AREC 736 D HW!

Innocent Vomitadyo

2022-10-17

``

#1. Simulate a dataset with one X variable and one Y variable in R or Stata that does not contain a confounder or a collider. Let the relationship between Y and X be 5, that is the population parameter 𝛽5 = 5 (you can set 𝛽0 to anything you want). Assuming X causes Y, recover the coefficient on X using OLS.

rm(list =ls())  
X <- floor(runif(1000, min=0, max=45))  
E <- rnorm(1000, mean=0, sd=0.01)  
B0 <- 2  
B1 <- 5  
Y <- B0 + B1\*X + E  
df <- data.frame(Y, X)  
head(df)

## Y X  
## 1 207.011235 41  
## 2 101.999213 20  
## 3 136.997111 27  
## 4 197.009842 39  
## 5 47.006366 9  
## 6 2.000554 0

summary(df)

## Y X   
## Min. : 1.986 Min. : 0.00   
## 1st Qu.: 57.005 1st Qu.:11.00   
## Median :112.000 Median :22.00   
## Mean :112.201 Mean :22.04   
## 3rd Qu.:167.008 3rd Qu.:33.00   
## Max. :222.015 Max. :44.00

#Recover the coefficient on X using OLS.

OLS1 <- lm(Y~X, data=df)  
OLS1

##   
## Call:  
## lm(formula = Y ~ X, data = df)  
##   
## Coefficients:  
## (Intercept) X   
## 2 5

summary(OLS1)

##   
## Call:  
## lm(formula = Y ~ X, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.033439 -0.007181 0.000033 0.006655 0.036174   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.000e+00 6.287e-04 3182 <2e-16 \*\*\*  
## X 5.000e+00 2.454e-05 203734 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01013 on 998 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 4.151e+10 on 1 and 998 DF, p-value: < 2.2e-16

#2. Alter your program from step 1 to include a confounder (a variable Z that is correlated with the X variable and the error term in your model)

Z <- floor(runif(1000, min=0, max=45))  
X <- 4+ 0.6\*Z + rnorm(1000, mean=0, sd=0.1)  
B0 <- 2  
B1 <- 5  
B2 <- 3  
Y <- B0 + B1\*X + B2\*Z + rnorm(1000, mean=0, sd=0.1)  
df1 <- data.frame(Y, X, Z)  
head(df1)

## Y X Z  
## 1 232.34110 25.089703 35  
## 2 274.23262 29.243074 42  
## 3 51.82815 6.935378 5  
## 4 21.09270 3.830235 0  
## 5 141.57975 15.945292 20  
## 6 201.86288 21.973238 30

summary(df1)

## