BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part VII

Loglinear Model

Introduction

Previously, we have learnt some models for XYZ three-way contingency table. Specifically,

- Y: Response of interest
- X: Treatment
- Z: Center

If Y takes two levels, we can apply logistic model.

If Y takes more than two levels, we can apply multicategory logit model.

In the above three-way table case, both models describe how a categorical response depends on a set of categorical explanatory variables.

Question: what to do when there is no clear distinction between response and explanatory variables?

Motivating Example

Data is collected from a survey conducted by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked students in their final year of a high school near Dayton, Ohio whether they had ever used alcohol (A), cigarettes(C), or marijuana(M).

Alcohol Use	Cigarette Use	Drug Use: Yes	Drug Use: No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Each of A, C, and M is a binary variable, and the cross-tabulate leads to this $2 \times 2 \times 2$ table.

- Are A, C, and M independent of each other?
- If not, how to measure the strength of association?

Overview

Here we introduce loglinear models, which focus on associations between categorical response variables and do not distinguish response variable and explanatory variables.

- Loglinear model for two-way tables
  - Independence model
  - Saturated model
- Loglinear model for three-way tables
  - Mutual independence model
  - Joint independence model
  - Conditional independence model
  - Homogeneous model
  - Saturated model
- Loglinear-Logistic connection

The emphasis is to understand loglinear models and make connection with contingency analysis and logistic model. The inference and model selection have already studied.

Loglinear model

Loglinear models model cell counts for contingency tables. The focus of loglinear models is on statistical independence and dependence. Therefore, there is no clear distinction between response and explanatory variables.

When cross-classified n subjects, the cell counts can be modeled by multinomial distribution.

Recall the connection between multinomial and Poison distribution. If $X_1, X_2, \ldots, X_c$ are independent Poisson variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_c$, respectively. Then the joint conditional distribution of $X_1, X_2, \ldots, X_c$ given $\sum X_i = n$, i.e., $X_1, X_2, \ldots, X_c | \sum X_i = n$, is multinomial with parameter n and $\pi_i = \lambda_i/(\lambda_1 + \lambda_2 + \ldots + \lambda_c)$.

Therefore, we can model those multinomial cell counts using Poisson models. Because of the $\sum X_i = n$ constraint, the intercept is not a meaningful parameter, but a normalizing constant to ensure that the cell probabilities add up to 1.

Loglinear model is just Poisson model for contingency tables. Therefore, codes, inference, model comparison should not look foreign.

Loglinear model

Loglinear models **model cell counts** for contingency tables. The focus of loglinear models is on statistical independence and dependence. Therefore, there is no clear distinction between response and explanatory variables.

When cross-classified n subjects, the cell counts can be modeled by **multinomial distribution**.

Recall the **connection between multinomial and Poison distribution**. If $X_1, X_2, \ldots, X_c$ are independent Poisson variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_c$, respectively. Then the joint conditional distribution of $X_1, X_2, \ldots, X_c$ given $\sum X_i = n$, i.e., $X_1, X_2, \ldots, X_c | \sum X_i = n$, is multinomial with parameter n and $\pi_i = \lambda_i/(\lambda_1 + \lambda_2 + \ldots + \lambda_c)$.

Therefore, we can model those multinomial cell counts using **Poisson models**. Because of the $\sum X_i = n$ constraint, the intercept is not a meaningful parameter, but a normalizing constant to ensure that the cell probabilities add up to 1.

Loglinear model is just Poisson model for contingency tables. Therefore, codes, inference, model comparison should not look foreign.

Independence model for $I \times J$ two-way tables

Independence model $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$.

- λ: normalizing constant
- λ_i^X: row effect for $X = i$. Only $I - 1$ are non-redundant. Use dummy coding scheme and set the last level as reference $\rightarrow \lambda_I^X = 0$.
- λ_j^Y: column effect for $Y = j$. Only $J - 1$ are non-redundant. Use dummy coding scheme and set the last level as reference$\rightarrow \lambda_J^Y = 0$.
- Differences between two parameters for a given variables relate to the log odds of making one response, relative to another, on that variable.

In this model

- Number of cells: IJ
- Number of parameters in the model: $1 + (I - 1) + (J - 1) = I + J - 1$
- Degree of freedom: $IJ - (I + J - 1) = (I - 1)(J - 1)$

That is, the goodness of fit test for this independence model is $(I - 1)(J - 1)$. This goodness of fit test is exactly the X^2 and G^2 tests of independence for two-way table that we introduced in Chapter 2.

