STOR 556: ADV METH DATA ANAL
Instructor: Richard L. Smith

Class Notes #16:
March 5, 2019

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Seeking Volunteers...

- UNC Science Expo
- April 6, 11:00 am to 4:00 pm, Morehead Planetarium
- STOR booth, organized by Dr. Olvera-Cravioto and Dr. Nobel
- They would like three or four STOR majors to help with the booth!
- If interested, please email or talk to me and I'll pass on the message

Homework 5

- Chapter 5, Problems 2 and 5 (pages 99–101)
- Due date: Tuesday March 5
- I have posted the TA's solutions to HW1–4. Please treat these as *for personal use only* and do not pass on to anyone outside the class!

Comments on Midterm

- Full solutions now available on “Resources” page on sakai (includes comments on grading on the last page)
- Also an updated version of the exam (includes grades scheme)
- Summary of results:
  - Mean 86.3, median 88, first and third quartiles were 84 and 92
  - 63 students scored 80 or better
  - If your score was below 70, please arrange an appointment with the instructor

Mantel-Haenszel Test

- Objective: test independence of 2×2 tables across K strata
- Data written $\{y_{ijk},\ i = 1, 2,\ j = 1, 2,\ k = 1, ..., K\}$
- Test is conditional on marginal totals in each table — therefore, it suffices to base the test on the values of $y_{11k},\ k = 1, ..., K$
- $T = \frac{(|\sum_k y_{11k} - \sum_k \mathsf{E}(y_{11k})| - 1/2)^2}{\sum_k \mathsf{Var}(y_{11k})}$
- Expectation and variance are computed under null hypothesis of independence in each table, given the marginal totals
- Test statistic T is approximately χ^2_1 for large samples; exact p-value calculation is possible for small datasets

Independence Model

- Assume 3-way table with cell probabilities p_{ijk}.
- Mutual independence: $p_{ijk} = p_i p_j p_k$.
- If n total observations, $\mathsf{E}(Y_{ijk}) = np_{ijk}$

$$\log \mathsf{E}(y_{ijk}) = \log n + \log p_i + \log p_j + \log p_k$$

- Fit as a glm with main effects only
- For smoking dataset, implies independence of all three variables, which is implausible

Joint Independence Model

- $p_{ijk} = p_{ij}p_k.$

- If n total observations, $\mathrm{E}(Y_{ijk}) = np_{ijk}$

$$\log \mathrm{E}(y_{ijk}) \;=\; \log n + \log p_{ij} + \log p_k$$

- In smoking example, allows for smoking and death status to be dependent, but only if they are independent of age — unlikely

- Doesn't fit the data

Conditional Independence Model

- Let $p_{ij|k}$ be the probability that an observation falls in the (i, j) cell conditional that the third variable is k
- Conditional independence assumption is

$$p_{ij|k} = p_{i|k} p_{j|k}$$

- Equivalent to

$$p_{ijk} = \frac{p_{ik} p_{jk}}{p_k}.$$

- Model is then

$$\log \mathsf{E}(y_{ijk}) = \log n + \log p_{ik} + \log p_{jk} - \log p_k$$

- The text implies this model could plausibly fit the data but I think this is wrong — explanation to follow

Uniform Association Model

- Model is then

$$\begin{aligned}\log \mathsf{E}(y_{ijk}) &= \log n + \log p_i + \log p_j + \log p_k \\ &\quad + \log p_{ij} + \log p_{ik} + \log p_{jk}\end{aligned}$$

- No three-way association, not saturated

- Odds ratio the same for every group (but doesn't have to be 1)

- Odds ratio for k'th group is

$$\frac{\mathsf{E}(Y_{11k})\mathsf{E}(Y_{22k})}{\mathsf{E}(Y_{12k})\mathsf{E}(Y_{21k})}$$

- This model does appear to fit the data — implies smoking–death interaction within each age group

Comparison Between Conditional Independence and Uniform Association Models

- The text doesn't note that the C.I. model is nested inside the U.A. model — the latter has a model term smoker:death which is not present in the C.I. model
- Therefore, we can do an anova test of one against the other:

```
> anova(modc,modu,test='Chi')
Analysis of Deviance Table

Model 1: y ~ smoker * age + age * dead
Model 2: y ~ (smoker + age + dead)^2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         7     8.3269
2         6     2.3809  1    5.946  0.01475 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Conclude U.A. is a statistically significant better fit