STA4030: Categorical Data Analysis
Loglinear Models for Contingency Tables

Instructor: Bojun Lu

School of Data Science
CUHK(SZ)

December 1, 2020

Agenda

1. 10.1 Introduction
2. 10.2 Loglinear Model for Two-way Tables
3. 10.3 Loglinear Models for Three-way Tables
4. 10.4 The Loglinear-Logit Connection
5. 10.5 Model Selection and Comparison

10.1 Introduction

In Chapter 7, we introduced loglinear models as generalized linear models (GLMs) using the log link function with a Poisson response. Their most common use is modeling cell counts in contingency tables.

The models specify how the expected count depends on levels of the categorical variables for that cell as well as associations and interactions among those variables.

The purpose of loglinear modeling is the analysis of associations and interactions patterns. Loglinear models are of use primarily when at least two variables are response variables.

10.2 Loglinear Model for Two-way Tables

10.2.1 Independence model in two way table

Loglinear models express logarithms of expected cell frequency $\mu_{ij} = n\pi_{ij}$ in additive terms. Consider an $I \times J$ table:

$X\backslash Y$	1	2	$\cdots$	J	
1	μ_{11}	μ_{12}	$\cdots$	μ_{1J}	μ_{1+}
2	μ_{21}	μ_{22}	$\cdots$	μ_{2J}	μ_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
I	μ_{I1}	μ_{I2}	$\cdots$	μ_{IJ}	μ_{I+}
Total	μ_{+1}	μ_{+2}	$\cdots$	μ_{+J}	μ

10.2 Loglinear Model for Two-way Tables

Under an independence model, we have

for $\forall i, j$, $\pi_{ij} = \pi_{i+}\pi_{+j}$, or equivalently
$\mu_{ij} = \mu_{i+}\mu_{+j}/n$, or equivalently
$\log \mu_{ij} = \log \mu_{i+} + \log \mu_{+j} - \log n.$

This additive form suggests we should express $\log \mu_{ij}$ as

$$\log \mu_{ij} = \mu + \lambda_i^X + \lambda_j^Y, \; i = 1, \ldots, I; \; j = 1, \ldots, J. \tag{10.1}$$

The parameter λ_i^X represents the effect of classification in row i for variable X. The larger the value of λ_i^X, the larger each expected frequency is in row i of the table. When $\lambda_h^X = \lambda_i^X$, each expected frequency in row h equals the corresponding expected frequency in row i.

Similarly, the parameter λ_j^Y represents the effect of classification in column j for variable Y.

10.2 Loglinear Model for Two-way Tables

10.2.2 Parameter constraints

For the independence model (10.1), one $\{\lambda_i^X\}$ parameter is redundant, and one $\{\lambda_j^Y\}$ parameter is redundant. This is analogous to ANOVA and multiple regression models with factors, which require one fewer dummy variable than the number of factor levels.

Section 8.3.2 showed three ways of eliminating the redundancy:

(1) set the parameter for the last level of each factor equal to 0;

(2) set the parameter for the first level of each factor equal to 0;

(3) let the parameters for each factor sum to 0, e.g. for 2×2 tables, $\lambda_1^X + \lambda_2^X = 0$, $\lambda_1^Y + \lambda_2^Y = 0$.

The choice of constraints is arbitrary. Some software for loglinear models set a parameter equal to 0, others have parameters sum to 0.

10.2 Loglinear Model for Two-way Tables

10.2.3 Interpretation of parameters

Loglinear models for contingency tables are examples of GLMs. For $I \times J$ tables, this GLM treats the $N = IJ$ cell counts as N independent observations of a Poisson random component.

For loglinear GLMs, the data are the N cell counts rather than the individual classifications of the n subjects. The expectations $\{\mu_{ij}\}$ of the cell counts are linked to the explanatory terms using the log link.

As formula (10.1) illustrates, loglinear models for contingency tables do not distinguish between response and explanatory classification variables. They treat all variables jointly as responses, modeling the cell count for all combinations of their levels.

