STATISTICS 536B, Lecture #1

February 24, 2015

A few highlights: logistic regression

logit function: $\text{logit}(p) = log\{p/(1-p)\}$

expit function: $\text{expit}(z) = \text{logit}^{-1}(z) = 1/\{1 + \exp(-z)\}$

Say have a binary disease variable Y, a binary exposure variable X and some potential confounding variables $C = (C_1, \ldots, C_p)$.

Sampling: either joint (Y, X, C) or conditional $(Y|X, C)$.

Model specification

$$\text{logit} Pr(Y = 1|X, C) \quad = \quad \beta_0 + \beta_1 X + \text{other terms}$$

implies

$\log OR(Y, X|C) =$

Some examples

$$\text{logit} Pr(Y = 1|X, C) \qquad \text{OR}(Y, X|C)$$

Prediction Accuracy and ROC Curve

Write model as logit$Pr(Y = 1|X, C) = \beta^T W$

Predict outcome for a subject with $W = w$ as $\tilde{y} = I\{\hat{\beta}^T w > k\}$, for some choice of **threshold** k.

(in-sample prediction, out-of-sample prediction)

Sensitivity

Specificity

ROC Curve

Unmatched case-control study

Sample some controls $(X, C|Y = 0)$ and some cases $(X, C|Y = 1)$

Potentially **huge** savings in cost/time.

Think real-world distribution of (X, C, Y) given by $f(x, c, y)$.

Think of distribution giving rise to the data as $f^*(x, c, y)$.

So $f^*(y)$ is **directly controlled by the study investigator**, while $f^*(x, c|y) = f(x, c|y)$.

Are we justified in collecting data according to f^* but analyzing it as if it were collected according to f?

$\text{logit}\,Pr^*(Y = 1|X, C) =$

$\text{logit}\,Pr(Y = 1|X, C) + \log \text{Odds}^*(Y = 1) - \log \text{Odds}(Y = 1)$

What about joint sampling of (X,Y)?

	X=1	X = 0	
Y=1	Z_{11}	Z_{01}	
Y=0	Z_{10}	Z_{00}	
			n

$$(Z_{00}, Z_{01}, Z_{10}, Z_{11}) \quad \sim \quad \text{Multinomial}(n; p_{00}, p_{01}, p_{10}, p_{11})$$

$$\log \hat{OR} = \log Z_{11} + \log Z_{00} - \log Z_{10} - \log Z_{01}$$

$$SE \quad = \quad \sqrt{\sum_{i=0}^{1}\sum_{j=0}^{1}\frac{1}{z_{ij}}}$$

Still valid?

Today is 'use the Delta method day!'

$$
\begin{aligned}
\log \hat{OR} &= \log n^{-1} Z_{11} + \log n^{-1} Z_{00} - \log n^{-1} Z_{10} - \log n^{-1} Z_{01} \\
&= \sum_{i=0}^{1} \sum_{j=0}^{1} \pm \log n^{-1} Z_{ij} \\
&\approx \sum_{i=0}^{1} \sum_{j=0}^{1} \pm \left\{ \log p_{ij} + \left(\frac{1}{p_{ij}} \right) (n^{-1} Z_{ij} - p_{ij}) \right\}
\end{aligned}
$$

So the approximation to $Var \log \hat{OR}$ will have variance and covariance terms...

Recall multinomial moments

If

$$(Z_{00}, Z_{01}, Z_{10}, Z_{11}) \quad \sim \quad \text{Multinomial}(n; p_{00}, p_{01}, p_{10}, p_{11})$$

Then

$E(Z_{ij}) = np_{ij}$

$Var(Z_{ij}) = np_{ij}(1 - p_{ij})$

$Cov(Z_{ij}, Z_{rs}) = -np_{ij}p_{rs}$

Variance terms

$$\sum \left(\frac{1}{p_{ij}}\right)^2 \mathrm{Var}\left(n^{-1} Z_{ij}\right) =$$

Covariance terms

$$\sum(\pm)\mathrm{Cov}\left(\frac{1}{p_{ij}}\frac{Z_{ij}}{n}, \frac{1}{p_{rs}}\frac{Z_{rs}}{n}\right) =$$