introduction of the loglinear model provided them with a formal and rigorous method for selecting a model or models for describing associations between variables.

Overview

When to use loglinear models:

The loglinear model is one of the specialized cases of generalized linear models for Poisson-distributed data. Loglinear analysis is an extension of the two-way contingency table where the conditional relationship between two or more discrete, categorical variables is analyzed by taking the natural logarithm of the cell frequencies within a contingency table. Although loglinear models can be used to analyze the relationship between two categorical variables (two-way contingency tables), they are more commonly used to evaluate multiway contingency tables that involve three or more variables. The variables investigated by log linear models are all treated as "response variables". In other words, no distinction is made between independent and dependent variables. Therefore, loglinear models only demonstrate association between variables. If one or more variables are treated as explicitly dependent and others as independent, then logit or logistic regression should be used instead. Also, if the variables being investigated are continuous and cannot be broken down into discrete categories, logit or logistic regression would again be the appropriate analysis. For a complete discussion on logit and logistic regression consult Agresti (1996) or Tabachnick and Fidell (1996).

Example of data appropriate for loglinear models:

Suppose we are interested in the relationship between sex, heart disease and body weight. We could take a sample of 200 subjects and determine the sex, approximate body weight, and who does and does not have heart disease. The continuous variable, body weight, is broken down into two discrete categories: not over weight, and over weight. The contingency table containing the data may look like this:

		Heart Disease		Total
Body Weight	**Sex**	**Yes**	**No**	
Not over weight	**Male**	15	5	20
	Female	40	60	100
Total		55	65	120
Over weight	**Male**	20	10	30
	Female	10	40	50
Total		30	50	80

In this example, if we had designated heart disease as the dependent variable and sex and body weight as the independent variables, then logit or logistic regression would have been the appropriate analysis.

Basic Strategy and Key Concepts:

The basic strategy in loglinear modeling involves fitting models to the observed frequencies in the cross-tabulation of categoric variables. The models can then be represented by a set of expected frequencies that may or may not resemble the observed frequencies. Models will vary in terms of the marginals they fit, and can be described in terms of the constraints they place on the associations or interactions that are present in the data. The pattern of association among variables can be described by a set of odds and by one or more odds ratios derived from them. Once expected frequencies are obtained, we then compare models that are hierarchical to one another and choose a preferred model, which is the most parsimonious model that fits the data. It's important to note that a model is not chosen if it bears no resemblance to the observed data. The choice of a preferred model is typically based on a formal comparison of goodness-of-fit statistics associated with models that are related hierarchically (models containing higher order terms also implicitly include all lower order terms). Ultimately, the preferred model should distinguish between the pattern of the variables in the data and sampling variability, thus providing a defensible interpretation.

The Loglinear Model

The following model refers to the traditional chi-square test where two variables, each with two levels (2 x 2 table), are evaluated to see if an association exists between the variables.

$$Ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

$Ln(F_{ij})$ = is the log of the expected cell frequency of the cases for cell ij in the contingency table.
μ = is the overall mean of the natural log of the expected frequencies
λ = terms each represent "effects" which the variables have on the cell frequencies
A and B = the variables
i and j = refer to the categories within the variables

Therefore:
λ_i^A = the main effect for variable A
λ_j^B = the main effect for variable B
λ_{ij}^{AB} = the interaction effect for variables A and B

The above model is considered a Saturated Model because it includes all possible one-way and two-way effects. Given that the saturated model has the same amount of cells in the contingency table as it does effects, the expected cell frequencies will always exactly match the observed frequencies, with no degrees of freedom remaining (Knoke and Burke, 1980). For example, in a 2 x 2 table there are four cells and in a saturated model involving two variables there are four effects, μ, λ_i^A, λ_j^B, λ_{ij}^{AB}, therefore the expected cell frequencies will exactly match the observed frequencies. Thus, in order to find a more parsimonious model that will isolate the effects best demonstrating the data patterns, a non-saturated model must be sought. This can be achieved by setting some of the effect parameters to zero. For instance, if we set the effects parameter λ_{ij}^{AB} to zero (i.e. we assume that variable A has no effect on variable B, or vice versa) we are left with the unsaturated model.

$$Ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B$$

This particular unsaturated model is titled the Independence Model because it lacks an interaction effect parameter between A and B. Implicitly, this model holds that the variables are unassociated. Note that the independence model is analogous to the chi-square analysis, testing the hypothesis of independence.

