STATISTICS 536B, Lecture #4

March 5, 2015

Recap: Interested in population-level association between (X, Y).

Various study designs:

- prospective, sample $(Y|X)$
- retrospective, sample $(X|Y)$
- cross-sectional, sample (X, Y)

Same analysis:

```
       X=1  X=0
---------------
Y=1  |   q    r
Y=0  |   s    t
```

$$\widehat{\log OR} = \log q + \log t - \log r - \log s$$

$$SE\left[\widehat{\log OR}\right] = \sqrt{1/q + 1/r + 1/s + 1/t}$$

And with confounders

Logistic regression of Y on $(1, X, C)$ [or more generally basis functions of (X,C)] is appropriate, whether the actual data acquisition is prospective [sampling $(Y \mid X, C)$], cross-sectional [sampling (Y, X, C)], or retrospective [sampling $(X, C \mid Y)$]

Caveat about estimating intercept - intuitively sensible - case-control data cannot tell you how common the disease is in the population.

Matched Case-Control Data with binary exposure: Cross-Classify the *n* **PAIRS** of subjects

		Case: Not exposed	Case: Exposed	
Control:	Not exposed	a	b	
	Exposed	c	d	
				n

Think about the source population having a distribution over disease status, exposure, and matching factors (confounders) jointly: $f(y, x, m)$

Think about sampling a pair of individuals yielding:
$\{(Y_0, X_0, M_0), (Y_1, X_1, M_1)\}$
but with the **constraints** $Y_0 + Y_1 = 1$ and $M_0 = M_1$.

A likelihood for parameters describing the source population based on how the data were actually sampled??? Activity.

So we have a likelihood function for β, the $(Y, X|M)$ log-OR

$$L(\beta) = (1/2)^a (1/2)^d \{1/(1+\exp(-\beta))\}^b \{1/(1+\exp(\beta))\}^c,$$

$$l(\beta) = -b\log(1+\exp(-\beta)) - c\log(1+\exp(\beta)) + \text{constant}$$

And we can apply our usual tools to the log-likelihood, to get inference procedures

Also a **score test** - mentioned in reading

General idea of score test. Have log-likelihood $l(\beta)$ for scalar parameter β.

Under the hypothesis $\beta = 0$,

$$\frac{l'(0)}{\sqrt{-l''(0)}} \overset{\text{approx}}{\sim} N(0,1)$$

Matching is inefficient/efficient when the matching factor(s) is a weak/strong confounder???

```
### set up a population
t <- 100000
set.seed(17)

m.pop <- sample(0:6, size=t, prob=rep(1/7, 7), replace=T)

x.pop <- rbinom(t, size=1,
           prob=expit(logit(.1)+(m.pop/6)*(logit(.9)-logit(.1))))

y.pop <- rbinom(t, size=1, prob=expit(-2 + 0.5*m.pop + 0.5*x.pop))
```

Sample some cases and controls

```
n <- 400
cs <- sample((1:t)[y.pop==1], size=n)

### one possible way to complete the study - unmatched controls
cn.unmt <- sample((1:t)[y.pop==0], size=n)

### another possible way to complete the study - matched controls
cn.mtch <- rep(NA,n)
for (i in 1:n) {
  cn.mtch[i] <- sample((1:t)[(y.pop==0)&(m.pop==m.pop[cs[i]])], size=1)
}
```

Do the matched study analysis

```
dat.pair <- table(x.pop[cn.mtch], x.pop[cs])

> dat.pair

     0   1
 0  66  72
 1  65 197

> c(log(dat.pair[1,2])-log(dat.pair[2,1]),
    sqrt(1/dat.pair[1,2] + 1/dat.pair[2,1]))

 0.102 0.171
```

Do the unmatched study analysis

```
y <- y.pop[c(cs, cn.unmt)]
x <- x.pop[c(cs, cn.unmt)]
m <- m.pop[c(cs, cn.unmt)]

ft <- glm(y~x+as.factor(m), family=binomial)

> c(coef(ft)[2], sqrt(vcov(ft)[2,2]))
 0.260 0.191
```

And repeat...