Independence model for $I \times 2$ table = intercept-only logit model

When $J = 2$, this independence model corresponds to the logit model with only intercept.

$$\text{logit}(\text{P(Y=1)}) = \alpha, \quad \text{where} \quad \alpha = \lambda_1^Y - \lambda_2^Y$$

To see this, we take row i

$$\begin{aligned}
\text{logit}(\text{P(Y=1)}) &= \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) \\
&= \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) \\
&= \log(\mu_{i1}) - \log(\mu_{i2}) \\
&= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) \\
&= \lambda_1^Y - \lambda_2^Y
\end{aligned}$$

This logit does not dependent on i, that is, does not depend on level of X.

In each row, the odds of response in column 1 equal $e^{\alpha} = e^{\lambda_1^Y - \lambda_2^Y}$

Example of Independence Model: Afterlife

	Belief in Afterlife	
Race	Yes	No
White	1339	300
Black	260	55
Other	88	22

Table 7.1. Results of Fitting Independence Loglinear Model to Cross-Classification of Race by Belief in Life after Death

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	2	0.3565
Pearson Chi-Square	2	0.3601

Parameter		DF	Estimate	Standard Error
Intercept		1	3.0003	0.1061
race	white	1	2.7014	0.0985
race	black	1	1.0521	0.1107
race	other	0	0.0000	0.0000
belief	yes	1	1.4985	0.0570
belief	no	0	0.0000	0.0000

Example of Independence Model: Afterlife (cont.)

Reading the output, it is dummy coding scheme where *race* = *others* and *belief* = *no* are the reference.

Therefore, the independence model

$$\log(\mu_{ij}) = \lambda + \lambda_i^R + \lambda_j^B$$

can be written as

$$\log(\mu_{ij}) = \lambda + \lambda_1^R R_1 + \lambda_2^R R_2 + \lambda_1^B B_1,$$

where

$$R_1 = \begin{cases} 1 & \text{white} \\ 0 & \text{otherwise} \end{cases} \qquad R_2 = \begin{cases} 1 & \text{black} \\ 0 & \text{otherwise} \end{cases} \qquad B_1 = \begin{cases} 1 & \text{belief=yes} \\ 0 & \text{belief=no} \end{cases}$$

The prediction model

$$\log(\hat{\mu}_{ij}) = 3.00 + 2.70R_1 + 1.05R_2 + 1.50B_1$$

Deviance = 0.36 with df =2 , no evidence of lack of fit.

For each race, the estimated odds of belief in afterlife is $e^{1.5} = 4.5$

Saturated model for $I \times J$ two-way tables

Independence model $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$.

- λ: normalizing constant
- λ_i^X: row effect for $X = i$. Only $I - 1$ are non-redundant. Use dummy coding scheme and set the last level as reference $\rightarrow \lambda_I^X = 0$.
- λ_j^Y: column effect for $Y = j$. Only $J - 1$ are non-redundant. Use dummy coding scheme and set the last level as reference$\rightarrow \lambda_J^Y = 0$.
- λ_{ij}^{XY}: association parameters. Using the above coding scheme, $\lambda_{iJ} = \lambda_{Ij} = 0$ for $i = 1, \ldots, I; j = 1, \ldots, J$.
- There is direct relationship between log odds ratios and $\{\lambda_{ij}^{XY}\}$ association parameters.

In this model

- Number of cells: IJ
- Number of parameters in the model: $1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = IJ$
- Degree of freedom: 0

Relationship between log odds ratios and $\{\lambda_{ij}^{XY}\}$

Consider log odds ratio comparing levels i and i' of X and j and j' of Y,

$$\begin{aligned}
\log\left(\frac{\mu_{ij}\mu_{i'j'}}{\mu_{i'j}\mu_{ij'}}\right) &= \log(\mu_{ij}) + \log(\mu_{i'j'}) - \log(\mu_{i'j}) - \log(\mu_{ij'}) \\
&= (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) + (\lambda + \lambda_{i'}^X + \lambda_{j'}^Y + \lambda_{i'j'}^{XY}) \\
&\quad -(\lambda + \lambda_{i'}^X + \lambda_j^Y + \lambda_{i'j}^{XY}) - (\lambda + \lambda_i^X + \lambda_{j'}^Y + \lambda_{ij'}^{XY}) \\
&= \lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{i'j}^{XY} - \lambda_{ij'}^{XY}
\end{aligned}$$

Hence, the odds ratio is $e^{\lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{i'j}^{XY} - \lambda_{ij'}^{XY}}$.

