Wiki Assignment: 2009 and 2010 (2010 changes highlighted)

Chapter 7

Summary

Loglinear models analyze the conditional relationships of two or more categorical variables, by examining the relationship between cell frequencies. Thus, unlike previous techniques pertaining predicting group membership, loglinear models focus on the counts for the various cells within the contingency table. They can be used for large contingency tables and are especially useful when there is not a single outcome that needs to be predicted.

Model Design

Loglinear models examine contingency tables and have a different fixed, random and link function from other analyses. The random component (DV) of loglinear models is the cell counts. These are Poisson distributed and this loglinear models use a log link. The systematic component of the models involve the classifications from the contingency tables.

The saturated model is a model with both variables frequencies and cell frequencies estimated in the model. In a model with two variables x and y, for the expected cell frequency μ_{ij}

The equation for the saturated model is:

$$\log(\mu_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy}$$

The first term is the intercept and represents the overall cell frequencies. The second and third terms test if the frequencies in the rows and columns are equal or unequal. The fourth term tests if the proportions in the rows are conditional on proportions in the columns. More restricted models can be tested by comparing model fit.

Model Fit

Loglinear analyses examine all possible main effects and interactions in a similar manner as logistic regression. Through the use of likelihood ratio (LR) tests of model fit, comparing restricted models to the saturated model, we can determine whether a more parsimonious model can adequately explain the data by gradually including and excluding variables, depending upon their effect in predicting cell count. To test model fit, we compare the deviance value to the chi-square significance value with the degrees of freedom equal to the number of degrees of freedom that change between the two models. If the deviance value is smaller than the chi-square significance value, then we can infer that the more saturated model does not do a significantly better job at explaining the data trends. Just like in previous models, we would like the model to not be significantly different from the saturated model but significantly different from at least the null model with only the intercept.

Model Selection

Model selection can use deviance scores to test if models are different. However, in large samples, models that are similar may still be significantly different. One method is to look at the standardized residuals from the analysis. Large standardized residuals indicate that additional conditional effects should be considered as part of the model. Another method of model selection is to use the dissimilarity index (D), described further in the Definitions and Concepts section of this wiki. The larger the dissimilarity index, the larger the difference between the current model and the ideal model respective to the data.

Definitions & Concepts: Loglinear Models for Contingency Tables

Two-Way Tables

Log linear models are useful for predicting cell frequencies of contingency tables. These models are based on the table's marginal probabilities. The marginal probability is the sum of the probability for the cell's row and the probability for the cell's column and is stated as:

$$\Pi_{ij} = \Pi_{i+}\Pi_{+j}$$

The expected frequency of the cell is the marginal probability of the cell multiplied by the cell count and is represented as:

$$\mu_{ij} = n\Pi_{ij}$$

Thus, we can substitute the marginal probability for it's component parts.

$$\mu_{ij} = n\Pi_{i+}\Pi_{+j}$$

If we take the log of the expected cell frequency of a cell, we can develop the equation for the loglinear model (see below).

$$\log(\mu_{ij}) = \log(n\Pi_{i+}\Pi_{+j}) = \log(n) + \log(\Pi_{i+}) + \log(\Pi_{+j})$$
$$= \lambda + \lambda_i^X + \lambda_j^Y$$

Thus, for a two-way table, where the row variable is indicated by X and the column variable by Y:

The loglinear independence model is: $\log(\mu_{ij}) = \lambda + \lambda^X + \lambda^Y$

The loglinear saturated (dependence) model is: $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$

Therefore, the λ_{ij}^{XY} term determines the log odds ratio. You want to calculate all the columns without redundant information, so you either want to pick one category and set it to zero or make the sum of the categories equals zero. To determine whether X and Y are independent, conduct a likelihood ratio (LR) test with *df*=2 by making the λ_{ij}^{XY} term zero. Compare the deviance value to the χ^2 critical value of 5.991. If the deviance value is not significant, then the λ_{ij}^{XY} is not significantly different from 0 and the condition of independence holds.

Three-Way Tables

A three-way table is merely an expansion of a two-way table to include a third variable, Z:

The loglinear independence model is: $\log(\mu_{ijk}) = \lambda + \lambda^{X} + \lambda^{Y} + \lambda^{Z}$

The loglinear saturated model is: $\log(\mu_{ijk}) = \lambda + \lambda_i^{X} + \lambda_j^{Y} + \lambda_k^{Z} + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$

As you can tell, there are many intermediate models that are possible. For instance, you might be testing the relationship between exercise (X), diet (Y), and vitamins (Z). In this example, suppose that once you control for diet, there is no relationship between exercise and vitamins. In this case, there are two two-way interactions (XY, YZ) but no interaction between XZ and no three-way interaction (XYZ). The model would look like:

$\log(\mu_{ijk}) = \lambda + \lambda_i^{X} + \lambda_j^{Y} + \lambda_k^{Z} + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$

To compare models, you can look at the change in deviance using the LR test as long as the models are nested. To test that models are nested, the setting of a term to 0 should result in equivalent models. If you do not have nested models, then you will conduct model comparison using the AIC. For the AIC, if the value decreases when the additional term is included, then the more complex model better fits the data. If the AIC does not decrease by a large enough degree, then the simpler model is assumed to be an equally good fit for the data.

