STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

March 15, 2018

Class 18- Summary of Case Study VI

Framingham Heart Study
A Project of the National Heart, Lung, and Blood Institute and Boston University

2 ×2 CT

(2 factors) I×J CT

I,J > 2

(3 factors) I×J×K CT

Eg. 2×2×2

I

	Diff. in prop	LRT	Log-linear
Assume	Row totals fixed	Overall total fixed	Totals are random
Dist. of Y	Binomial	Multinomial	Poisson
H_0	$\pi_1 = \pi_2$	$\pi_{ij} = \pi_{i.}\pi_{.j}$	Additive model
Test Stat.	Z	$\chi^2_{(I-1)(J-1)}$	$\chi^2_{(I-1)(J-1)}$

Non-parametric method : Fisher's Exact Test.

table (f1, f2, f3) ← dim (f3) 2-way tables.

A Three-way Contingency Table

Case Study VII Data:

- 1992 survey of high-school seniors in Ohio
- Table of counts of seniors who used alcohol, cigarettes and marijuana.

(3 factors).

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

Q: Are alcohol (A), cigarettes (C) and marijuana (M) use associated?

Forms of independence in $I \times J \times K$ Tables

Models

	Independence		π_{ijk}	Short form	
	Mutually indep.	(1)	$\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$	(X,Y,Z)	Complete
1 2way	Jointly indep.	(3)	$\pi_{ijk} = \pi_{ij.}\pi_{..k}$	($\boxed{XY}$,Z)	Block
2 2-way	Conditionally indep.	(3)	$\pi_{ijk} = \pi_{i.k}\pi_{.jk}/\pi_{..k}$	(XZ,YZ)	Partial
All 2-way	Uniform assoc.	(1)	$\pi_{ijk} = \pi_{ij.}\pi_{i.k}\pi_{.jk}$	(XZ,YZ, XY)	Homo
All 2-way & 3-way	Saturated	(1)	π_{ijk}	XYZ	

Eg

(AC, M)
(AM, C)
(CM, A)

(XY, XZ)
(YZ, XY)

Three-way Tables

$$\log(\mu_{ijk}) = \beta_0 + \beta_1 1_A + \beta_2 1_C + \beta_3 1_M$$

$$\cdots 1_A 1_C$$

- Learning Objectives
  - Write out the models used and the assumptions for inference
  - Carry out the inference procedures completely
  - Interpret the respective R outputs

Wald
G-O-F
Global LRT

$$\cdots 1_A 1_C 1_M$$

Model 1: Complete Independence

- $P(ACM) = P(A)P(C)P(M)$; Alcohol, cigarette and marijuana use are mutually independent
- Hypotheses:

$$H_0 : \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k} \text{ for all } i, j, k$$
$$H_a : \pi_{ijk} \neq \pi_{i..}\pi_{.j.}\pi_{..k}$$

- Short form: (A,C,M) -all 3 main effects only
- $I = J = K = 2$

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M}$$

Additive

where $\mathbf{I} : \{1 = Yes, 0 = No\}$

Model 1: Complete Independence

- In general, we have the constraint $n = \sum_i \sum_j \sum_k y_{ijk}$ or $\sum_i \sum_j \sum_k \hat{\pi}_{ijk} = 1$. Then by ML estimation,

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

$$\Longrightarrow \hat{\pi}_{ijk} = \frac{y_{ijk}}{n} \text{ or } \hat{\mu}_{ijk} = y_{ijk}$$

LRT

(Deviance G-o-F)

H_0: Fitted

H_a: Saturated

- For complete independence model, using an additional $(I-1)+(J-1)+(K-1)$ constraints

$$\begin{aligned}\hat{\mu}_{ijk} = n\hat{\pi}_{ijk} &= n\hat{\pi}_{i\cdot\cdot}\hat{\pi}_{\cdot j\cdot}\hat{\pi}_{\cdot\cdot k} \\ &= n\frac{y_{i\cdot\cdot}}{n}\frac{y_{\cdot j\cdot}}{n}\frac{y_{\cdot\cdot k}}{n}\end{aligned}$$

Model Class 2: Block Independence

- $P(AC|M) = P(AC)$; Joint probability of alcohol and cigarette use is independent of marijuana use; Alcohol and cigarette use are associated
- Hypotheses:

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

- Short form: (AC,M) - all 3 main effects and 1 interaction

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AC}}$$

1 2-way interaction term.

where $\mathbf{I}_{AC} = \mathbf{I}_A * \mathbf{I}_C$

- Others in this class: (AM, C), (CM, A)

Model 2: Block Independence

- By ML estimation, for block independence model

$$
\begin{aligned}
\hat{\mu}_{ijk} = n\hat{\pi}_{ijk} &= n\boxed{\hat{\pi}_{ij\cdot}}\hat{\pi}_{\cdot\cdot k} \\
&= n\frac{y_{ij\cdot}}{n}\frac{y_{\cdot\cdot k}}{n}
\end{aligned}
$$

Model Class 3: Partial Independence

- $P(AC|M) = P(A|M)P(C|M)$; Alcohol and cigarette use are conditionally independent given marijuana use; Alcohol and marijuana use are associated, and cigarette and marijuana use are associated
- Hypotheses:

$$H_0 : \pi_{ijk} = \pi_{i \cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}$$
$$H_a : \pi_{ijk} \neq \pi_{i \cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}$$

- Short form: (AM,CM) - all 3 main effects and 2 interactions

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{CM}}$$

2 2-way interaction terms.