Under saturated model, the expected cell counts are original observations. Therefore, this estimated odds ratio is also the same as

$$\frac{n_{ij}n_{i'j'}}{n_{i'j}n_{j'i}}.$$

Example of Saturated Model: Afterlife

	Belief in Afterlife	
Race	Yes	No
White	1339	300
Black	260	55
Other	88	22

Table 7.2. Estimates for Fitting Saturated Loglinear Model to Cross-Classification of Race by Belief in Life after Death

Parameter			DF	Estimate	Standard error
Intercept			1	3.0910	0.2132
race	white		1	2.6127	0.2209
race	black		1	0.9163	0.2523
race	other		0	0.0000	0.0000
belief	yes		1	1.3863	0.2384
belief	no		0	0.0000	0.0000
race*belief	white	yes	1	0.1096	0.2468
race*belief	white	no	0	0.0000	0.0000
race*belief	black	yes	1	0.1671	0.2808
race*belief	black	no	0	0.0000	0.0000
race*belief	other	yes	0	0.0000	0.0000
race*belief	other	no	0	0.0000	0.0000

Example of Saturated Model: Afterlife (cont.)

Using the same coding scheme, the saturated model

$$\log(\mu_{ij}) = \lambda + \lambda_i^R + \lambda_j^B + \lambda_{ij}^{RB}$$

can be written as

$$\log(\mu_{ij}) = \lambda + \lambda_1^R R_1 + \lambda_2^R R_2 + \lambda_1^B B_1 + \lambda_{11}^{RB} R_1 B_1 + \lambda_{21}^{RB} R_2 B_1,$$

The prediction model

$$\log(\hat{\mu}_{ij}) = 3.09 + 2.61 R_1 + 0.92 R_2 + 1.39 B_1 + 0.11 R_1 B_1 + 0.17 R_2 B_1$$

The estimated odds ratios between belief and race are

- $e^{0.11} = 1.12$ for white and other
- $e^{0.17} = 1.18$ for black and other
- $e^{0.11-0.17} = 0.94$ for white and black. The estimated odds of belief in afterlife for whites are 0.94 times the estimated odds for blacks.

Recall that the independence model fitted well, none of these estimated odds ratios differ significantly from 1.

Loglinear Model for Three-way Table

With three-way contingency tables, loglinear models can represent various independence and association patterns.

- Model (X, Y, Z): This model only has main effects for X, Y, and Z. This says that variables are mutually independent.
- Model (XY, Z): This model has all main effects and the XY association. This says that X and Y could be related, but Z is unrelated to X or Y. That is XY is jointly independent with Z.
- Model (XY, XZ): This model includes all main effects, and XY and XZ association. This says that Y and Z are conditionally independent given X. That is, the $Y \times Z$ odds ratios at each level of X are 1.
- Model (XY, XZ, YZ): This models includes all main effects and two-way associations. This model says that the association between any pair of variables is identical across all levels of the third variable.
- Model (XYZ): This model is the saturated model. This allows the relationship between any pair of variables to vary across the levels of the third.

Model for Various Independence

Model (X, Y, Z)

Mutual Independence $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$

Mutual Independence Model $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$

Model (XY, Z)

Joint Independence $\pi_{ijk} = \pi_{ij+}\pi_{++z}$

Joint Independence Model $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}.$

Model (XY, XZ)

Conditional Independence $\pi_{ijk|i} = \pi_{ij+|i}\pi_{i+k|i}$

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

Recall that two-factor terms describes conditional association.

Homogeneous Association Model (XY, XZ, YZ)

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

Model interpretation refers to the highest-order parameters. To understand those two-factor terms, we consider log odds ratio comparing levels i and i' of X and j and j' of Y given $Z = k$.

$$\begin{aligned}
\log(\theta_{XY(k)}) &= \log\left(\frac{\mu_{ijk}\mu_{i'j'k}}{\mu_{i'jk}\mu_{ij'k}}\right) \\
&= \log(\mu_{ijk}) + \log(\mu_{i'j'k}) - \log(\mu_{i'jk}) - \log(\mu_{ij'k}) \\
&= (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}) \\
&\quad +(\lambda + \lambda_{i'}^X + \lambda_{j'}^Y + \lambda_k^Z + \lambda_{i'j'}^{XY} + \lambda_{i'k}^{XZ} + \lambda_{j'k}^{YZ}) \\
&\quad -(\lambda + \lambda_{i'}^X + \lambda_j^Y + \lambda_k^Z + \lambda_{i'j}^{XY} + \lambda_{i'k}^{XZ} + \lambda_{j'k}^{YZ}) \\
&\quad -(\lambda + \lambda_i^X + \lambda_{j'}^Y + \lambda_k^Z + \lambda_{ij'}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}) \\
&= \lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{i'j}^{XY} - \lambda_{ij'}^{XY}
\end{aligned}$$

The expression of log odds ratio does not depend on k, so the odds ratio is the same at every level of Z. Similarly, this model has equal XZ odds ratios at different levels of Y, and equal YZ odds ratios at different levels of X.