Dissimilarity Index

There is an issue to consider when we have a very large sample size (in the thousands). While you are quite lucky to have access to such a large sample, you are also likely to find that almost any reduced model will not fit the data as well as the saturated model. This is because a test statistic is related to both effect size and sample size. So, use the dissimilarity index to get an idea of how closely two models fit the data, regardless of sample size:

$$D = \frac{\sum |p_i - \hat{\pi}_i|}{2}$$

where p_i are the observed cell probabilities, and π_i are the cell probabilities estimated in the model. D ranges from 0 to 1, and smaller values of D indicate more similar models. The dissimilarity index indicates the proportion of cases needed to be moved to achieve perfect fit. Thus, the larger the dissimilarity index, the farther the current model is from the ideal model.

The Connection between Loglinear and Logistic Models

Logistic models have a specific outcome that you are trying to predict (a particular DV), so logistic models do not examine the relationship between all of the variables. In loglinear models, however, there are no particular outcomes that are being modeled, so all possible relationships are included in the model.

Equivalent models:

$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

logit $[P(Y=1)]$

$= \log [P(Y=1)/(1-P(Y=1)]$

$= \log [P(Y=1| X=i, Z=k) / (1 - P(Y=1| X=i, z = k)]$

So, $\log(\mu_{i1k}/\mu_{i2k}) = (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} + \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} + \lambda_{2k}^{YZ}) = \alpha + \beta X + \beta Z$

For instance, the loglinear model of Y, XZ is equivalent to the logistic model logit $[P(Y=1)] = \alpha$ because there is no relationship between Y and X or between Y and Z. Similarly, the loglinear model of XYZ is equivalent to the logistic model logit $[P(Y=1)] = \alpha + \beta X + \beta Z + \beta XZ$.

Independence Graphs and Collapsibility

Independence graphs allow us to identify conditionally independent relationships by presenting them in a visual form. These can be particularly helpful when complex (e.g., four-way) tables are analyzed by generating a visual format of the relationships between the variables. As the table becomes more complex, it is easier to become confused due to the number of variables and relationships. Having an independence graph allows the user to clarify exactly what they are trying to model. The user can then run the analysis that best represents the graph.

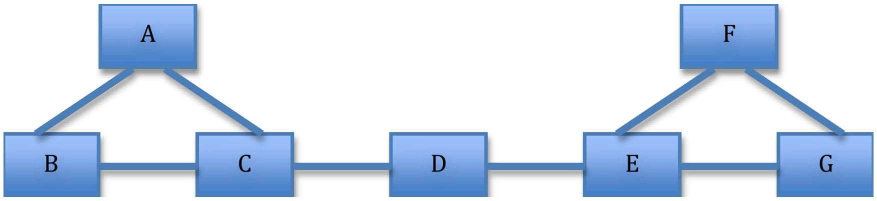
For instance, your model contains the following relationships: WX, XY, and YZ. Draw links between any proposed connections:

W ------- X --------- Y ---------- Z

Conditional independence occurs when there is no direct connection between points, but they are connected indirectly through intermediary points. So, W and Y are conditionally independent on X, and W and Z are independent conditional on both X and Y.

You can only collapse a three-way model if the variables are conditionally independent. For a three-way table ABC, A and B have the same marginal and conditional odds ratios if either (1) A and C are conditional independent or (2) B and C are conditionally independent.

In the next model, the following relationships exist: ABC, EFG, CD, and DE. The collapsibility chart is depicted below:



As chart above shows, to examine the relationship between ABC and D, you can collapse over EFG.

Log Linear Models with Ordinal Data

Although it is not the most ideal situation, you might have ordinal data on your hand, and you can use log linear models to analyze it. The equation for log linear models with nominal data is: $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$

For a contingency table with ordinal data for the two variables, the equation is

$$\log\left(\mu_{ij}\right) = \lambda + \lambda_i^X + \lambda_j^Y + \beta a_i b_j$$

where we replace the interaction term with parameter estimate multiplied by the row and column frequencies of the cell.

