simulated_endogeneity-1.R

danny

2020-02-05

```r
# Author: Gordon Burtch and Gautam Ray
# Course: MSBA 6440
# Session: Causality and Endogeneity
# Topic: Simulating Endogeneity

set.seed(100)

## Some examples of what happens when we ignore different kinds of endogeneity

# 1) Measurement Error

# We build our variable X, and then also an erroneously measured version of X.
X <- rnorm(200, mean = 50, sd=7)
X_m <- X + rnorm(200,mean=4, sd=15)

# Now we simulate Y using the true data generating process (accurately measured X)
Y <- 0.5*X + rnorm(200,mean=0,sd=1)

# You can see that the estimate is hugely deflated when we ignore the measurement error.
summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3031 -0.6666  0.0320  0.6496  2.8931
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06088    0.59601  -0.102    0.919
## X            0.50114    0.01181  42.423   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.065 on 198 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.9004
## F-statistic:  1800 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Y~X_m))
```

```
##
## Call:
## lm(formula = Y ~ X_m)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4062 -1.8631  0.1169  1.6514  7.1989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.40168    0.68256  29.890  < 2e-16 ***
## X_m          0.08631    0.01211   7.125  1.9e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.018 on 198 degrees of freedom
## Multiple R-squared:  0.2041, Adjusted R-squared:    0.2
## F-statistic: 50.76 on 1 and 198 DF,  p-value: 1.895e-11
```

```r
# 2) Omitted Variables // Correlated Unobservable

# Let's add in a confounder for X that we will "not observe" in our regression and see what it does.

Z <- rnorm(200, mean=3, sd=.5) - X
Y <- 0.5*X + 2*Z + rnorm(200,mean=0,sd=1)

# You can see that ignoring Z causes X to be downward biased (because Z is negatively correlated with X
summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0211 -1.0238  0.0938  1.0383  3.6412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6731     0.8376   7.967 1.25e-13 ***
## X            -1.5111     0.0166 -91.026  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.497 on 198 degrees of freedom
## Multiple R-squared:  0.9767, Adjusted R-squared:  0.9765
## F-statistic:  8286 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Y~X+Z))
```

```
##
## Call:
## lm(formula = Y ~ X + Z)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.57143 -0.78364  0.02501  0.77308  3.02907
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9656     0.7206   1.340 0.181755
## X             0.4844     0.1429   3.389 0.000848 ***
## Z             2.0025     0.1429  14.009  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.062 on 197 degrees of freedom
## Multiple R-squared:  0.9883, Adjusted R-squared:  0.9882
## F-statistic:  8326 on 2 and 197 DF,  p-value: < 2.2e-16
```