Hierarchical Approach to Loglinear Modeling

The following equation represents a 2 x 2 x 2 multiway contingency table with three variables, each with two levels – exactly like the table illustrated on page 1 of this article. Here, this equation is being used to illustrate the hierarchical approach to loglinear modeling.

$$Ln(F_{ij}) = \mu + \lambda_i^{A} + \lambda_j^{B} + \lambda_k^{C} + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

A hierarchy of models exists whenever a complex multivariate relationship present in the data necessitates inclusion of less complex interrelationships (Knoke and Burke, 1980). For example, in the above equation if a three-way interaction is present (ABC), the equation for the model must also include all two-way effects (AB, AC, BC) as well as the single variable effects (A, B, C) and the grand mean (μ). In other words, less complex models are nested within the higher-order model (ABC). Note the shorter notation used here to describe models. Each set of letters within the braces indicates a highest order effect parameter included in the model and by virtue of the hierarchical requirement, the set of letters within braces also reveals all lower order relationships which are necessarily present (Knoke and Burke, 1980).
SPSS uses this model to generate the most parsimonious model; however, some programs use a non-hierarchical approach to loglinear modeling. Reverting back to the previous notation, a non-hierarchical model would look like the following: $Ln(F_{ij}) = \mu + \lambda_i^{A} + \lambda_{ij}^{AB}$. Notice that the main effect term λ_i^{B} is not included in the model therefore violating the hierarchical requirement. The use of non-hierarchical modeling is not recommended, because it provides no statistical procedure for choosing from among potential models.

Choosing a model to Investigate

Typically, either theory or previous empirical findings should guide this process. However, if an a priori hypothesis does not exist, there are two approaches that one could take:

1. Start with the saturated model and begin to delete higher order interaction terms until the fit of the model to the data becomes unacceptable based on the probability standards adopted by the investigator.
2. Start with the simplest model (independence model) and add more complex interaction terms until an acceptable fit is obtained which cannot be significantly improved by adding further terms.

Fitting Loglinear Models

Once a model has been chosen for investigation the expected frequencies need to be tabulated. For two variable models, the following formula can be used to compute the direct estimates for non-saturated models.

(column total) * (row total)/grand total

For larger tables, an iterative proportional fitting algorithm (Deming-Stephan algorithm) is used to generate expected frequencies. This procedure uses marginal tables fitted by the model to insure that the expected frequencies sum across the other variables to equal the corresponding observed marginal tables (Knoke and Burke, 1980).

For Example: In the following contingency tables, the observed marginal table totals (each column and row) are equal to the expected marginal table totals, even though the actual expected frequencies are different from the observed frequencies.

Observed Frequencies

		Membership		
		One or More	None	Total
Vote Turnout	Voted	689	298	987
	Not Voted	232	254	486
Total		921	552	1473

Expected Frequencies

		Membership		
		One or More	None	Total
Vote Turnout	Voted	617.13	369.87	987
	Not Voted	303.87	182.13	486
Total		921	552	1473

(Note: The above contingency tables were taken from Knoke and Burke, 1980 and represent data collected on voluntary membership association and voter turnout in the 1972 and 1976 Presidential elections in the United States.)

The iterative proportional fitting process generates maximum likelihood estimates of the expected cell frequencies for a hierarchical model. In short, preliminary estimates of the expected cell frequencies are successfully adjusted to fit each of the marginal sub-tables specified in the model. For example, in the model AB, BC, ABC, the initial estimates are adjusted to fit AB then BC and finally to equal the ABC observed frequencies. The previous adjustments become distorted with each new fit, so the process starts over again with the most recent cell estimate. This process continues until an arbitrarily small difference exists between the current and previous estimates. Consult Christensen (1997) for a numerical explanation of the iterative computation of estimates.