10.2 Loglinear Model for Two-way Tables

Parameter interpretation is simplest for binary responses. Consider the independence model (10.1) for $I \times 2$ tables, with two columns being levels of a response Y. In row i, the logit for the probability π that $Y = 1$ equals

$$\begin{aligned}\log\left(\frac{\pi}{1-\pi}\right) &= \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) = \log \mu_{i1} - \log \mu_{i2} \\ &= (\mu + \lambda_i^X + \lambda_1^Y) - (\mu + \lambda_i^X + \lambda_2^Y) = \lambda_1^Y - \lambda_2^Y.\end{aligned}$$

The final term is a constant that does not depend on i; that is, the logit for Y does not depend on the level of X. The loglinear model of independence corresponds to the simple logit model having form

$$\text{logit}(\pi) = \alpha, \text{ where } \alpha = \lambda_1^Y - \lambda_2^Y.$$

In each row, the odds of response in column 1 equal $\exp(\alpha)$. The chance of classification in a particular column is the same in all rows.

10.2 Loglinear Model for Two-way Tables

Example 10.1: Belief in Afterlife.

The following Table 10.1 shows a 2×2 table of cell counts and the corresponding fitted value $\hat{\mu}_{ij}$ for the independence loglinear model:

$$\log \mu_{ij} = \mu + \lambda_i^X + \lambda_j^Y,$$

where X = gender (Female, Male),
Y = belief in the afterlife (Yes, No or Undecided)

Since the fitted values satisfy independence, they have an odds ratio of 1 $(\theta = (432.10 \times 131.10)/(377.90 \times 149.90) = 1)$.

They are close to the observed counts, and the goodness-of-fit statistics for testing that the independence loglinear model holds equal $X^2 = 0.2$ and $G^2 = 0.2$, with $df = 1$, the chi-squared P-value = 0.66. The independence model is plausible for the data.

10.2 Loglinear Model for Two-way Tables

Table 10.1 Loglinear parameter estimates for the independence model, relating belief in the afterlife (columns) to gender (rows)

Observed Frequency		Fitted Value		Log Fitted Value	
435	147	432.10	149.90	6.069	5.010
375	134	377.90	131.10	5.935	4.876

Parameter	Set 1	Set 2	Set 3
μ	4.876	6.069	5.472
λ_1^X	0.134	0	0.067
λ_2^X	0	-0.134	-0.067
λ_1^Y	1.059	0	0.529
λ_2^Y	0	-1.059	-0.529

10.2 Loglinear Model for Two-way Tables

The difference between two main effect parameters of a particular type is not changed!

E.g. for each set of estimates in Table10.1and for each i,

$$\log\left(\frac{\hat{\mu}_{i1}}{\hat{\mu}_{i2}}\right) = (\hat{\mu} + \hat{\lambda}_i^X + \hat{\lambda}_1^Y) - (\hat{\mu} + \hat{\lambda}_i^X + \hat{\lambda}_2^Y) = \hat{\lambda}_1^Y - \hat{\lambda}_2^Y = 1.059.$$

Thus, for each gender, the estimated odds of belief in the afterlife equal $\exp(1.059) = 2.9$.

Let $\pi = P(Y = 1) = P(\text{with belief in the afterlife})$, then

$$\text{logit}(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 1.059.$$

Thus, the estimated probability of belief in the afterlife

$$\hat{\pi} = \frac{\exp(1.059)}{1+\exp(1.059)} = 0.74.$$

10.2 Loglinear Model for Two-way Tables

10.2.4 Saturated model

Saturated model: when a model has as many independent parameters as table has cells. Saturated model provides a perfect fit for any set of frequencies. It is the most general model for two-way contingency tables.

The saturated model for $I \times J$ table is given by

$$\log \mu_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \ I = 1, \ldots, I; j = 1, \ldots, J.$$

Number of parameters:

μ	λ_i^X	λ_j^Y	λ_{ij}^{XY}
1	$I-1$	$J-1$	$IJ-I-J+1$

$\therefore$ Total number of parameters $= IJ =$ number of cells.

$\therefore$ Model saturated, $\hat{\mu}_{ij} = n_{ij}$.

10.2 Loglinear Model for Two-way Tables

The λ_{ij}^{XY} can be interpreted as the association parameters or the interaction terms that reflect deviations from independence of X and Y. If X and Y are independent, $\lambda_{ij}^{XY} = 0 \; \forall \; i$ and j.

Direct relationships exist between log odds ratios and the λ_{ij}^{XY} association parameters. For instance, the saturated model for 2×2 tables has log odds ratio

$$\begin{aligned}
\log \theta &= \log \left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} \right) \\
&= \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21} \\
&= (\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\mu + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\
&\quad - (\mu + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\mu + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\
&= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}.
\end{aligned}$$

Thus, λ_{ij}^{XY} determine the log odds ratio.

