4/20/2010 36-402/608 ADA-II H. Seltman

Breakout #23: Mediation 1

Simulation of an experiment

```
x = rnorm(n=100, mean=5, sd=1)
x2 = rnorm(n=100, mean=5, sd=1)
y = rnorm(n=100, mean=15+3*x+4*x2, sd=2.5)

summary(lm(y ~ x))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  39.1052     2.7368  14.289  < 2e-16
# x             2.1867     0.5406   4.045 0.000104

summary(lm(y ~ x2))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  32.5712     1.6107   20.22   <2e-16
# x2            3.4515     0.3109   11.10   <2e-16

summary(lm(y ~ x + x2))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  16.8382     1.8540   9.082 1.29e-14
# x             2.8418     0.2690  10.563  < 2e-16
# x2            3.7677     0.2152  17.506  < 2e-16
```

Question 1: Draw a "directed acyclic graph" (DAG) in the form of a simple diagram of the variables x, x2, and y connected with arrows showing causality, i.e. A→B means changes in A cause changes in B. Compare the estimated (causal) effects to the true effects. What happens when x and x2 are correlated?

$x \rightarrow y \leftarrow x2$

The x coefficients (2.1867 and 2.8418) are estimates of the true causal effect of x on y (when x goes up by 1, y goes up by 3). The x2 coefficients similarly estimate the true x2 causal effect of 4.

Here is an example with correlated x's:

```
library(MASS)
# 0.9 * 1 * 1 = 0.9 # covariance for cor=0.9, vars=1
x34 = mvrnorm(30, mu=c(3,4), Sigma=matrix(c(1,0.9,0.9,1),2))
x3 = x34[,1]
```

```
x4 = x34[,2]
cor(x3,x4) # 0.89
y34 = rnorm(30, mean=15+3*x3+4*x4, sd=7)
summary(lm(y34~x3))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   23.816      3.585   6.642 3.31e-07
# x3             5.686      1.075   5.291 1.25e-05

summary(lm(y34~x4))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   15.968      4.924   3.243  0.00305
# x4             6.108      1.133   5.390 9.55e-06

summary(lm(y34~x3+x4))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   18.357      5.302   3.462   0.0018
# x3             2.765      2.367   1.168   0.2529
# x4             3.475      2.519   1.379   0.1791
```

If x and x2 are correlated, then either or both may be "nonsignificant" in the combined model. This is because with sufficient "shared" information between the x's, neither adds information about y beyond what is provided by the other.

Simulation of an observational study

```
z = rnorm(n=100, mean=5, sd=1)
x = rnorm(n=100, mean=20+2*z, sd=2)
y = rnorm(n=100, mean=15+3*z, sd=1.5)

summary(lm(y ~ x))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  7.35008    2.95870   2.484   0.0147
# x            0.76111    0.09902   7.687 1.18e-11
```

Question 2: Draw the DAG. Explain why this shows that observational studies can't be used to claim causal relationships.

$$x \leftarrow z \rightarrow y$$

Even though x and z are highly correlated it would be a mistake to conclude that x causes y. In fact z cause x and y, and if we could/would manipulate x, that would have no effect on y. Variable z is a confounder (lurking variable). One or more confounding z's is always possible (and not unlikely) in any observational study. In a randomized experiment the average of z (and therefore the average causal effect of z on y) is the same for for each level of x, so we *can* attributed any observed change in y to the manipulation of x.

Simulation of a mediator (causal) model

```
x = rnorm(n=100, mean=20, sd=2)
m = rnorm(n=100, mean=10+3*x, sd=1.5)
y = rnorm(n=100, mean=15+2*m, sd=1)

summary(lm(m ~ x))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 10.97590    1.85094    5.93 4.55e-08
# x            2.94580    0.09072   32.47  < 2e-16

summary(lm(y ~ m))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 15.74659    1.18391    13.3   <2e-16
# m            1.99179    0.01666   119.5   <2e-16

summary(lm(y ~ x))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   37.431      3.775   9.915   <2e-16
# x              5.876      0.185  31.758   <2e-16

summary(lm(y ~ m + x))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 15.91940    1.22443  13.002   <2e-16
# m            1.95986    0.05733  34.188   <2e-16
# x            0.10280    0.17654   0.582    0.562
```

