STATISTICS 536B, Lecture #3

March 3, 2015

General options for binary Y, binary X, confounders $C = (C_1, \ldots, C_p)$

- Build a logistic regression model for Y in terms of X and C
- Think of the data "carved up" into a bunch of 2×2 (Y, X) mini data tables, one for each distinct value of C manifested in the data.

For today's lecture, will write m as the index to keep track of these mini-tables.

Demo: Generating and analyzing unmatched case-control data

```
### 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(2 - (4/6)*m.pop))

y.pop <- rbinom(t, size=1, prob=expit(-3 + (4/9)*(m.pop-3)^2 + 0.5*x.pop))
```

Misspecified model messes up the controlling for confounders

```
### do a (balanced, unmatched) case-control study
n <- 2000
set.seed(133)
cs <- sample((1:t)[y.pop==1], size=n)
cn <- sample((1:t)[y.pop==0], size=n)

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

### adjustment using main effect of m
ft <- glm(y~x+m, family=binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.07383    0.08813  -0.838 0.402178
x            0.26231    0.07781   3.371 0.000749 ***
m           -0.02001    0.01792  -1.117 0.264118
```

Aside: what about no control, proper control

```
glm(formula = y ~ x, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.15838    0.04516  -3.508 0.000452 ***
x            0.31269    0.06344   4.929 8.28e-07 ***
```

```
glm(formula = y ~ x + m + I(m^2), family = "binomial")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.84813    0.13277  13.920  < 2e-16 ***
x            0.49335    0.10175   4.849 1.24e-06 ***
m           -2.81596    0.08521 -33.046  < 2e-16 ***
I(m^2)       0.46643    0.01356  34.397  < 2e-16 ***
```

Adjust for confounding without modelling?

```
dat.strfd <- table(y,x,m)

> dat.strfd
, , m = 0
   x
y     0   1
  0  12  60
  1  86 615

, , m = 1
   x
y     0   1
  0  62 227
  1  44 230

...

, , m = 6
   x
y     0   1
  0 108   5
  1 535  77
```

So have stratum-specific log-OR and SE

```
### function to get log-OR and SE from 2 by 2 table
myfun.lor <- function(dattbl) {
  est <- log(dattbl[1,1])+log(dattbl[2,2])-log(dattbl[1,2])-log(dattbl[2,1])
  se <- sqrt(sum(1/as.vector(dattbl)))
  c(est,se)
}


inf.strfd <- apply(dat.strfd, 3, myfun.lor)
```

```
      m
            0     1     2     3     4     5     6
  [1,] 0.358 0.356 0.563 0.713 0.661 0.328 1.134
  [2,] 0.337 0.218 0.291 0.323 0.276 0.212 0.473
```

General statistical strategy for combining multiple estimators of the same target

```
ests <- inf.strfd[1,]
vars <- inf.strfd[2,]^2
whts <- (1/vars)/sum(1/vars)

est.cmbd <- sum(whts*ests)
se.cmbd <- sqrt(sum(whts^2*vars))


c(est.cmbd, se.cmbd)
[1] 0.496 0.105
```

Homogeneity of (X,Y) odds-ratio across strata?

We've *assumed* this. Empirically testable?

```
> sum( (ests-est.cmbd)^2/vars )
[1] 3.891315
```

Or find a package...

```
> library(epicalc)
> mhor(mhtable=dat.strfd)

Stratified analysis by  m
                OR lower lim. upper lim.  P value
m 0           1.43      0.672       2.82 2.68e-01
m 1           1.43      0.912       2.25 1.07e-01
m 2           1.75      0.968       3.30 5.98e-02
m 3           2.04      1.045       4.11 3.08e-02
m 4           1.93      1.086       3.45 2.20e-02
m 5           1.39      0.896       2.15 1.35e-01
m 6           3.11      1.232      10.07 9.33e-03
M-H combined 1.67      1.362       2.05 8.22e-07

M-H Chi2(1) = 24.3 , P value = 0
Homogeneity test, chi-squared 6 d.f. = 3.91 , P value = 0.689
```

Is the Odds-Ratio also a **collapsible** measure of association?

Example. Say the joint distribution of (C, X, Y) is characterized by:

- X and C are independent and identically distributed as Bernoulli(0.5)
- $Pr(Y = 1X = 0, C = 0) = 0.2$ and $Pr(Y = 1|X = 0, C = 1) = 0.5$
- The stratified odds-ratios are $OR(Y, X|C = c) = 4$, for both $c = 0$ and $c = 1$.

Is C a confounder?

What is the crude odds-ratio $OR(Y, X)$?