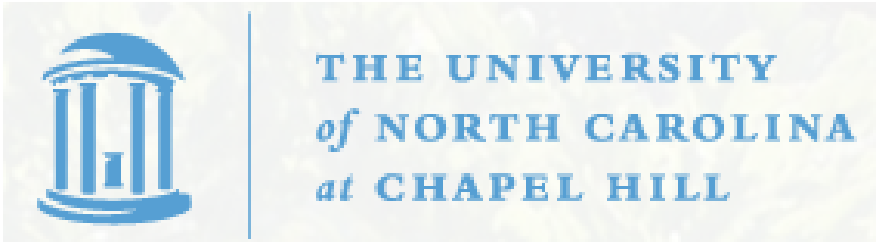
STOR 556: ADV METH DATA ANAL
Instructor: Richard L. Smith

Class Notes #12:

February 19, 2019



1

Scheduling a Take-home Midterm/Final

- Midterm, posted noon Feb 24, email solutions no later than 6pm Feb 25
- Final, posted noon Apr 30, email solutions no later than 6pm May 1
- Dates are confirmed but will I work with any individual students who have difficulties with those dates

Homework 4

- Chapter 3, Problems 1 and 3
- Hint for problem 1: you can test for interactions by including terms like

  glm(cbind(ncases,ncontrols)~agegp+alcgp+tobgp+agegp*alcgp +agegp*tobgp+alcgp*tobgp,family=binomial,esoph)

  The * terms denote interactions between factor variables.

  Part (c) is open-ended: try to find some model that fits better than the best model from (b)
- Problem 3: data(seeds)
- In both problems, also answer part (i): would the fit be improved by using a quasi-binomial model?
- Due date: Tuesday, February 19.

Homework 5

- Chapter 5, Problems 2 and 5 (pages 99–101)
- Due date: To be changed

Rate Models

- Example: count cancer cases in Raleigh, Durham and Chapel Hill
- Expect cases(Raleigh) $>$ cases(Durham) $>$ cases(Chapel Hill)
- However that just reflects differences in population — a more reasonable hypothesis would be that the *rate* of cancer is the same in each of the three cities
- Model as

$$\text{Number of cases} \approx \text{Population of city} \times \text{Rate}$$

- $\mathrm{E}(\log y_i) \approx \log N_i + \sum_{j=1}^{q} x_{ij}\beta_j$ where N_i is population of ith city
- Like a GLM with one parameter constrained to be 1
- Can fit that using an *offset*

Use of Offset in R

"dicentric" example

```
rmod=glm(ca∼offset(log(cells))+log(doserate)*dosef,
family=poisson,dicentric)
```

Negative Binomial Model

- Number of Bernoulli trials needed to get k'th success
- $\Pr\{Z = z\} = \binom{z-1}{k-1} p^k (1-p)^{z-k}$.
- Alternative: $Y = Z - k$, $p = \frac{1}{1+\alpha}$ so $\Pr\{Y = y\} = \binom{y+k-1}{k-1} \frac{\alpha^y}{(1+\alpha)^{y+k}}$.
- $\mathrm{E}(Y) = \mu = k\alpha$ and $\mathrm{Var}(Y) = k\alpha + k\alpha^2 = \mu + \frac{\mu^2}{k}$.
- Log likelihood is

$$\ell = \sum_{i=1}^{n} \left(y_i \log \frac{\alpha}{1+\alpha} - k \log(1+\alpha) + \sum_{j=0}^{y_i - 1} \log(j+k) - \log(y_i!) \right).$$

- $\eta = \sum_{j=0}^{p} x_{ij}\beta_j = \log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{\mu+k}$

Fitting in R

Venables-Ripley method with k fixed:

```
library(MASS) modn=glm(skips .,negative.binomial(k),solder)
```

Alternative: determine k by maximum likelihood

```
modn=glm.nb(skips .,solder)
```

Zero-inflated counts models

- First load package "pscl"
- Hurdle model:

$$\begin{aligned} P(Y=0) &= f_1(0), \\ P(Y=j) &= \frac{1-f_1(0)}{1-f_2(0)} f_2(j),\ j>0, \end{aligned}$$

  where (by default) f_2 is Poisson. Fit in R: `hurdle(y~.,data=dataframe)`

- ZIP model:

$$\begin{aligned} P(Y=0) &= \phi + (1-\phi) f(0), \\ P(Y=j) &= (1-\phi) f(j),\ j>0. \end{aligned}$$

  Fit in R: `zeroinfl(y~.,data=...)`

- Possibility of using different covariates for the two components, e.g. `zerolinf(y~x1+x2+x3|x4+x5,data=...)`