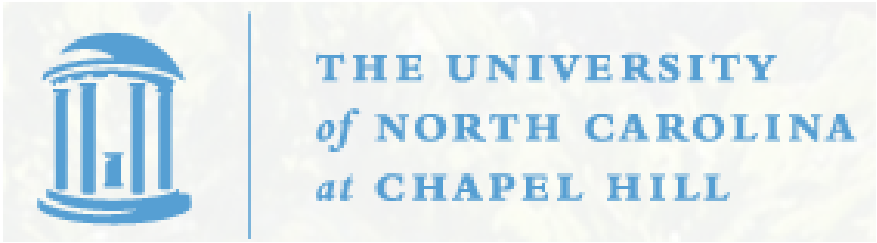
STOR 556: ADV METH DATA ANAL
Instructor: Richard L. Smith

Class Notes #9:

February 7, 2019

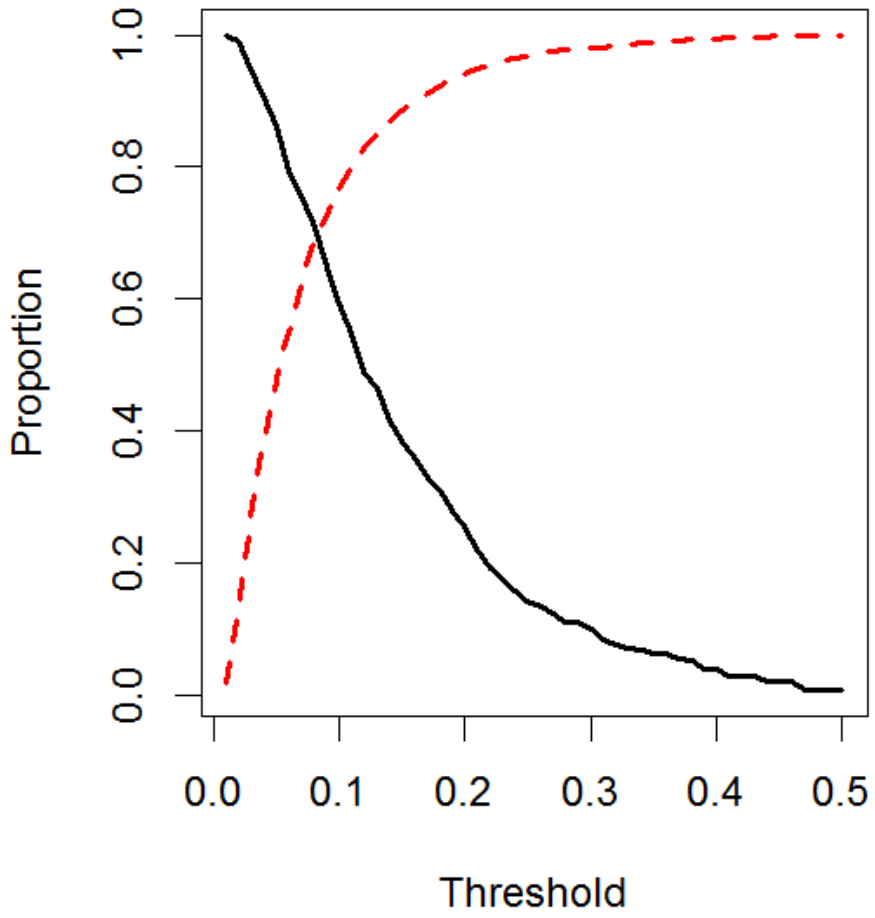


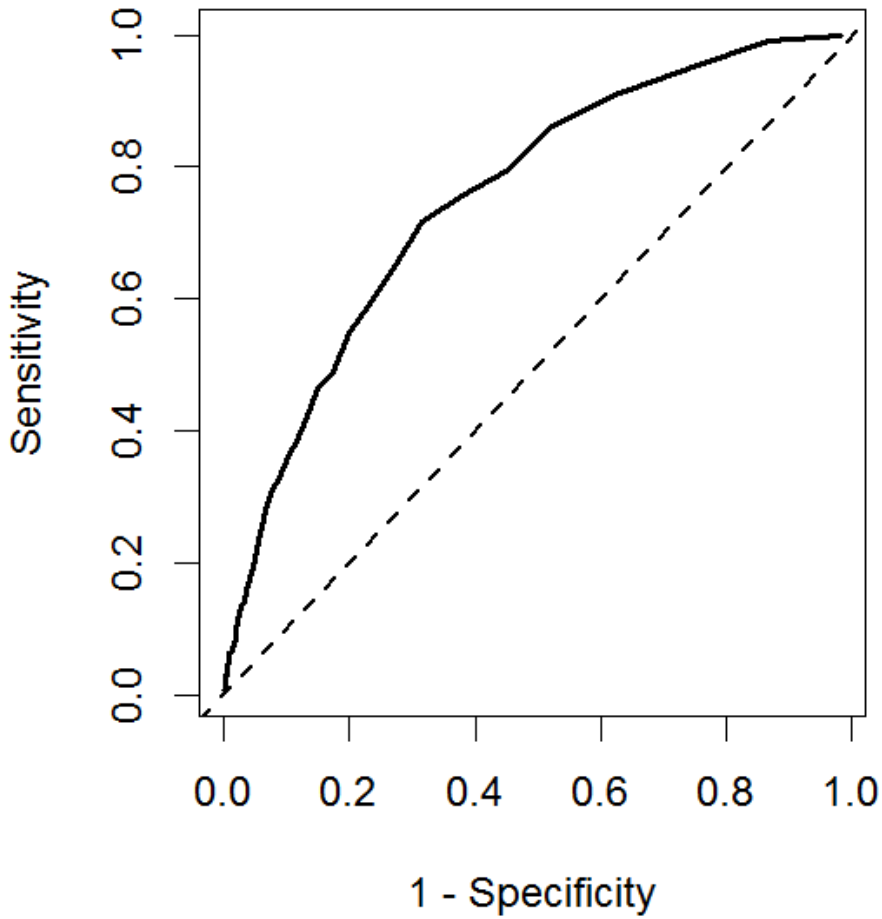
Scheduling a Take-home Midterm/Final

- 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

Review Sensitivity and Specificity

- Assume we are testing for a disease or some specific health outcome, and we use a diagnostic test to predict the outcome
- Specificity: the probability that a person who *does not have* the disease is correctly predicted to not have the disease
- Sensitivity: the probability that a person who *does have* the disease is correctly predicted to have the disease
- After subtracting from 1, these are analogous to type I error and type II error, respectively
- Sensitivity is also the *power of the test*
- As the threshold for detection rises, the specificity increases but the sensitivity decreases
- The plot of Sensitivity against 1-Specificity is called the *Receiver Operating Characteristic* or ROC curve


Proportion
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
Threshold


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity

CHAPTER 3: BINOMIAL AND PROPORTION DATA

Model and Likelihood Function

- $\Pr\{Y_i = y_i\} = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i - y_i},\ \ i = 1, ..., n,\ \ 0 \le y_i \le m_i.$
- $\eta_i = \log \frac{p_i}{1-p_i} = \sum_{j=0}^{q} \beta_j x_{ij}$
- Write ℓ for log likelihood,

$$\begin{aligned}
\ell &= \sum \left\{ y_i \log p_i + (m_i - y_i) \log(1-p_i) + \log \binom{m_i}{y_i} \right\} \\
&= \sum \left\{ y_i \log \frac{p_i}{1-p_i} + m_i \log(1-p_i) + \log \binom{m_i}{y_i} \right\} \\
&= \sum \left\{ y_i \eta_i - m_i \log(1 + e^{\eta_i}) + \log \binom{m_i}{y_i} \right\}.
\end{aligned}$$

- Hence derive likelihood equations $\frac{\partial \ell}{\partial \beta_k} = 0$ for $k = 0, ..., q$.

Deviance

- Compare model H_0 with fitted parameters $\beta_0, ..., \beta_q$ ($DF = q+1$) with alternative in which p_i's are unrestricted ($DF = n$)
- Under H_1, estimate $p_i = \frac{y_i}{m_i}$, fitted values same as y_i
- Under H_0, assume estimates $\widehat{p}_i$ and fitted values $\widehat{y}_i$
- Therefore, the deviance statistic is

