STA 303 / 1002

Note Title 3/19/2012

TA Office hours: Today 2:00-4:00 RW 107 / 109
Friday 2:00-4:00 " "

Loglinear Models for Three-Dimensional Tables

Example: Alcohol, Cigarette & Marijuana Use

1992 survey of high-school seniors in Ohio

Have you ever used alcohol (A), cigarettes (C), marijuana (M)

DATA:

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

Is there a response variable? not a clear choice

Question of interest: Are variables associated?

General setup and notation: for 3-way contingency tables

Have data that are counts in $I \times J \times K$ table

Let π_{ijk} = probability an observation is classified in cell (i, j, k)

Our example $I = J = K = 2$

π_{ijk} denotes joint probability of categorical random variables A, C, M

y_{ijk} = observed count in cell (i, j, k)

$\pi_{i..}$ / $y_{i..}$ = probability observation is in level i of 1st variable / count of observations in

level i of 1st category

$\pi_{ij.}$ = prob an observation is in level i of 1st variable and j of 2nd var.

etc.

$\pi_{...} = 1$

$y_{...} = n$ = total of all counts

y_{ijk} is a realization of a Poisson random variable with mean μ_{ijk}

$\hat{\mu}_{ijk} = n \hat{\pi}_{ijk}$

Hierarchy of Models:

Model 1: Complete Independence Short form: (A, C, M)

H_0: $\pi_{ijk} = \pi_{i..}\,\pi_{.j.}\,\pi_{..k}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$

H_a: $\pi_{ijk} \neq \pi_{i..}\,\pi_{.j.}\,\pi_{..k}$

If independent:

$$\mu_{ijk} = n\,\pi_{ijk} = n\,\pi_{i..}\,\pi_{.j.}\,\pi_{..k}$$

$$\log(\mu_{ijk}) = \log n + \log \pi_{i..} + \log \pi_{.j.} + \log \pi_{..k}$$

additive model

Alternative notation: $\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M$ (Agresti)

(1 = Yes, 2 = No)

Independence Model:

$$\log(\mu_{ijk}) = \beta_0 + \beta_1 I_{[A=1]} + \beta_2 I_{[C=1]} + \beta_3 I_{[M=1]}$$

Find parameter estimates using maximum likelihood estimation, subject to constraint $\sum_k \sum_j \sum_i \hat{\pi}_{ijk} = 1$

$$\Rightarrow \sum_k \sum_j \sum_i \hat{\mu}_{ijk} = n$$

$$= \sum_k \sum_j \sum_i y_{ijk}$$

Can show $\hat{\mu}_{ijk} = \left(n \hat{\pi}_{ijk} = n \hat{\pi}_{i\cdot\cdot} \hat{\pi}_{\cdot j\cdot} \hat{\pi}_{\cdot\cdot k} \right)$

$$= n \frac{y_{i\cdot\cdot}}{n} \frac{y_{\cdot j\cdot}}{n} \frac{y_{\cdot\cdot k}}{n}$$

Estimates of β's make this true

Is the model assuming complete independence adequate?

Compare it to saturated model.

Saturated model includes all possible interactions:

Number of parameters in saturated model:

$$1 + 3 + 3 + 1 = 8$$

1: β_0; 3: coef of $I_{A=1}, I_{C=1}, I_{M=1}$; 3: 2-way interactions; 1: 3-way interaction

In general: $I \times J \times K$ table

$$1 + (I-1) + (J-1) + (K-1) \quad \leq \text{intercept + main effects}$$

$$+ (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) \quad \leq \text{2-way interactions}$$

$$+ (I-1)(J-1)(K-1) \quad \leq \text{3-way interactions}$$

$$= IJK$$

= number of observed counts.

Saturated model fits the data perfectly.

To compare fitted model (complete independence) to saturated model use LRT = Deviance goodness-of-fit test

Example Deviance = 1286.02

Under H_0, this is an observation from a chi-square distribution with $8 - 4 = 4$ df

Very strong evidence that saturated model fits better

Model 2: Block independence

- add a 2-way interaction into the model
- 3 choices for which:

(AC, M) – indicates model has main effect for M, A, C and interaction between A and C

or (AM, C)

or (A, CM)

The presence of a 2-way interaction between 2 variables indicates there is an association between those variables

eg (AC, M) indicates alcohol & cigarette use are associated but they jointly independent of M

So - joint distribution of A, C, M factors into 2 blocks

For (AC, M) model:

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

vs $H_a: \pi_{ijk} \neq \pi_{ij.}\,\pi_{..k}$

$i = 1, .., I$

$j = 1, \ldots, J$

$k = 1, \ldots, K$

Model: (AC, M)

$$\log(\mu_{ijk}) = \beta_0 + \beta_1 I_{A=1} + \beta_2 I_{C=1} + \beta_3 I_{M=1} + \beta_4 I_{A=1} * I_{C=1}$$

Find est using ML estimation:

Can show: $\hat{\mu}_{ijk} \left(= n\hat{\pi}_{ijk} = n\hat{\pi}_{ij.}\hat{\pi}_{..k} \right)$

$$= n \frac{y_{ij.}}{n} \frac{y_{..k}}{n}$$

Is this an adequate model?
Use Deviance G-O-F test to compare to saturated model

Model (class) #3: Partial Independence

Model contains 2 of possible 2-way interactions, so 2 pairs of variables are associated

Possibilities: (AC, AM) *
or (AC, CM)
or (AM, CM)

(AC, AM) indicates alcohol & cigarette use are associated and alcohol and marijuana use are associated but cigarette and marijuana use are conditionally independent, conditional on alcohol use

(Possible Mechanisms: A → C, A → M ; C → A, M → A ; M → A → C, C → A → M)

Conditional independence: $P(CM \mid A) = P(C \mid A)\, P(M \mid A)$

Translate to π's

$$\frac{\pi_{ijk}}{\pi_{i..}} = \frac{\pi_{ij.}}{\pi_{i..}} * \frac{\pi_{i.k}}{\pi_{i..}}$$

$P(A \mid B) \neq \frac{P(A \cap B)}{P(B)}$

or $\pi_{ijk} = \frac{\pi_{ij.}\, \pi_{i.k}}{\pi_{i..}}$

MLEs give: $\hat{\mu}_{ijk} = n\, \hat{\pi}_{ijk} = n\, \frac{\hat{\pi}_{ij.}\, \hat{\pi}_{i.k}}{\hat{\pi}_{i..}}$

(can show)

$$= n\, \frac{y_{ij.}/n \;\; y_{i.k}/n}{y_{i..}/n}$$

$$= \frac{y_{ij} \cdot y_{i \cdot k}}{y_{i \cdot \cdot}}$$

Model 4: Uniform Association
(AC, AM, CM)

- model contains all 2-way interactions
- there is an association between each pair of variables, but no 3-way interaction
  - $\Rightarrow$ the way a pair of variables is associated is the same for all levels of the 3rd variable

For this model, There is no simple interpretation in terms of independence structure ; no nice form for $\hat{\mu}_{ijk}$'s

Need numerical procedures to solve for MLEs

Model 5: Saturated model
(ACM)
- perfectly fits table

Example

Model	G^2 (Deviance)	df	p
A, C, M	1286	4	< .0001
A, CM	534	3	< .0001
AM, C	940	3	< .0001
AC, M	844	3	< .0001
AC, AM	491	2	< .0001
AC, CM	92	2	< .0001
AM, CM	184	2	< .0001
AC, AM, CM	.4	1	.54

SAS doesn't give

Want simplest model that fits data adequately:

Uniform association

Key to loglinear model analysis: examine association between variables (NOT how some variables predict / explain a response)

Example: Uniform Association

Fitted equation:

$$\log(\hat{\mu}_{ijk}) = 5.63 + .049 I_A - 1.89 I_C - 5.31 I_M + 2.05 I_A * I_C + 2.99 I_A * I_M + 2.85 I_C * I_M$$

Example calculation of a predicted value:

$\hat{\mu}_{111}$ $(A = \text{Yes}, C = \text{Yes}, M = \text{Yes}, I_A = 1, I_C = 1, I_M = 1)$

$= \exp\{5.63 + .049 - 1.89 + \ldots + 2.85\} = 910.38$

$\hat{\mu}_{222}$ $(A = \text{No}, C = \text{No}, M = \text{No}, I_A = I_C = I_M = 0)$

$= \exp\{5.63\}$

Interpreting $\hat{\beta}$'s – use them to calculate odds

$\hat{\mu}_{ijk} = n\hat{\pi}_{ijk}$

$\frac{\hat{\pi}_{ij1}}{\hat{\pi}_{ij2}}$ gives odds of marijuana use for alcohol & cigarette use = i, j

$= \frac{\hat{\mu}_{ij1}}{\hat{\mu}_{ij2}}$

Example For students use alcohol & cigarettes, estimated odds of marijuana use are $\frac{910.38}{538.62}$ ← from SAS, predicted count for A=1, C=1, M=2

$= 1.7$

For students who used neither alcohol or cigarette,

estimated odds of marijuana use

1.38

279.62