10.2 Loglinear Model for Two-way Tables

Example 10.2 (Example 10.1 continued): In Table 10.1

$$\hat{\theta} = (435 \times 134)/(147 \times 375) = 1.057 \text{ and } \log \hat{\theta} = 0.056.$$

The following Table10.2shows possible association parameter estimates that have a log odds ratio of

$$\hat{\lambda}_{11}^{XY} + \hat{\lambda}_{22}^{XY} - \hat{\lambda}_{12}^{XY} - \hat{\lambda}_{21}^{XY} = 0.056.$$

Table 10.2 Equivalent association parameter estimates for saturated loglinear model

Association Parameter	Set 1	Set 2	Set 3
λ_{11}^{XY}	0.056	0.0	0.014
λ_{12}^{XY}	0.0	0.0	-0.014
λ_{21}^{XY}	0.0	0.0	-0.014
λ_{22}^{XY}	0.0	0.056	0.014

10.2 Loglinear Model for Two-way Tables

In $I \times J$ tables, only $(I-1)(J-1)$ association parameters are nonredundant. These "interaction" parameters in the saturated model are coefficients of cross-products of $(I-1)$ dummy variables for X with $(J-1)$ dummy variables for Y. Tests of independence analyze whether these $(I-1)(J-1)$ parameters are equal to zero, so they have $df = (I-1)(J-1)$.

When $I = J = 2$, a single nonredundant parameter determines the odds ratio. For example, for Set 3 of eliminating the redundancy we have the constraints:

$$\lambda_{1j}^{XY} + \lambda_{2j}^{XY} = \lambda_{i1}^{XY} + \lambda_{i2}^{XY} = 0.$$

Then

$$\lambda_{11}^{XY} = -\lambda_{12}^{XY} = -\lambda_{21}^{XY} = \lambda_{22}^{XY}.$$

Thus,

$$\log\theta = 4\lambda_{11}^{XY} \quad \text{or} \quad \lambda_{11}^{XY} = \frac{1}{4}\log\theta.$$

10.3 Loglinear Models for Three-way Tables

10.3.1 The general three-way model

Any type of model for three-way contingency tables can be expressed as a special case of the following general loglinear model:

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz}.$$

Parameter interpretation:

μ:	overall mean
λ_i^x, λ_j^y, λ_k^z:	marginal effect
λ_{ij}^{xy}, λ_{ik}^{xz}, λ_{jk}^{yz}:	partial association, 2-factor effect, or the first order interaction
λ_{ijk}^{xyz}	3-factor effect, or the second order interaction

10.3 Loglinear Models for Three-way Tables

When certain parameters in the general model are set to zero, models with special structures are obtained.

We will consider only hierarchical models: higher order terms may be included only if the related lower order terms are included.

e.g.

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ijk}^{xyz}$$

is not a hierarchical model.

There are five types of these models in a three dimensional table.

10.3 Loglinear Models for Three-way Tables

10.3.2 The five types of model

Model 1: *Complete Independence Model:* (X, Y, Z)

$$\Pr(X = x, Y = y, Z = z) = \Pr(X = x)\Pr(Y = y)\Pr(Z = z)$$

This suggests the following loglinear model:

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z$$

with appropriate constraints and $df = IJK - I - J - K + 2$.

Model 2: *Independence of* X *and* (Y, Z) *jointly:* (X, YZ)

$$\Pr(X = x, Y = y, Z = z) = \Pr(X = x)\Pr(Y = y, Z = z)$$

This suggests the following loglinear model:

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{jk}^{yz}$$

with appropriate constraints and $df = (I - 1)(JK - 1)$.

10.3 Loglinear Models for Three-way Tables

There are two other versions of Model 2:

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy}$$
$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ik}^{xz}$$

Model 3: *Conditional Independence Model:* (XZ, YZ)

$$\Pr(X = x, Y = y \mid Z = z) = \Pr(X = x \mid Z = z) \Pr(Y = y \mid Z = z)$$

This suggests the following loglinear model:

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}$$

with appropriate constraints and $df = (I-1)(J-1)K$.

