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?

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.

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.

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 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?