Loglinear Models of Association

American Evaluation Association Conference 2010
Eric Canen
Nanette Nelson
University of Wyoming, Wyoming Survey & Analysis Center

What are loglinear models?

Log Linear models are statistical models that can be used on contingency tables to determine relationships between categorical variables. Most often log linear models are used with higher order contingency tables because they use terminology and concepts that are similar to regression and ANOVA models. These models test relationships between variables in a similarly conceptual manner as do correlational measures of association.

What are the assumptions that I should be aware of when using loglinear models?

- Count data broken down according to the variables used in the models and can be represented in a contingency table format.
- Smallest sample size should be better than five times the number of cells in the contingency table (e.g. contingency table is 2 X 3 X 3 then total sample size should not be less than 90)
- All cells should have at least one case and no more than 20% of the cells should have expected counts of less than five cases.
- Use the right distribution for the models
  - ◊ If total sample size is fixed, then data should be fitted to a multinomial distribution
  - ◊ If total sample size is free to vary, then data should be fitted to a Poisson distribution

What software can I use to analyze loglinear models?

Many standard statistical software packages have the capabilities to analyze loglinear models. Most software has at least two ways of analyzing the data using loglinear analysis. The most general way is the generalized linear models commands, however most stats packages also have specific commands for loglinear models.

SAS
- Proc CatMod procedure
- Proc GenMod procedure

R
- loglin() function
- glm() function

Stata
- poisson command (Poison regression)
- glm command

SPSS/PASW
- GENLOG
- GENLIN

NOTE: This handout will focus on loglinear models from SPSS using

How do I perform loglinear analysis using the software?

(Example from SPSS)
1. Open a dataset
2. Click Analyze → Loglinear → General
3. This will open an analysis dialog box where you can enter the variables you will be analyzing and select the distribution type (multinomial or Poisson)
4. Under options, click to see the expected counts
5. Under design you can add main effects, and interaction effects among the chosen variables.

What is the program trying to do under different model designs?

Under a simple situation where there are only two variables being modeled there are only two real ways of modeling the data. The first is to have a null hypothesis where you assume that both variables are independent of one another. Under that situation the probabilities for the joint distribution between the two variables are assumed to have the following expected probabilities:

$H_0: \pi_{ij} = \pi_{i.}\,\pi_{.j}$

In other words multiplying the marginal probabilities between the two variables will yield the expected probabilities of the joint cells. This corresponds to a loglinear model that only includes the main effects for the two variables and no interaction effects between the two.

If all possible main effects and interactions are in the model, this generates a saturated model, which predicts the cell frequencies perfectly but leaves no degrees of freedom. The saturated model represents the standard by which all goodness-of-fit tests are compared in loglinear models. Putting the two-way interaction effect into the two factor model will yield a saturated model.

When the loglinear model involves three or more variables, then the models can correspond to the following model types:

Complete Independence-This model tests whether all the variables are independent of one another. It only includes the main effects in the model. The null hypothesis (for three variables) is:

$H_0: \pi_{ijk} = \pi_{i..}\,\pi_{.j.}\,\pi_{..k}$

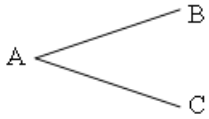
Block Independence-This model tests whether there is a relationship between two variables, but the third variable remains independent. The model includes all of the main effect terms, and only one of the two-way interaction terms. The null hypothesis (for three variables is):

$H_0: \pi_{ijk} = \pi_{i..}\,\pi_{.jk}$

—*Continued on next page*

—continued from previous page

Partial Association Models-These models test a meditational relationship of one variable against two other variables. In diagram form they test the following idea:



These models involve all the main effects and two two-way interaction effects. These interaction effects share a common variable. The null hypotheses for these models involve conditioning the probabilities based on categories of the common variable:

$$H_0: \pi_{ijk} = (\pi_{ij.} \pi_{i.k}) / \pi_i$$

Uniform Association Models-These models test whether the association between any two of the variables is the same at all levels of the third variable. These models have all the main effects and three or more two-way effects.

For these models, the null hypothesis cannot be represented by describing the combination of marginal effects as was done for the other types of models. However if these models fail to adequately fit the data then associations between any two variables are not the same at all levels of the third variable.

How do I understand the output from loglinear models?