Parameter Estimates

Once estimates of the expected frequencies for the given model are obtained, these numbers are entered into appropriate formulas to produce the effect parameter estimates (λ's) for the variables and their interactions (Knoke and Burke, 1980). The effect parameter estimates are related to odds and odds ratios. Odds are described as the ratio between the frequency of being in one category and the frequency of not being in that category. For example, in the above contingency table for observed frequencies, the odds that a person voted is 987/486 = 2.03. The odds ratio is one conditional odds divided by another for a second variable, such as the odds of having voted for the second variable Membership. Based on the same contingency table, the conditional odds for having voted and belonging to one or more groups is 2.97 (689/232), and the conditional odds

for having voted and not belonging to any groups is 1.17 (289/254). Then the odds ratio for voting for people belonging to more than one group to belonging to none is 2.97/1.17 = 2.54. This is also called the "cross-product ratio" and in a 2x2 table can be computed by dividing the product of the main diagonal cells (689*254) by the product of the off diagonal cells (232*298). An odds ratio above 1 indicates a positive association among variables, while odds ratios smaller than one indicate a negative association. Odds ratios equaling 1 demonstrate that the variables have no association (Knoke and Burke, 1980). Note that odds and odds ratios are highly dependent on a particular model. Thus, the associations illustrated by evaluating the odds ratios of a given model are informative only to the extent that the model fits well.

Testing for Fit

Once the model has been fitted, it is necessary to decide which model provides the best fit. The overall goodness-of-fit of a model is assessed by comparing the expected frequencies to the observed cell frequencies for each model. The Pearson Chi-square statistic or the likelihood ratio (L^2) can be used to test a models fit. However, the (L^2) is more commonly used because it is the statistic that is minimized in maximum likelihood estimation and can be partitioned uniquely for more powerful test of conditional independence in multiway tables (Knoke and Burke, 1980). The formula for the L^2 statistic is as follows:

$$L^2 = 2\Sigma fij \ln(f_{ij}/F_{ij})$$

L^2 follows a chi-square distribution with the degrees of freedom (df) equal to the number of lambda terms set equal to zero. Therefore, the L^2 statistic tests the residual frequency that is not accounted for by the effects in the model (the λ parameters set equal to zero). The larger the L^2 relative to the available degrees of freedom, the more the expected frequencies depart from the actual cell entries. Therefore, the larger L^2 values indicate that the model does not fit the data well and thus, the model should be rejected. Consult Tabachnick and Fidell (1996) for a full explanation on how to compute the L^2 statistic.

It is often found that more than one model provides an adequate fit to the data as indicated by the non-significance of the likelihood ratio. At this point, the likelihood ratio can be used to compare an overall model within a smaller, nested model (i.e. comparing a saturated model with one interaction or main effect dropped to assess the importance of that term). The equation is as follows:

$$L^2_{comparison} = L^2_{model1} - L^2_{model2}$$

Model 1 is the model nested within model 2. The degrees of freedom (df) are calculated by subtracting the df of model 2 from the df of model 1.
If the L^2 comparison statistic is not significant, then the nested model (1) is not significantly worse than the saturated model (2). Therefore, choose the more parsimonious (nested) model.

Following is a table that is often created to aid in the comparison of models. Based on the above equation, if we wanted to compare model 1 with model 11 then we would compute L^2 comparison = 66.78 – 0.00 which yields a L^2 comparison of 66.78. The df would be computed by subtracting 0 from 1 yielding a df of 1. The L^2 comparison figure is significant, therefore we cannot eliminate the interaction effect term VM from the model. Thus, the best fitting model in this case is the saturated model.

Comparisons Among Models								
		Effect Parameters				**Likelihood Ratio**		
Model	Fitted Marginals	η	T_1^V	T_1^M	T_{11}^{VM}	$\mathbf{L^2}$	d.f.	p
1	{VM}	331.66	1.37	0.83	0.80	**0.00**	0	-
11	{V{}M}	335.25	1.43	0.77	1.00*	**66.78**	1	<.001
12	{V}	346.3	1.43	1.00*	1.00*	**160.22**	2	<.001
13	{M}	356.51	1.00*	1.29	1.00*	**240.63**	2	<.001
14	{ }	368.25	1.00*	1.00*	1.00*	**334.07**	3	<.001

* Set to 1.00 by hypothesis (Note: Table is taken from Knoke and Burke, 1980)

Loglinear Residuals

In order to further investigate the quality of fit of a model, one could evaluate the individual cell residuals. Residual frequencies can show why a model fits poorly or can point out the cells that display a lack of fit in a generally good-fitting model (Tabachnick and Fidell, 1996). The process involves standardizing the residuals for each cell by dividing the difference between frequencies observed and frequencies expected by the square root of the frequencies expected ($F_{obs} – F_{exp} / \sqrt{} F_{exp}$). The cells with the largest residuals show where the model is least appropriate. Therefore, if the model is appropriate for the data, the residual frequencies should consist of both negative and positive values of approximately the same magnitude that are distributed evenly across the cells of the table.

