STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

March 20, 2018

Three-way Tables

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

Results from R output

Test C.I

Model	df	G^2=Deviance	p-value	AIC
(A,C,M)	4	1286.02	< 0.0001	1343.06
(AC,M)	3	843.83	< 0.0001	902.87
(AM, C)	3	939.56	<0.0001	
(A,CM)	3	534.21	<0.0001	
(AC,AM)	2	497.37	<0.0001	558.41
(AC,CM)	2	92.02	<0.0001	
(AM,CM)	2	187.75	<0.0001	
(AC,AM,CM)	1	0.37	0.5408	63.42
(ACM)	0	0.00	-	65.04

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

$I,J=2$

$I,J>2$

$I,J,k=2$

$I,J,k>2$

Exercise: Fitted values and Interpretations

Q: Complete the fitted equation and prove the fitted values ~~above~~ below

Fitted equation:

$$\log(\hat{\mu}_{ijk}) = 6.81 - 5.53 \mathbf{I}_{A_2} - 3.02\,1_{C2} - 0.52\,1_{M2} + 2.98\,1_{A2} \cdot 1_{M2} + 2.05\,1_{A2} \cdot 1_{C2} + 2.85\,1_{C2} \cdot 1_{M2}$$

Some fitted values, $\hat{\mu}_{ijk}$: $\hat{\mu}_{ijk} = e^{\hat{\beta}_0 + \hat{\beta}_1 1_{A2} + \cdots}$

				Log-linear models	
	A use	C use	M use	(AC, AM, CM,)	(ACM)
$\hat{\mu}_{111}$	Yes, 1	Yes, 1	Yes, 1	910.4 $= e^{6.81}$	911
$\hat{\mu}_{112}$	Yes, 1	Yes, 1	No, 2	538.6	538
$\hat{\mu}_{121}$	Yes, 1	No, 2	Yes, 1	44.6	44
$\hat{\mu}_{122}$	Yes, 1	No, 2	No, 2	$e^{6.81-3.02-0.52+2.85} = 455.38$	456

$\boxed{\hat{\mu}_{ijk} \equiv y_{ijk}}$

Observed counts.

$$1_{A_2} = \begin{cases} 1 & \text{if Alcohol was NOT used} \\ 0 & \text{if '' was used} \end{cases}$$

Estimation

Inference: S.I, Test

Fitted values and Interpretations

- **Use estimates of β's to calculate odds.**
  - Eg, the odds of marijuana use for alcohol and cigarette use at (i, j) are:

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

$$\frac{\hat{\pi}}{1-\hat{\pi}}$$

- *Example 1*: For students who use alcohol and cigarettes, the estimated odds of using marijuana are:

$(A = 1 = C = M) \longrightarrow$ $\hat{\mu}_{111} = 910.38$

$(A = 1 = C, M = 2) \longrightarrow$ $\hat{\mu}_{112} = 538.61$

Odds of Marijuana Use,

$$\frac{\hat{\mu}_{111}}{\hat{\mu}_{112}} = \frac{910.38}{538.61} > 1$$

$$= 1.69$$

- *Example 2*: For students who use neither alcohol nor cigarettes, the estimated odds of using marijuana are:

$(A = 2 = C, M = 1) \longrightarrow$ $\hat{\mu}_{221} = 1.38$

$(A = 2 = C = M) \longrightarrow$ $\hat{\mu}_{222} = 279.62$

$$\text{Odds of Marij. Use} = \frac{1.38}{279.62} \ll 1$$

$$= 0.005$$

Inference for Log-linear models

- Q: What procedures do we use to:
  - Estimate parameters in log-linear models?
    - A: Maximum likelihood estimation
  - Carry out inference (significance tests and C.I.s)?
    - A: Wald tests and C.I.s, and LRT

(R) (F)
Global (Null vs Fitted)
LRT (Reduced vs Full).
Deviance G-o-F.
(Fitted vs Saturated)
(R) (F)

What are the conditions for inference to be valid?

1. Independent quantities being counted
2. Large enough sample sizes for MLE asymptotic tests to hold.
   - RULE-OF-THUMB: (Most) $\widehat{\mu}_{ijk} \geq 5$ for all i, j, k.
3. Cross-classified counts follow a Poisson distribution, i.e., $\boxed{\text{Var}(y_{ijk}) = \mu_{ijk}}$.
   - If not, then the deviance is very large ("extra-Poisson" variation).
   - Deviance/df should be about 1.
4. Correct form of the model / Model fits the data.
   - $\log(E(Y))$ is linear in the β's
   - All relevant variables included.
   - No outliers
   - Agreement of predicted and observed counts
   - Check deviance goodness-of-fit test

What is the frame of a Likelihood Ratio Test?

