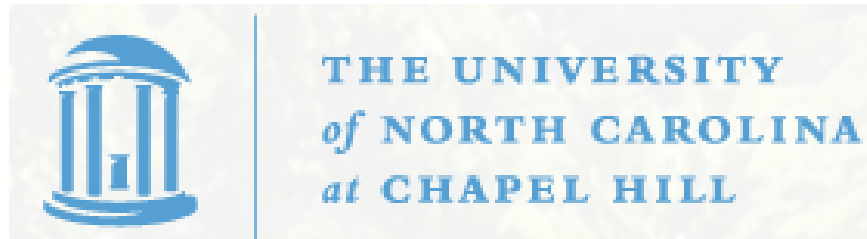


STOR 556: ADV METH DATA ANAL

Instructor: Richard L. Smith

Class Notes #17:

March 7, 2019



Homework 6

- Chapter 6, Problems 2 and 4 (pages 126–127). You can omit part (e) of question 2.
- Due date: Tuesday March 19

Uniform Association Model

- Model is then

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

- 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

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

- 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

Saturated Model

- Same as U.A. model *plus* three-way interactions
- Model is then

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

- Allows different odds ratios in different groups
- Can drop three-way interaction — then reverts to U.A. model

Binomial Model

- Treat one variable as the response, e.g. “alive” or “dead”
- View as a binomial distribution within each smoker/age group
- Most general model allows for interaction between smoker and age
- We can drop interaction but still see marginal effects due to smoker and age
- This is actually equivalent to the U.A. model — reason isn't obvious, but it's confirmed by the deviance
- The text also discussed the “null model for Binomial GLM” but this doesn't fit the data

Conclusion for Smoking Dataset

- Smoking is associated with increased mortality after adjusting for age
- Three different tests lead to this conclusion:
 - Mantel-Haenszel
 - Uniform Association Model
 - Binomial response model
- I think the conditional independence model is misleading — the uniform association model is a better fit, and confirms the smoking–mortality interaction

Ordinal Data

- Sometimes, data are categorical in the sense that they do not correspond to numerical values, but there is still a natural ordering to the categories
- *Ordinal data* techniques take advantage of the ordering
- Linear association model of form

$$\log EY_{ij} = \log n + \alpha_i + \beta_j + \gamma u_i v_j$$

where u_i and v_j are predetermined numerical ordering variables

- Test of $\gamma = 0$ is a test of association between the ordered variables

Application to Voting Trends Dataset

- Educational level and party affiliation (two variables part of a much larger dataset)
- Each measured on a 7-point scale
- Analysis as a two-way table does not indicate dependence
- But, maybe we can get better information by exploiting the natural ordering of both variables

Recoding as a Mixed Factor-Numerical Dataset

- For marginal effects, keep both PID and educ as factor variables
- For the interaction term, recode both variables as numerical on a scale of 1–7 using `unclass`
- Reduces interaction to a single variable γ and this *is* significant

Estimate	Std. Error	z value	Pr(> z)
0.028744615	0.009061742	3.172084969	0.001513487

- Conclusion: Higher education level is associated with increased support for Republicans
- Some suggestion there's a pattern in the residuals (I'm not convinced of this)

Alternative Models

- Alternative numerical codings (not necessarily 1,2,...,7) — makes slight difference to numerical results, not to overall conclusion
- Mixed factor-numerical analysis (factor for education, numerical for political affiliation)
- Alternative: recode education level into two classes (below HS or HS grad — only for interaction term not marginal distribution)
- This model has the best deviance but may be due to “data snooping”

Conclusions

- If we use a numerical 1–7 scale for both variables and then look for interactions, there *is* a statistically significant effect — indicates higher-educated people are more likely to support Republicans
- Alternative numerical codings are possible but don't make much difference
- A mixed factor-numerical scale for the interactions doesn't improve on this (my interpretation)
- The analysis does not account for gender/race/age or geographic variables — *possible Simpson bias here?*