$$\begin{aligned} D &= 2\sum_i \left\{ y_i \log \frac{y_i}{m_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i} + \log \binom{m_i}{y_i} \right\} \\ &\quad -2\sum_i \left\{ y_i \log \frac{\widehat{y}_i}{m_i} + (m_i - y_i) \log \frac{m_i - \widehat{y}_i}{m_i} + \log \binom{m_i}{y_i} \right\} \\ &= 2\sum_i \left\{ y_i \log \frac{y_i}{\widehat{y}_i} + (m_i - y_i) \log \frac{(m_i - y_i)}{(m_i - \widehat{y}_i)} \right\} \end{aligned}$$

Deviance and Pearson Residuals

- See page 53. If H_0 is correct, D is approximately χ^2_{n-q-1}.
  - Assumes m_i not too small. Maybe $m_i \geq 5$ could be guideline.
- An alternative formula (page 55) is

$$X^2 = \sum_i \frac{(y_i - m_i\widehat{p}_i)^2}{m_i\widehat{p}_i(1-\widehat{p}_i)}.$$

- Pearson residuals are

$$r_i^P = \frac{(y_i - m_i\widehat{p}_i)}{\sqrt{\text{var } \widehat{y}_i}}$$

- In R: `residuals(lmod,type='pearson')`
- X^2 should be close to the deviance but not always (p. 55)
- Discrepancy may suggest *overdispersion*

Example with orings data

- Binomial model with temperature as covariate

  `lmod=glm(cbind(damage,6-damage)~temp,family=binomial,orings)`

- Deviance is 16.9 but $X^2 = 28.1$ with 21 DF

- `pchisq(28.1,21,lower=F)` gives 0.137 so no problem with goodness of fit — test statistic X^2 is "not significant"

- Nevertheless $\frac{28.1}{21} = 1.34$ implies some overdispersion

- If we correct for this, the standard error of the temperature term is increased

```
> lmod=glm(cbind(damage,6-damage)~temp,family=binomial,orings)
> sumary(lmod)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 11.662990   3.296263  3.5382 0.0004028
temp        -0.216234   0.053177 -4.0663 4.777e-05
.
.
> lmodod=glm(cbind(damage,6-damage)~temp,family=quasibinomial,
> orings)
>
> sumary(lmodod)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.662990   3.810774  3.0605 0.005938
temp        -0.216234   0.061477 -3.5173 0.002047

Dispersion parameter = 1.33654
```

Binomial or quasibinomial?

- Both give same regression coefficients
- Quasibinomial allows for overdispersion (here 1.34) — more "robust" but leads to higher standard errors for coefficients (lower t statistics)
  - Side issue — p-values for binomial are based on normal distribution but p-values for quasibinomial are based on t distribution. I don't know the reason for that.
- If we were confident the binomial model was correct, that would be right thing to do
- However there are also reasons why overdispersion might be a factor, e.g. other variations in experimental conditions