There are two other versions of this type of model:

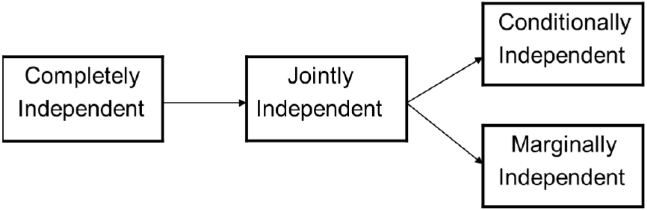
$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{jk}^{yz}$$
$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz}$$

10.3 Loglinear Models for Three-way Tables

The relationship of three independence:

Table 10.3 Summary of loglinear independence models

Model	Probabilistic form of π_{ijk}	Association terms in loglinear model	Interpretation
1	$\pi_{i++}\pi_{+j+}\pi_{++k}$	None	Variables are mutually independent
2	$\pi_{i++}\pi_{+jk}$	λ_{jk}^{YZ}	X is independent of Y and Z
3	$\pi_{i+k}\pi_{+jk}/\pi_{++k}$	$\lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	X and Y are independent, given Z



Completely Independent

Jointly Independent

Conditionally Independent

Marginally Independent

10.3 Loglinear Models for Three-way Tables

Model 4: *No second order interaction model:* (XY, YZ, XZ)

Pairwise associations exist among the three variables, with each two variables' interaction unaffected by the value of the third variable. That is

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}$$

with appropriate constraints and $df = (I-1)(J-1)(K-1)$.

Note:

For a $2 \times 2 \times 2$ table, the model of no second order interaction is equivalent to the model that specifies the equality of the odds ratios in levels of each variable.

10.3 Loglinear Models for Three-way Tables

Since $\mu_{ijk} = \exp(\mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz})$, we have
$\frac{\mu_{111}\mu_{221}}{\mu_{121}\mu_{211}} = \frac{\mu_{112}\mu_{222}}{\mu_{122}\mu_{212}}$,
$\theta_{XY(1)} = \theta_{XY(2)}$, equality of odds ratios for levels of 3rd variable

$\frac{\mu_{111}\mu_{212}}{\mu_{112}\mu_{211}} = \frac{\mu_{121}\mu_{222}}{\mu_{122}\mu_{221}}$,
$\theta_{XZ(1)} = \theta_{XZ(2)}$, equality of odds ratios for levels of 2nd variable

$\frac{\mu_{111}\mu_{122}}{\mu_{112}\mu_{121}} = \frac{\mu_{211}\mu_{222}}{\mu_{212}\mu_{221}}$,
$\theta_{YZ(1)} = \theta_{YZ(2)}$, equality of odds ratios for levels of 1st variable

The equivalence can be verified by direct substitution.

For this model, each pair has homogeneous association. Model 4 is called the loglinear model of *homogeneous association*.

10.3 Loglinear Models for Three-way Tables

Model 5: *Second-order interaction model:* (XYZ)

A second-order interaction (three-factor effect) relates all three variables, so the interaction between any two variables does depend on the value of the third variable. The model is

$$\log \mu_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz}$$

with appropriate constraints and $df = 0$.

This is a saturated model, it describes all possible $\{\mu_{ijk}\}$: $\hat{\mu}_{ijk} = n_{ijk}$

Total No. of parameters $= IJK =$ No. of cells.

$\mu : 1$

$\lambda_i^x : I-1$	$\lambda_j^y : J-1$	$\lambda_k^z : K-1$
$\lambda_{ij}^{xy} : (I-1)(J-1)$	$\lambda_{ik}^{xz} : (I-1)(K-1)$	$\lambda_{jk}^{yz} : (J-1)(K-1)$

$\lambda_{ijk}^{xyz} : (I-1)(J-1)(K-1)$

10.3 Loglinear Models for Three-way Tables

The λ_{ijk}^{XYZ} term in Model 5 refers to the three-factor interaction. It describes how the odds ratio between two variables changes across categories of the third.

Table 10.4 Summary of loglinear models for three-dimensional tables

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

10.3 Loglinear Models for Three-way Tables

10.3.3 Alcohol, Cigarette, and Marijuana Use Example

Example 10.3: In a survey study, 2,276 students are asked whether they had ever used alcohol (A), cigarettes (C), or marijuana (M) in their final year of high school in a nonurban area near Dayton, Ohio.

Table 10.5 Alcohol (A), cigarette (C), and marijuana (M) use for high school seniors.

