STATISTICS 536B, Lecture #8

March 24, 2015

Back to Confounding

Let's set up a really nasty problem.

```
n <- 10000

cnf <- mvrnorm(n, mu=rep(0,4), Sigma=.7*diag(4) + .3*matrix(1,4,4))

x <- rbinom(n, size=1, prob=expit(0.5*cnf[,1] + 0.5*cnf[,2]))

y <- rnorm(n,
           mean = x + 20*exp( - (3*(cnf[,1]-.5)^2 + 6*(cnf[,2]-.5)^2))
                    + 15*exp(-(cnf[,3]-1)^2),  sd=1)
```

Pity the poor statistician who has to deal with this confounding...

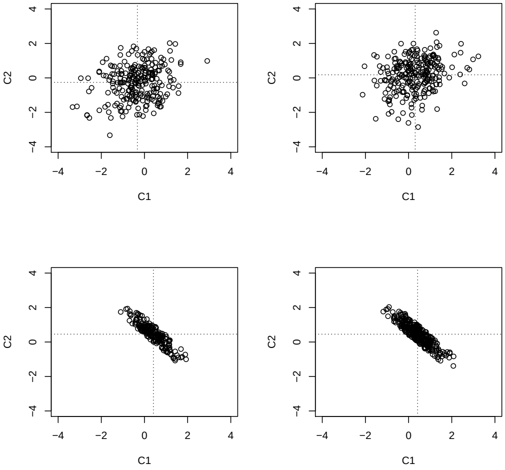
```
lm(formula = y ~ x)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.00612    0.09856   71.08   <2e-16 ***
x            3.01746    0.13905   21.70   <2e-16 ***


lm(formula = y ~ x + cnf)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.93694    0.07807 101.671   <2e-16 ***
x            1.25967    0.11393  11.057   <2e-16 ***
cnf1         0.52298    0.05966   8.766   <2e-16 ***
cnf2         0.62451    0.05968  10.465   <2e-16 ***
cnf3         4.14476    0.05749  72.100   <2e-16 ***
cnf4        -0.06884    0.05766  -1.194    0.233
```

What is this all about???

```
xpsrmod <- glm(x~cnf, family=binomial)
> coef(xpsrmod)
 (Intercept)         cnf1         cnf2         cnf3         cnf4
 0.011181043  0.521858073  0.538290344  0.009976775 -0.051287371


prpns <- fitted(xpsrmod, response=T)
grp <- cut(prpns, breaks=c(0, quantile(prpns, c(.2,.4,.6,.8)), 1))

est <- se <- rep(NA,5)
for (i in 1:5) {
  ft <- lm(y~x, subset=(grp==levels(grp)[i]))
  est[i] <- coef(ft)[2]
  se[i] <- sqrt(vcov(ft)[2,2])
}

> est
[1] 1.35 1.18 1.04 1.13 0.67
> se
[1] 0.26 0.24 0.26 0.38 0.32
```

And if we want to combine estimates, we know how...

```
est.combo <- sum(est/se^2) / sum(1/se^2)

se.combo <- sqrt(1/sum(1/se^2))

> c(est.combo, se.combo)
[1] 1.10 0.13


### some clue as to what is going on ???

ndx <- sample(1:n, size=500)
plot(cnf[ndx,1][x[ndx]==0], cnf[ndx,2][x[ndx]==0])
plot(cnf[ndx,1][x[ndx]==1], cnf[ndx,2][x[ndx]==1])

ndx <- sample((1:n)[grp==levels(grp)[4]], size=500)
plot(cnf[ndx,1][x[ndx]==0], cnf[ndx,2][x[ndx]==0])
plot(cnf[ndx,1][x[ndx]==1], cnf[ndx,2][x[ndx]==1])
```


C1
C2
−4
−2
0
2
4

Yet another estimate of the unconfounded relationship between Y and X?

```
> mean( y * (x/prpns - (1-x)/(1-prpns)) )
[1] 1.16

### SE?
> sqrt(var(y*(x/prpns - (1-x)/(1-prpns))) / n)
[1] 0.24
```

Why estimate $E\left\{Y\left(\frac{X}{Z(C)}-\frac{1-X}{1-Z(C)}\right)\right\}$?