Two-Factor Parameters Describe Conditional Associations

In the loglinear model (XY, XZ, YZ), we show that the two-factor terms describe conditional odds ratio. Specifically,

$$\begin{aligned} \theta_{XY(k)} &= e^{\lambda_{ij}^{XY}+\lambda_{i'j'}^{XY}-\lambda_{i'j}^{XY}-\lambda_{ij'}^{XY}} \\ \theta_{XZ(j)} &= e^{\lambda_{ik}^{XZ}+\lambda_{i'k'}^{XZ}-\lambda_{i'k}^{XZ}-\lambda_{ik'}^{XZ}} \\ \theta_{YZ(i)} &= e^{\lambda_{jk}^{YZ}+\lambda_{j'k'}^{YZ}-\lambda_{j'k}^{YZ}-\lambda_{jk'}^{YZ}} \end{aligned}$$

This conclusion can be generalized to other model whose highest order is two-factor. For example,

- For model (XY, Z) and (XY, XZ), you can interpret parameters by

$$\theta_{XY(k)} = e^{\lambda_{ij}^{XY}+\lambda_{i'j'}^{XY}-\lambda_{i'j}^{XY}-\lambda_{ij'}^{XY}}.$$

- For model (XY, XZ), you can interpret parameters by

$$\theta_{XZ(j)} = e^{\lambda_{ik}^{XZ}+\lambda_{i'k'}^{XZ}-\lambda_{i'k}^{XZ}-\lambda_{ik'}^{XZ}}.$$

Homogeneous Association Loglinear Model for $I \times 2 \times K$ Table $=$ Logit Model With Main Effects

When $J = 2$, we may treat it as a response and X and Z are explanatory.

$$\text{logit}(\text{P(Y=1)}) = \alpha + \beta_i^X + \beta_k^Z,$$

where $\alpha = \lambda_1^Y - \lambda_2^Y$, $\beta_i^X = \lambda_{i1}^{XY} - \lambda_{i2}^{XY}$, and $\beta_k^Z = \lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}$.

To see this, take X at its level i and Z at its level k,

$$\begin{aligned}
\text{logit}(\text{P(Y=1)}) &= \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) \\
&= \log\left(\frac{P(Y=1|X=i,Z=k)}{1-P(Y=1|X=i,Z=k)}\right) \\
&= \log\left(\frac{\mu_{i1k}}{\mu_{i2k}}\right) = \log(\mu_{i1k}) - \log(\mu_{i2k}) \\
&= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}) \\
&\quad -(\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}) \\
&= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}) \\
&= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ})
\end{aligned}$$

In the above derivation, the term λ_{ik}^{XZ} cancels out.

- The logistic model does not describe relationships among explanatory variables, so it assumes nothing about their association structure.

It might seem that loglinear model (XY, YZ) leads to the same logistic model form. However, to obtain exactly same model fit, we need the loglinear model (XY, YZ, XZ).

Equivalent Loglinear and Logistic Models for a Three-Way Table With Binary Response Variable Y

Loglinear Symbol	Logistic Model
(Y, XZ)	α
(XY, XZ)	$\alpha + \beta_i^X$
(YZ, XZ)	$\alpha + \beta_k^Z$
(XY, YZ, XZ)	$\alpha + \beta_i^X + \beta_k^Z$
(XYZ)	$\alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$

- In each pairing of models in this table, the loglinear model contains the XZ association term relating the variables that are explanatory in the logistic models.

Example: Drug Use

Alcohol Use	Cigarette Use	Drug Use Yes	Drug Use No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Table 7.6. Output for Fitting Loglinear Model to Table 7.3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	1	0.3740
Pearson Chi-Square	1	0.4011

Parameter			Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			5.6334	0.0597	8903.96	<.0001
a	1		0.4877	0.0758	41.44	<.0001
c	1		−1.8867	0.1627	134.47	<.0001
m	1		−5.3090	0.4752	124.82	<.0001
a*m	1	1	2.9860	0.4647	41.29	<.0001
a*c	1	1	2.0545	0.1741	139.32	<.0001
c*m	1	1	2.8479	0.1638	302.14	<.0001

LR Statistics

Example: Drug Use

Model (AC, AM, MC) permits all pairwise associations but has homogeneous odds ratios. For example,

- The AC fitted conditional odds ratios for this model equal 7.8.

$$e^{2.0545} = 7.8$$

  For each level of M, students who have smoked cigarettes have estimated odds of having drunk alcohol that are 7.8 times the estimated odds for students who have not smoked cigarettes.