- Others in this class: (AC, CM), (AC, AM)

Model 3: Partial Independence

- We have $P(AC|M) = P(A|M)P(C|M)$.

$$\Longrightarrow \frac{\pi_{ijk}}{\pi_{\cdot\cdot k}} = \frac{\pi_{\cdot jk}}{\pi_{\cdot\cdot k}} \frac{\pi_{i\cdot k}}{\pi_{\cdot\cdot k}}$$

$$\text{or } \pi_{ijk} = \frac{\pi_{\cdot jk}\pi_{i\cdot k}}{\pi_{\cdot\cdot k}}$$

- Then by ML estimation

$$\begin{aligned}\hat{\mu}_{ijk} = n\hat{\pi}_{ijk} &= n\frac{\hat{\pi}_{\cdot jk}\hat{\pi}_{i\cdot k}}{\pi_{\cdot\cdot k}} \\ &= n\frac{(y_{\cdot jk}/n)(y_{i\cdot k}/n)}{(y_{\cdot\cdot k}/n)} \\ &= \frac{y_{\cdot jk}y_{i\cdot k}}{y_{\cdot\cdot k}}\end{aligned}$$

Model 4: Uniform association

- There is an association among all pairs
- For all levels of the 3rd variable, the association between the pair is the same
- Short form: (AM,AC,CM) - all 3 main effects and 3 two-way interactions but no three-way interaction

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{AC} + \beta_6 \mathbf{I}_{CM}}$$

All 2-way interaction terms.

- Solutions for π_{ijk} (μ_{ijk}) are found numerically with no simple expression in terms of y_{ijk}'s
- No simple interpretation ito. independence structure

Saturated Model

- Total number of parameters:

$$1 + \underbrace{3}_{\text{1-way}} + \underbrace{3}_{\text{2-way}} + \underbrace{1}_{\text{3-way}} = \boxed{8}$$

- Total number of observed counts:

$$\begin{aligned} &1 + (I-1) + (J-1) + (K-1) \\ &\quad + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) \\ &\quad + (I-1)(J-1)(K-1) = IJK = 2*2*2 = 8 \end{aligned}$$

$$\boxed{\begin{aligned} \log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M \\ + \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{AC} + \beta_6 \mathbf{I}_{CM} + \beta_7 \mathbf{I}_{ACM} \end{aligned}}$$

- Saturated model always fits the data perfectly

On the Saturated Model

$$\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{AC} + \beta_6 \mathbf{I}_{CM} + \beta_7 \mathbf{I}_{ACM}$$

- Total # of parameters=Total # of observed counts
- Has a separate parameter for each observation
- Always gives a perfect fit
- Explains all the variation by its systematic component
- Sounds good but not a helpful model
- Does not smooth the data or is not parsimonious
- Serves as a baseline for checking model fit

Deviance G-O-F

H_0: Fitted (Reduced)

H_a: Saturated (Full).

T.S = $\text{Deviance}_R - \text{Deviance}_S$ → 0

Results from R output

Use R codes provided to show these → Add AIC statistics to the table

Model	df	G^2=Deviance	p-value
(A,C,M)	4	1286.02	< 0.0001
(AC,M)	3	843.83	< 0.0001
(AM, C)	3	939.56	<0.0001
(A,CM)	3	534.21	<0.0001
(AC,AM)	2	497.37	<0.0001
(AC,CM)	2	92.02	<0.0001
(AM,CM)	2	187.75	<0.0001
(AC,AM,CM)	1	0.37	0.5408
(ACM)	0	0.00	-

fitted model is not adequate

Adequate.

The simplest model that fits the data adequately is the "Uniform Association" model (AC,AM,CM).

Class 18 Summary

- Three-way contingency tables:
  - Log-linear model approach
  - Types of independence or association/ interactions
    - (i) Complete
    - (ii) Block
    - (iii) Partial
    - (iv) Uniform association
    - (v) 3-way interaction
  - Deviance goodness-of-fit test
- Next Class: Model Diagnostics

Interpretation, Estimate

- Things to do:
  - Assignment #3
  - Participation 6
  - Participation 7
  - Practice Problems on Poisson Regression (Log-linear models)

Final (Apr. 25)