Limitations to Loglinear Models

Interpretation

The inclusion of so many variables in loglinear models often makes interpretation very difficult.

Independence

Only a between subjects design may be analyzed. The frequency in each cell is independent of frequencies in all other cells.

Adequate Sample Size

With loglinear models, you need to have at least 5 times the number of cases as cells in your data. For example, if you have a 2x2x3 table, then you need to have 60 cases. If you do not have the required amount of cases, then you need to increase the sample size or eliminate one or more of the variables.

Size of Expected Frequencies

For all two-way associations, the expected cell frequencies should be greater than one, and no more than 20% should be less than five. Upon failing to meet this requirement, the Type I error rate usually does not increase, but the power can be reduced to the point where analysis of the data is worthless. If low expected frequencies are encountered, the following could be done:

1. Accept the reduced power for testing effects associated with low expected frequencies.
2. Collapse categories for variables with more than two levels, meaning you could combine two categories to make one "new" variable. However, if you do this, associations between the variables can be lost, resulting in a complete reduction in power for testing those associations. Therefore, nothing has been gained.
3. Delete variables to reduce the number of cells, but in doing so you must be careful not to delete variables that are associated with any other variables.
4. Add a constant to each cell (.5 is typical). This is not recommended because power will drop, and Type I error rate only improves minimally.
   Note: Some packages such as SPSS will add .5 continuity correction under default.

References

Agresti, A.1996. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc. New York, New York, USA. *

Christensen, R. 1997. *Log-Linear Models and Logistic Regression*. Springer-Verlag Inc. New York, New York, USA.

Everitt, B.S. 1977. *The Analysis of Contingency Tables*. John Wiley & Sons, Inc. New York, New York, USA.

Knoke, D. and P.J. Burke 1980. *Log-Linear Models*. Sage Publications, Inc. Newberry Park, California, USA. *

Read, T.R.C. and N.A.C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag Inc. New York, New York, USA.

Tabachnick, B.G. and L.S. Fidell 1996. *Using Multivariate Statistics*. 3rd Edition. Harper Collins. New York, New York, USA.

* References that were the most informative

References on the Internet

http://asio.jde.aca.mmu.ac.uk/new_gis/analysis/loglin.htm - A brief tutorial on log-linear models.

http://www.statsoftinc.com/textbook/stloglin.html - A tutorial on log linear analysis of frequency tables.

http://www.math.yorku.ca/SCS/Courses/grcat/grc8.html - A brief discussion on log linear models and how to use the statistical software, SAS, for log linear modeling.

http://wizard.ucr.edu/~rhannema/soc203a/loglin.html - A comprehensive article on hierarchical log-linear models.

http://www2.chass.ncsu.edu/garson/pa765/logit.htm - A brief tutorial on log-linear, logit and probit models. This article provides a good glossary of terms that apply to all three analyses.

Articles that use Log Linear Models

Brunkow, P.E., Collins, J.P. 1996. Effects if Individual Variation in Size on Growth and Development of Larval Salamanders. Ecology, 77, 1483-1492.

Cords, M. 1986. Interspecific and Intraspecific Variation in Diet of Two Forest Guenons, Cercopithecus ascanius and C. mitis. Journal of Animal Ecology, 55, 811-827.

Quesada, M., Bollman, K., and Stephenson, A.G. 1995. Leaf Damage Decreases Pollen Production and Hinders Pollen Performance in Cucurbita texana. Ecology, 76, 437-443.

Whittam, T.S., and Siegel-Causey, D. 1981. Species Interactions and Community Structure in Alaskan Seabird Colonies. Ecology, 62, 1515-1524.

Following are statistical packages that perform loglinear analysis: SPSS, SAS, and BMPD