Let us say that we have the following contingency table with each variable being ordinal with two levels.

	n	p
m	μ_{mn}	μ_{mp}
o	μ_{on}	μ_{op}

If we want to compare the expected frequency relations between the different columns across the rows, we would have the following equation:

$$\frac{\mu_{mn}/\mu_{mp}}{\mu_{on}/\mu_{op}}$$

To accomplish this, we would take the log of this equation and break it down into it's component parts as is seen below.

$$\log\left[\frac{\mu_{mn}/\mu_{mp}}{\mu_{on}/\mu_{op}}\right] = \log\left(\mu_{mn}/\mu_{mp}\right) - \log\left(\mu_{on}/\mu_{op}\right) =$$
$$\log\left(\mu_{mn}\right) - \log\left(\mu_{mp}\right) - \log\left(\mu_{on}\right) + \log\left(\mu_{op}\right) =$$
$$\beta\left(\left(1*1\right) - \left(1*2\right) - \left(2*1\right) + \left(2*2\right)\right) = \beta\left(1\right) =$$
$$\beta\left(p-n\right)\left(o-m\right)$$

Thus, we take the log of numerator and denominator and then break up and take the log of the numerators and denominators. You would then multiply the parameter estimate by the sum of the multiplication of the row and column numbers for each component, which should sum to one. This leaves only the parameter estimate, which you the difference between each column and each row. The local odds ratio for the different cells is directly related to β if the difference between the different levels is the same. You can calculate the odds ratios by taking the exponent of the parameter estimates.

Additionally, if you there are multiple levels for the variables, you only have to calculate the odds ratio regarding four cells to know the odds ratios for other combinations of cells if it is an uniform association model, which indicates that the odds are the same regardless which of the four cells you look at due to the equal spacing between the variable levels.

Example

The data from this example come from Schoemann & Seta (2005). Eighty-nine People participated in an experiment about selecting a leader. Participants were placed in a group, told their group would be completing either logic problems or a group discussion and asked to select a leader with was either prototypical of the group (a leader with the same values as group members) or a stereotypical leader (a leader who is intelligent, has good communication skills, good planning skills, and grace under pressure. The results of the study are shown below.

		Prototypical	Stereotypical
Males			
	Logic Problems	5	15
	Discussions	5	14
Females			
	Logic Problems	6	19
	Discussions	12	13

The equation for the saturated model is (g=gender, t=task, l=leader):

$$\log(\mu_{ijk}) = \lambda + \lambda_i^{g} + \lambda_j^{t} + \lambda_k^{l} + \lambda_{ij}^{gt} + \lambda_{ik}^{gl} + \lambda_{jk}^{tl} + \lambda_{ijk}^{gtl}$$

SAS code:

```
data leader;
input gender $ task $ leader$ count @@;
datalines;
M L P 5
M L S 15
M D P 5
M D S 14
```

```
F L P 6
F L S 19
F D P 12
F D S 13
;
```

Saturated Model:

SAS Code:

```
PROC catmod data=leader;
WEIGHT count;
MODEL gender*task*leader = _response_ /noresponse noiter;
LOGLIN gender*task*leader;
Run;
```

Results:

```
                                         Standard     Chi-
Parameter                     Estimate      Error   Square    Pr > ChiSq
_____________________________________________________________________________
gender               F          0.1525     0.1197     1.62        0.2028
task                 D          0.0306     0.1197     0.07        0.7984
gender*task          F D        0.0478     0.1197     0.16        0.6896
leader               P         -0.4201     0.1197    12.31        0.0005
gender*leader        F P        0.1119     0.1197     0.87        0.3499
task*leader          D P        0.1427     0.1197     1.42        0.2334
gender*task*leader   F D P      0.1255     0.1197     1.10        0.2948
```

Based on the Wald chi square the only significant parameter is that of leader, thus the number of participants selecting each type of leader is not equal. To determine odds of selecting a stereotypical leader over a prototypical leader we take the exponent of the estimate. So exp(-0.4201)=0.657.The odds of selecting a prototypical leader are 0.657 times that of selecting a stereotypical leader. To make this more interpretable, we can compute the odds of selecting a stereotypical leader by taking 1/(exp(-0.4201), which equals 1.522.

Restricted model:

We can test a more restricted model, which does not include any conditional terms (a main effects model).