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

10.3 Loglinear Models for Three-way Tables

Table 10.6 Fitted values for several loglinear models

			Loglinear Model				
A	C	M	(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

Only the fit for model (AC, AM, CM) is close to the observed data (ACM).

10.3 Loglinear Models for Three-way Tables

Table 10.7 illustrates association patterns for these models by presenting estimated odds ratios for their marginal and conditional associations.

Table 10.7 Estimated odds ratios for the loglinear models in Table 10.6.

	Conditional Association			Marginal Association		
Model	A-C	A-M	C-M	A-C	A-M	C-M
(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			

10.3 Loglinear Models for Three-way Tables

For (AM, CM) model, from Table 10.6,

(M = Yes):

A\ C	Yes	No
Yes	909.24	45.76
No	4.76	0.24

(M = No):

A\ C	Yes	No
Yes	438.84	555.16
No	142.16	179.84

$$\hat{\theta}_{AC(1)} = \frac{909.24 \times 0.24}{45.76 \times 4.76} = 1 = \frac{438.84 \times 179.84}{555.16 \times 142.16} = \hat{\theta}_{AC(2)}.$$

The model (AM, CM) implies conditional independence between alcohol use and cigarette use, controlling for marijuana use.

The entry 2.7 for the A-C marginal association for this model is the fitted odds ratio for the marginal A-C fitted table. i.e., For (AM, CM) model:

$$\hat{\theta}_{AC} = \frac{(909.24 + 438.84)(0.24 + 179.84)}{(45.76 + 555.16)(4.76 + 142.16)} = 2.7.$$

10.3 Loglinear Models for Three-way Tables

Model (AC, AM, CM) permits all pairwise associations but maintains homogeneous odds ratios between two variables at each level of the third variable. For (AC, AM, CM) model:

(M = Yes):

A\ C	Yes	No
Yes	910.4	44.6
No	3.6	1.4

(M = No):

A\ C	Yes	No
Yes	538.6	455.4
No	42.4	279.6

$$\hat{\theta}_{AC(1)} = \frac{910.4 \times 1.4}{44.6 \times 3.6} \doteq 7.8 = \frac{538.6 \times 279.6}{455.4 \times 42.4} = \hat{\theta}_{AC(2)}$$

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.

For (AC, AM, CM) model, ignoring the third factor (M), the A-C marginal odds ratio is

$$\hat{\theta}_{AC} = \frac{(910.4 + 538.6)(1.4 + 279.6)}{(44.6 + 455.4)(3.6 + 42.4)} = 17.7.$$

10.3 Loglinear Models for Three-way Tables

For (ACM) model:

(M = Yes):

A\ C	Yes	No
Yes	911	44
No	3	2

(M = No):

A\ C	Yes	No
Yes	538	456
No	43	279

$$\hat{\theta}_{AC(1)} = \frac{911 \times 2}{44 \times 3} = 13.8, \quad \hat{\theta}_{AC(2)} \frac{538 \times 279}{456 \times 43} = 7.7.$$

Table 10.7 shows that estimated conditional odds ratios equal 1.0 for the A-C association in the model (AM, CM). But for that model, the estimated marginal A-C odds ratio differs from 1.0, which reconfirms that conditional independence does not imply marginal independence!

Table 10.7 also shows that estimates of conditional and marginal odds ratios are highly dependent on the model. This highlights the importance of model selection.

10.4 The Loglinear-Logit Connection

Loglinear models for contingency tables do not distinguish between response and explanatory variables. They treat all variables as response variables.

Logit models, by contrast, describe how a binary response depends on a set of explanatory variables.

The model types seem distinct, but strong connections exist between them. A logit model is always corresponding to a loglinear model. For a loglinear model, one can construct logits for one response to help interpret the model. Moreover, logit models with categorical explanatory variables have equivalent loglinear models.

