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

Class Notes #11:
February 14, 2019

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

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.

CHAPTER 5:
REGRESSION FOR COUNT DATA

1. Poisson Regression

Basics of Poisson model

- $\Pr\{Y = y\} \; = \; \frac{\mu^y e^{-\mu}}{y!}, \;\; y = 0, 1, 2, ...$
- Data: $y_1, ..., y_n$ Poisson with mean $\mu_1, ..., \mu_n$
- Log link: $\log \mu_i = \eta_i = \sum_j x_{ij}\beta_j$
- Log likelihood $\ell(\mu_1, ..., \mu_n) = \sum(y_i \log \mu_i - \mu_i - \log y_i!)$
- Unrestricted μ_i: maximized when $\mu_i = y_i$. Call this ℓ_1.
- With log link and regressors:

$$\begin{aligned}
\ell(\beta) &= \sum_i \left\{ y_i \sum_j x_{ij}\beta_j - \exp\left(\sum_j x_{ij}\beta_j\right) - \log(y_i!) \right\}, \\
\frac{\partial \ell(\beta)}{\partial \beta_k} &= \sum_i \left\{ y_i x_{ik} - x_{ik} \exp\left(\sum_j x_{ij}\beta_j\right) \right\}.
\end{aligned}$$

Maximum Likelihood Estimators

- Write the *likelihood equations* as

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta_k} \;=\; \sum_i \left\{ y_i x_{ik} - x_{ik} \exp\left(\sum_j x_{ij} \widehat{\beta}_j \right) \right\} \;=\; 0.$$

- If we write $\exp\left(\sum_j x_{ij}\widehat{\beta}_j\right) = \widehat{\mu}_i$ we get

$$\sum_i (y_i - \widehat{\mu}_i)\, x_{ik} \;=\; 0$$

  which leads to the *normal equations*

$$X^T y \;=\; X^T \widehat{\mu}.$$

- Note however we must still use numerical approximation to find $\widehat{\mu}$.

Alternatives to Poisson Regression

- We can also try a standard linear regression, ignoring the fact that y is a count. The text starts out this way with the Species dataset
  - Simple linear regression did not give a good fit — variance increased with fitted value
  - Box=Cox transformation suggested $\lambda = 0.3$ but $\lambda = 0.5$ was almost as good on the plot
  - In fact taking $\lambda = 0.5$ is a standard trick for count data — the reason is given on the next slide
  - This improves on the untransformed linear regression but it still isn't perfect
  - Another problem with the square root transformation is difficulty of interpreting the resulting model — Poisson regression with log link is much easier to understand

Rationale for Square Root Transformation

- Suppose Y is Poisson with mean μ moderately large (say $\mu \geq 10$)
- The mean and variance of Y are both μ
- Write $Y = \mu(1 + \mu^{-1/2}\epsilon)$ where ϵ has mean 0 and variance 1
- Then $Y^{1/2} = \mu^{1/2}(1 + \mu^{-1/2}\epsilon)^{1/2} \approx \mu^{1/2}\left(1 + \frac{1}{2}\mu^{-1/2}\epsilon\right)$.
- $Y^{1/2}$ has mean approximately $\mu^{1/2}$ and variance approximately $\frac{1}{4}$ — *independent of* μ
- Therefore, a regression with $Y^{1/2}$ as the response should have approximately constant variance (standard deviation $\approx$ 0.5)
- However in the Species example, the residual standard error is 2.77, so this doesn't seem to work well either
- May indicate *overdispersion*

Deviance and Pearson X^2

- As for binary case, compare log likelihood for a saturated model (μ_i unrestricted) with the linear model being fitted,
- $\ell_1 \; = \; \sum_i \left(y_i \log y_i - y_i - \log y_i!\right)$
- $\ell_0 \; = \; \sum_i \left(y_i \log \widehat{\mu}_i - \widehat{\mu}_i - \log y_i!\right)$
- Deviance is

$$D \; = \; 2(\ell_1 - \ell_0) \; = \; 2\sum_i \left(y_i \log \frac{y_i}{\widehat{\mu}_i} - (y_i - \widehat{\mu}_i) \right).$$

- We can also calculate the Pearson X^2 statistic

$$X^2 \; = \; \sum_i \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

Overdispersion

- Sometimes a more reasonable model may be $\mathrm{E}(y_i) = \mu_i$, $\mathrm{Var}(y_i) = \phi\mu_i$ where ϕ is a constant known as the *overdispersion* (usually but not necessarily $\phi > 1$

- How to spot?
  - Plots of squared residuals against fitted values as in Fig. 5.3 (right — note that the plot is on a log scale here!)
  - Formal test of fit based on deviance or Pearson residuals (here leads to decisive rejection of the null hypothesis)

- Remedy — use `family=quasipoisson`

- For the species example we get a huge value $\phi = 31.7$

- There are still some observations with large Cook statistic but not nearly so bad as with the regular Poisson model