SAS Code:

```
 PROC CATMOD data=leader;
WEIGHT count;
MODEL gender*task*leader = _response_ /noresponse noiter;
```

```
LOGLIN gender task leader;
Run;
```

Results:

```
 Source              DF   Chi-Square    Pr > ChiSq
______________________________________________
gender                1       1.35          0.2448
task                  1       0.01          0.9156
leader                1      11.64          0.0006

Likelihood Ratio      4       4.30          0.3674
```

The Likelihood ratio tests if the model fit is different from the saturated model. In this case the model does not fit differently from the saturated model. Thus, we would prefer this model as it is more parsimonious and fits as well as the saturated model.

```
                                  Standard      Chi-
 Parameter       Estimate         Error        Square    Pr > ChiSq

gender    F      0.1242         0.1068         1.35        0.2448
 task     D     -0.0112         0.1060         0.01        0.9156
 leader   P    -0.3893         0.1141        11.64        0.0006
```

Once again, only leader is significant. By taking the exponent of the estimated parameter we can determine the odds of selecting a prototypical versus a stereotypical leader.

Example 2

This example will use the same data as the previous example but will demonstrate how to do it in R. Here are the c.ode for uploading the data and creating the contingency table.

```
library(MASS)

task<-factor(rep(rep(c("Logic","Discussions"),c(2,2)),2))

leader<-factor(rep(c("Prototypical","Stereotypical"),4))

gender<-factor(rep(c("Male","Female"),c(4,4)))

freq<-c(5,15, 5,14, 6,19, 12,13)

example<-data.frame(task,leader,gender,freq)

example

examplet<-xtabs(freq~task+leader+gender,data=example)

examplet
```

You should see the data appear in a contingency table like the one in the first example except that "Female" will come before "Male."

Now we want to determine the marginal probabilities of the three variables using a loglinear model. Here is the code for this model. It does not provide you with the parameter estimates automatically, so you have to ask for them.

```
model1<-loglm(~task+leader+gender,data=examplet,param=TRUE)

summary(model1)

model1$param
```

```
                        X^2     df  P(> X^2)
Likelihood Ratio   4.296170  4   0.3674049
Pearson            4.482641  4   0.3446108

$`(Intercept)`
[1] 2.327485

$task
Discussions       Logic
-0.01123643  0.01123643

$leader
 Prototypical Stereotypical
   -0.3893347     0.3893347

$gender
    Female       Male
 0.1242307 -0.1242307
```

This model provides us with an overall X^2 value, the degrees of freedom, and p-value for the likelihood ratio test. The parameter estimates sum to zero, which makes sense since when the odds for one increases by some amount the odds for the other one are that much lower. The likelihood ratio test indicates that this model is not significantly different from the saturated model. Unfortunately, these models do not provide the X^2 estimates and p-values for the individual variables, so it only provides a limited amount of information.

The code also provides the parameter estimates for the variables, which are log odds as in previous chapters. To get the odds ratio for each variable, you will have to find the exponent of the log odds.

We can determine the adequacy of this model in comparison with the ideal model by calculating the dissimilarity index, which you can do using the following code.

```
sum(abs(examplet-fitted(model1)))/(2*sum(examplet))
```

```
Re-fitting to get fitted values
[1] 0.0917473
```

This dissimilarity index number indicates that 9% of the values need to be shifted to approximate the ideal model. In this case, this is not a bad number, but it's still not perfect.

We can next try a model with conditional probabilities, such as the interaction between task and leader. The code for the model and dissimilarity index would be as follows:

```
model2<-loglm(~task+leader+gender+task*leader,data=examplet,param = TRUE)
summary(model2)
model2$param
sum(abs(examplet-fitted(model2)))/(2*sum(examplet))
```

The results from this model are:

```
                      X^2 df  P(> X^2)
Likelihood Ratio 2.206616  3 0.5306464
Pearson          2.154245  3 0.5410169
> model3$param
$`(Intercept)`
[1] 2.312482

$task
Discussions       Logic
  0.0511986  -0.0511986

$leader
 Prototypical Stereotypical
   -0.3977722     0.3977722

$gender
    Female       Male
 0.1242307 -0.1242307

$task.leader
             leader
task          Prototypical Stereotypical
  Discussions    0.1664604    -0.1664604
  Logic         -0.1664604     0.1664604

Re-fitting to get fitted values
[1] 0.05504356
```

Unlike the first model, this model did not have a significant likelihood ratio test, indicating that this model is not significantly different from the saturated model. We can determine whether the current model is significantly different from the first model by using the following code.

```
anova(model1,model2)
```

```
          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   4.296170  4
Model 2   2.206616  3   2.089554        1        0.14831
Saturated 0.000000  0   2.206616        3        0.53065
```

Thus, the current model is not significantly different from the model with only the marginal probabilities and the saturated model, which indicates that the most parsimonious model for this data is one with task, leader, and gender only. Though the dissimilarity index did drop from the first model, but this not necessarily enough to prevent our using the first model.