This is a quick walkthrough and things to be aware of from the loglinear output using SPSS GENLOG command:

Loglinear models use iterative maximum likelihood to arrive at its estimates. If the model fails to converge then you need to check the assumptions, specifically make sure that you have chosen the correct distribution and that there are not too many small cells where n < 5.

Convergence Information[b,c]

Maximum Number of Iterations	20
Converge Tolerance	.00100
Final Maximum Absolute Difference	.00014[a]
Final Maximum Relative Difference	.00012
Number of Iterations	4

a. The iteration converged because the maximum absolute changes of parameter estimates is less than the specified convergence criterion.

b. Model: Poisson

c. Design: Constant + ordinance + PREPOSTORD + FUD1_dichot

The goodness-of-fit tests compares the residuals of the current model against the saturated model. If the numbers in the Sig column are greater than your chosen alpha level, then you can assume that the model fits as well as the saturated model. Even more importantly you can use the Likelihood Ratio values presented in this table to perform Likelihood Ratio tests to compare different models.

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	192.805	4	.000
Pearson Chi-Square	186.247	4	.000

a. Model: Poisson

b. Design: Constant + ordinance + PREPOSTORD + FUD1_dichot

Cell Counts and Residuals[a,b]

Presence of municipal smokefree ordinance within this county	If county contains municipality with smokefree ordinance, was this respondent …	Q24A_R. (Dichotomized) Think of your four best friends. In the past year, how many of your best…	Observed		Expected		Residual	Standardized Residual	Adjusted Residual	Deviance
			Count	%	Count	%				
Has a smokefree ordinance	Before municipal smokefree ordinance enacted (in ordinance county or matched control county)	0 friends	7712	25.7%	7666.464	25.6%	45.536	.520	.954	.520
		1, 2, 3, or 4 friends	5461	18.2%	5382.142	17.9%	78.858	1.075	1.750	1.072
	After municipal smokefree ordinance enacted (in ordinance county or matched control county)	0 friends	3166	10.6%	2859.746	9.5%	306.254	5.727	7.958	5.629
		1, 2, 3, or 4 friends	1577	5.3%	2007.648	6.7%	-430.648	-9.611	-12.274	-9.990
DOES NOT have a smokefree ordinance	Before municipal smokefree ordinance enacted (in ordinance county or matched control county)	0 friends	4760	15.9%	5168.744	17.2%	-408.744	-5.685	-9.141	-5.763
		1, 2, 3, or 4 friends	3913	13.0%	3628.650	12.1%	284.350	4.720	6.798	4.661
	After municipal smokefree ordinance enacted (in ordinance county or matched control county)	0 friends	1985	6.6%	1928.046	6.4%	56.954	1.297	1.643	1.291
		1, 2, 3, or 4 friends	1421	4.7%	1353.560	4.5%	67.440	1.833	2.187	1.818

a. Model: Poisson

b. Design: Constant + ordinance + PREPOSTORD + FUD1_dichot

This is a key table from the output because it presents the actual observed counts and the expected counts from the model. Specifically, this the main diagnostic table to find out where the model is fitting and where it is lacking. NOTE: The residuals are the difference between the observed counts and the expected counts. The standardized residuals are the z-scores for the residuals and represent the discrepancy in standard deviation units. Deviance statistics for each cell are also presented.

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	7.210	.015	486.540	.000	7.181	7.240
[ordinance = 1.00]	.394	.012	33.486	.000	.371	.417
[ordinance = 2.00]	0[a]	.	.	.	.	.
[PREPOSTORD = 1.00]	.986	.013	75.974	.000	.961	1.012
[PREPOSTORD = 2.00]	0[a]	.	.	.	.	.
[FUD1_dichot = 1.00]	.354	.012	30.162	.000	.331	.377
[FUD1_dichot = 2.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + ordinance + PREPOSTORD + FUD1_dichot

This presents the parameter estimates of the model. I use this primarily to make sure that the all the parameters are necessary and significant. If there are nonsignficant parameters which are not involved in higher order interaction terms, then I try to remove them to get to a simpler model that may describe the data just as well if not better.

Excellent Web Resources:

http://data.princeton.edu/wws509/notes/c5s2.html

http://userwww.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.htm

Contact information:
Eric Canen
Wyoming Survey & Analysis Center
University of Wyoming

ecanen@uwyo.edu
http://wysac.uwyo.edu/