10.4 The Loglinear-Logit Connection

10.4.1 Using logit models to interpret loglinear models

To understand implications of a loglinear model formula, it can help to form a logit that treats one variable as a response and the others as explanatory. We illustrate with the loglinear model of homogeneous association in three-way tables (XY, XZ, YZ):

$$\log \mu_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \qquad (10.2)$$

Suppose Y is binary, and we treat it as a response and X and Z as explanatory. Let π denote the probability that $Y = 1$, which depends on the levels of X and Z. The logit for Y is

$$\begin{aligned}\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) &= \log \frac{P(Y=1 \mid X=i, Z=k)}{P(Y=2 \mid X=i, Z=k)} \\ &= \log\left(\frac{\mu_{i1k}}{\mu_{i2k}}\right) = \log(\mu_{i1k}) - \log(\mu_{i2k}).\end{aligned}$$

10.4 The Loglinear-Logit Connection

Substitution (10.2), we have

$$\operatorname{logit}(\pi) = (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}).$$

The first parenthetical term is a constant, which does not depend on i or k. The second parenthetical term depends on the level i of X. The third parenthetical term depends on the level k of Z. The logit has the additive form

$$\operatorname{logit}(\pi) = \alpha + \beta_i^X + \beta_k^Z \tag{10.3}$$

When Y is binary, loglinear model (10.2) is equivalent to logit model (10.3).

10.4 The Loglinear-Logit Connection

Note: It might seem as if the model (XY, YZ) omitting term λ_{ik}^{XZ} is also equivalent to model (10.3). Indeed, forming the logit on Y for loglinear model (XY, YZ) results in a logit model of the same form. However, the loglinear model that has the same fit as the logit model is the one containing a general interaction term for relationships among the explanatory variables.

10.4 The Loglinear-Logit Connection

10.4.2 Equivalent loglinear and logit models

Table 10.8 summarizes equivalent logit and loglinear models for three-way tables when Y is a binary response.

Table 10.8

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

Loglinear models are most natural when at least two variables are response variables. When only one is response, it is more sensible to use logit models directly.

10.5 Model Selection and Comparison

10.5.1 Testing goodness-of-fit (Model comparison)

Chi-square statistic:

$$X^2 = \sum_{\text{cells}} \frac{(O_i - E_i)^2}{E_i}$$

Likelihood ratio test statistic:

$$G^2 = 2 \sum_{\text{cells}} O_i \log\left(\frac{O_i}{E_i}\right)$$

When n is large and the proposed model is correct, both X^2 and G^2 are chi-square distributed with

$$df = \text{No. of cells} - \text{No. of parameters.}$$

Reject the proposed model if X^2 or G^2 is too large. X^2 and G^2 are asymptotically equivalent.

10.5 Model Selection and Comparison

10.5.2 Model selection

Example 10.4 (Example 10.3 continued): For the student survey, the following table shows results of testing fit for several loglinear models. Models that lack any association term fit poorly. The model (AC, AM, CM) fits well (P-value = 0.54).

Table 10.9

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

10.5 Model Selection and Comparison

Example 10.5 (Automobile Accident Example): Table 10.10 refers to observations of 68,694 passengers in autos and light trucks involved in accidents in the state of Maine in 1991. The table classifies passengers by gender (G), location of accident (L), seat-belt use (S), and injury (I).

Table 6.10 Injury (I) by Gender (G), Location (L), and Seat Belt Use (S).

			Injury		(GLS,GI,IL,IS)	
Gender	Location	Seat Belt	No	Yes	No	Yes
Female	Urban	No	7287	996	7273.2	1009.8
		Yes	11587	759	11632.6	713.4
	Rural	No	3246	973	3254.7	964.3
		Yes	6134	757	6093.5	797.5
Male	Urban	No	10381	812	10358.9	834.1
		Yes	10969	380	10959.2	389.8
	Rural	No	6123	1084	6150.2	1056.8
		Yes	6693	513	6697.6	508.4

10.5 Model Selection and Comparison

Table 10.11 Goodness-of-fit tests for several loglinear models

Model	G^2	df	P-value
(G, I, L, S)	2792.8	11	$<$0.001
(GI, GL, GS, IL, IS, LS)	23.4	5	$<$0.001
(GIL, GIS, GLS, ILS)	1.3	1	0.25
(GIL, GS, IS, LS)	18.6	4	0.001
(GIS, GL, IL, LS)	22.8	4	$<$0.001
(GLS, GI, IL, IS)	7.5	4	0.11
(ILS, GI, GL, GS)	20.6	4	$<$0.001

Model (G, I, L, S) and (GI, GL, GS, IL, IS, LS) have lack of fit (P-value < 0.001). Model (GIL, GIS, GLS, ILS) seems to fit well (P-value = 0.25), but is quite complex and difficult to interpret. This suggests studying models that are more complex than (GI, GL, GS, IL, IS, LS) but simpler than (GIL, GIS, GLS, ILS). Thus, the selecting model is (GLS, GI, IL, IS).