Model for Various Independence

Model (X, Y, Z)

Mutual Independence $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$

Mutual Independence Model $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$

Model (XY, Z)

Joint Independence $\pi_{ijk} = \pi_{ij+}\pi_{++z}$

Joint Independence Model $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}.$

Model (XY, XZ)

Conditional Independence $\pi_{jk|i} = \pi_{j+|i}\pi_{+k|i}$

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

We will see later that two-factor terms describes conditional association.

Example: Drug Use

We can also calculate the odds ratio from estimated/fitted cell counts.

Table 7.4. Fitted Values for Loglinear Models Applied to Table 7.3

Alcohol Use	Cigarette Use	Marijuana Use	Loglinear Model (A, C, M)	(AC, M)	(AM, CM)	(AC, AM, CM)	(ACM)
Yes	Yes	Yes	540.0	611.2	909.24	910.4	911
		No	740.2	837.8	438.84	538.6	538
	No	Yes	282.1	210.9	45.76	44.6	44
		No	386.7	289.1	555.16	455.4	456
No	Yes	Yes	90.6	19.4	4.76	3.6	3
		No	124.2	26.6	142.16	42.4	43
	No	Yes	47.3	118.5	0.24	1.4	2
		No	64.9	162.5	179.84	279.6	279

Example: Drug Use

Table 7.5. Estimated Odds Ratios for Loglinear Models in Table 7.4

	Conditional Association			Marginal Association		
Model	AC	AM	CM	AC	AM	CM
(A, C, M)	1.0	1.0	1.0	1.0	1.0	1.0
(AC, M)	17.7	1.0	1.0	17.7	1.0	1.0
(AM, CM)	1.0	61.9	25.1	2.7	61.9	25.1
(AC, AM, CM)	7.8	19.8	17.3	17.7	61.9	25.1
(ACM) level 1	13.8	24.3	17.5	17.7	61.9	25.1
(ACM) level 2	7.7	13.5	9.7			

Goodness of Fit Tests

As Loglinear models deal with sets of categorical variables, we can apply the goodness of fit test.

$$G^2 = 2\sum n_{ijk} \log\left(\frac{n_{ijk}}{\hat{\mu}_{ijk}}\right), \quad X^2 = \sum \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$$

- Under the null hypothesis, both are distributed as χ^2_{df}.
- df is the number of cell counts minus the number of model parameters.
- Saturated model has $df = 0$
- Small p-value indicates poor model fit
- When it has poor model fit, we can investigate standardized residuals. Lack of fit is indicated by absolute values larger than about 2 when there are fewer cells or about 3 when there are many cells.

Testing Nested Models

We can apply likelihood ratio test to compare nested models.

For example, (AM, CM) is a nested model of (AC, AM, CM). Therefore, we can apply the likelihood ratio test. Essentially, we are testing for $\lambda^{AC} = 0$.

- Test statistic: $2L(AC, AM, CM) - 2L(AM, CM)$, here $L(\cdot)$ is the Log-likelihood function
- Test statistic: $G^2(AM, CM) - G^2(AC, AM, CM)$
- We can denote this as $G^2[(AM, CM)|(AC, AM, CM)]$
- Under the null hypothesis, the test statistic is distributed as χ^2_{df}, where df is the difference between the number of parameters of the two models, or the number of parameters tested in the null hypothesis.

Example: Drug Use

Table 7.7. Goodness-of-Fit Tests for Loglinear Models Relating Alcohol (A), Cigarette (C), and Marijuana (M) Use

Model	G^2	X^2	df	P-value*
(A, C, M)	1286.0	1411.4	4	<0.001
(A, CM)	534.2	505.6	3	<0.001
(C, AM)	939.6	824.2	3	<0.001
(M, AC)	843.8	704.9	3	<0.001
(AC, AM)	497.4	443.8	2	<0.001
(AC, CM)	92.0	80.8	2	<0.001
(AM, CM)	187.8	177.6	2	<0.001
(AC, AM, CM)	0.4	0.4	1	0.54
(ACM)	0.0	0.0	0	—

* P-value for G^2 statistic.