Question 3: Draw the DAG. Interpret each regression with respect to the DAG. The effects of X on M, M on Y, and X on Y ignoring M (with M not in the model) are called "direct" effects. Relate the X on M and M on Y direct estimates to the simulated (causal) values. The "indirect" effect of X on Y is defined as the product of the two direct effects. How does it relate to the direct effect of X on Y? Explain what happened to the X coefficient in the final model.

$$x \to m \to y$$

This is "complete" mediation when x has no effect on y except through its effect on m. According to the simulation, when x goes up by 1, m goes up by 3 on average. And when m goes up by 3, y goes up by 6 on average. So when x goes up by 1, y goes up by 6 on average. In general the indirect mediated effect of x on y is the product of the X on M effect (usually designated “a”) and the M on Y effect (“b”) which equals *ab*.

The x coefficient becomes non-signficant and falls to near zero when it is in a regression model with y because a change in x while holding m constant has no effect no y, while a change in m while holding x constant would change y. This is another way of stating that m mediates the effect of x on y.

Question 4: Construct a simple set of non-quantitative rules that are based on high (>0.05) vs. low (<=0.05) p-values and that could be used to assess mediated causation.

A common set of rules is:

1. the regression of y on x should have a significant (slope) coefficient
2. the regression of m on x should have a significant coefficient
3. the regression of y on m should have a significant coefficient
4. the coefficient of x in the regression of y on m *and* x should drop to near zero, and its p-value should become non-significant.

A partial mediation model

```
x = rnorm(n=100, mean=20, sd=2)
m = rnorm(n=100, mean=10+3*x, sd=1.5)
y = rnorm(n=100, mean=15+1.5*x+2*m, sd=1)

summary(lm(m ~ x))
#             Estimate Std. Error t value Pr(>|t|)    f
# (Intercept) 11.85906    1.51144   7.846 5.39e-12
# x            2.90992    0.07541  38.588  < 2e-16

summary(lm(y ~ m))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 10.30802    1.39136   7.409 4.53e-11
# m            2.49497    0.01983 125.796  < 2e-16

summary(lm(y ~ x))
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  38.4438     3.3605   11.44   <2e-16
```

```
# x             7.3329     0.1677   43.74   <2e-16

summary(lm(y ~ m + x))
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) 13.36256    1.32948  10.051  < 2e-16
# m            2.11494    0.06963  30.372  < 2e-16
# x            1.17863    0.20919   5.634 1.72e-07
```

Question 5: How would you modify the rules to accommodate partial mediation?

In the more common partial mediation (as opposed to complete mediation), the fourth rule becomes "the coefficient of x in the regression of y on m *and* x should drop, and its p-value should rise.

This additional example shows that use of mediation analysis does *not* protect against false causal conclusions in observational studies. Although the rules suggest that m partially mediates the effect of x on y, x actually has no causal effect on y.

```
> z = rnorm(n=100, mean=20, sd=2)
> x = rnorm(n=100, mean=20+z, sd=1)
> m = rnorm(n=100, mean=10+3*z, sd=1.5)
> y = rnorm(n=100, mean=15+2*m, sd=1)
> summary(lm(m~x))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.2894     5.9950  -4.385 2.92e-05 ***
x             2.4004     0.1507  15.930  < 2e-16 ***

> summary(lm(y~m))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8689     1.3036   11.41   <2e-16 ***
m             2.0006     0.0188  106.43   <2e-16 ***

> summary(lm(y~x))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.8781    11.8265  -3.372  0.00107 **
x             4.8564     0.2973  16.338  < 2e-16 ***

> summary(lm(y~x+m))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.18110    2.27895   4.906 3.74e-06 ***
x            0.19442    0.09922   1.959   0.0529 .
m            1.94220    0.03511  55.318  < 2e-16 ***
```