Global

Deviance G-o-F

- Idea: Compare likelihood of data under FULL (F) model, $\mathcal{L}_F$ to likelihood under REDUCED (R) model, $\mathcal{L}_R$ of same data.

$$\text{Likelihood ratio}: \frac{\mathcal{L}_R}{\mathcal{L}_F}, \quad \text{where } \mathcal{L}_R \leq \mathcal{L}_F$$

- Hypotheses: $H_0 : \beta_1 = \cdots = \beta_k = 0$

  (Reduced model is appropriate; fits data as well as Full model)

  H_a: at least one $\beta_1, \cdots, \beta_k \neq 0$

  (Full model is better)

- Test Statistic: $G^2 = -2\log\mathcal{L}_R - (-2\log\mathcal{L}_F) = -2\log\left(\frac{\mathcal{L}_R}{\mathcal{L}_F}\right)$

- For large n, under H_0, G^2 is an observation from a Chi-square distribution with k df.

Comparing models

- LRTs for models with and without set of indicator variables for effect of interest
- Particularly useful if > 2 levels in categorical explanatory variables
- *Example*: Suppose we have a $\underline{2} \times \underline{2} \times \underline{3}$ table and we fit the Uniform association model (XY, XZ, YZ)

Eg, ACM Uniform — used 6 indicator variables for the factors

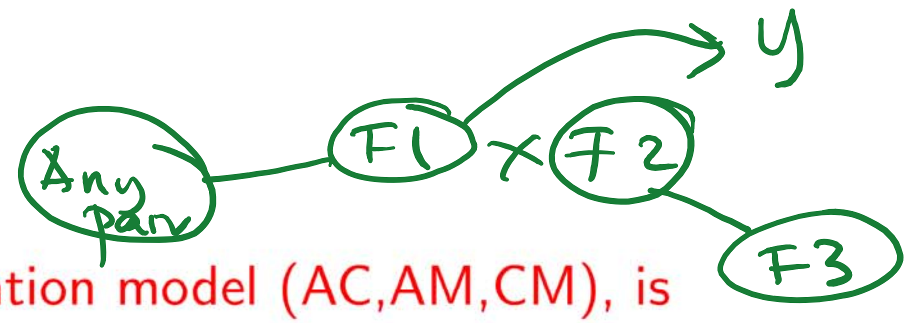
$$\begin{aligned}\log \mu_{ijk} = &\beta_0 + \beta_1 \mathbf{I}_{X=1} + \beta_2 \mathbf{I}_{Y=1} + \beta_3 \mathbf{I}_{\underline{Z=1}} + \beta_4 \mathbf{I}_{\underline{Z=2}} \\ &+ \beta_5 \mathbf{I}_{X=1} * \mathbf{I}_{Y=1} + \underbrace{\beta_6 \mathbf{I}_{X=1} * \mathbf{I}_{Z=1} + \beta_7 \mathbf{I}_{X=1} * \mathbf{I}_{Z=2}} \\ &+ \underbrace{\beta_8 \mathbf{I}_{Y=1} * \mathbf{I}_{Z=1} + \beta_9 \mathbf{I}_{Y=1} * \mathbf{I}_{Z=2}}\end{aligned}$$

$\rightarrow 4+6=10$ indicator variables.

Is the <u>YZ</u> interaction needed?

H_0 : $\underline{\beta_8 = \beta_9} = 0$ vs H_a: at least 1 of β_8, β_9 is not 0

Comparing models



- *Exercise*: For the Uniform association model (AC,AM,CM), is the CM interaction needed in the Uniform association model?/ Does the (AC, AM) model fit just as well?

$$\log \mu_{ijk} = \beta_0 + \beta_1 \mathbf{I}_{A=1} + \beta_2 \mathbf{I}_{C=1} + \beta_3 \mathbf{I}_{M=1} + \beta_{\cancel{5}\,4} \mathbf{I}_{A=1} * \mathbf{I}_{C=1} + \beta_{\cancel{6}\,5} \mathbf{I}_{A=1} * \mathbf{I}_{M=1} + \beta_{\cancel{7}\,6} \mathbf{I}_{C=1} * \mathbf{I}_{M=1}$$

- A: $H_0: \beta_6 = 0$ (AC, AM model fits as well)
  $H_a: \beta_6 \neq 0$ (Uniform asc. model fits better).

T.S :	Wald	LRT
Distn :	$(17.382)^2 = 302.13$	$G^2 = 497.37 - 0.37 = 497 \sim \chi^2_1$
p-value :	$p < 0.0001$	$p \approx 0$ or $p < 0.0001$
Cnc :	All pairs	of factors are associated with each other

What is the Deviance G-O-F test?

- Uses LRT: Compares
  - Model of Interest (REDUCED, R) model to
  - Saturated Model (FULL, F) model.
- Sometimes called "Drop-in-Deviance" test.

- Hypotheses:

  H_0: (Fitted model fits data as well as Saturated model)

  H_a: (Saturated model is better)
- Test Statistic:

$$Deviance = -2\log\left(\frac{\mathcal{L}_R}{\mathcal{L}_F}\right) = -2\log\left(\frac{\mathcal{L}_M}{\mathcal{L}_S}\right)$$

- Under H_0, *Deviance* is an observation from a chi-square distribution with $df = \#parameters(S) - \#parameters(M)$.

Deviance G-O-F statistic for 3-way tables

- The joint distribution of cell counts is

$$P(\mathbf{Y}=\mathbf{y}) = \prod_k \prod_j \prod_i \frac{\mu_{ijk}^{y_{ijk}} e^{-\mu_{ijk}}}{y_{ijk}!}$$

- Log-likelihood function:

$$\log \mathcal{L} = \sum_k \sum_j \sum_i (y_{ijk} \log \mu_{ijk} - \mu_{ijk} - \log y_{ijk}!)$$

- Likelihood ratio statistic: (Practice Question)

$$Deviance = 2 \sum_k \sum_j \sum_i y_{ijk} \log\left(\frac{y_{ijk}}{\hat{\mu}_{ijk}}\right)$$

Saturated

Fitted model.

- Hints: Under the saturated model, $\hat{\mu}_{ijk} = y_{ijk}$; $\sum_k \sum_j \sum_i y_{ijk} = n$

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

$\hat{\mu}_i = e^{\hat{\mu}}$ (log-linear)

$\hat{\pi}_i = \frac{e^{\hat{\mu}}}{1+e^{\hat{\mu}}}$ (logistic)

How to interpret Deviance?

- Is the form of the fitted model adequate or do I need something more complicated?
- Compares fitted model to saturated model
- Small deviance / Large p-values implies:
  - Fitted model is adequate, OR
  - Test is not powerful enough to detect inadequacies
- Large deviance / Small p-values implies:
  - Fitted model is not adequate; consider a more complex model OR
  - Underlying distribution is not adequately modelled by the Poisson distribution / Poisson model not correct / $Var(y_{ijk}) > \mu_{ijk}$ OR
  - There are severe outliers in the data

Are there outliers?

- Check residuals

1. Raw residual: $y_{ijk} - \hat{\mu}_{ijk}$
2. Pearson residual: sum of squares gives Pearson chi-square test statistic
$$\frac{y_{ijk} - \hat{\mu}_{ijk}}{\sqrt{\hat{\mu}_{ijk}}}$$
easy to interpret
3. Deviance residual: sum of the squares is the Deviance
$$\text{sign}(y_{ijk} - \hat{\mu}_{ijk})\sqrt{2\left\{y_{ijk}\log\left(\frac{y_{ijk}}{\hat{\mu}_{ijk}}\right) - y_{ijk} + \hat{\mu}_{ijk}\right\}}$$

Pearson and Deviance residuals

- Easier to interpret: Pearson
- More reliable: Deviance
- Usually similar? Yes
- Differences are more prominent when used to compare models
- If Poisson means are large, the sampling distributions are ... Approx Normal
- Rule-of-thumb: Outlier if Pearson or Deviance residual > 3 (if sample size is small, consider those > 2)

Presence of "Extra-Poisson Variation"

- Check if $\frac{Deviance}{df} > 1$

- Q: How much > 1 is important?
- A: If Deviance GOF test is statistically significant.

- If other problems are ruled out, then include a dispersion parameter in the model, i.e.,

$$\text{Var}(Y_{ijk}) = \psi\mu_{ijk}$$

quasi

- OR use Negative Binomial regression

$$Var(Y_{ijk}) = \mu_{ijk}(1 + \psi\mu_{ijk})$$

(Agresti, Chp. 14)

Summary of models

	OLS	Logistic	Log-linear
Link	Identity	Logit	Log
Regression $\mu\{Y\|\mathbf{X}\}$ is	Linear linear in β's	Non-linear not linear in β's	Non-linear not linear in β's
Models	Mean of Y	Log odds	Log of means
Natural response Response is	Yes Normal	Yes Binomial	No Poisson
Indep. Obs.	Yes	Yes	Yes
$Var(Y_i\|\mathbf{X}) =$	σ^2	$\pi_i(1-\pi_i)$	μ_i
	(constant	(changes with i)	

Class 19 Summary

- Log-linear models for three-way contingency tables:
  - Assumptions
  - LRT: Deviance goodness-of-fit test
  - Using fitted equation to find odds
  - Model diagnostics
- Next Class: Mixed Models
- Things to do:
  - Assignment #3
  - Participation 6
  - Practice Problems on Poisson Regression (Log-linear models)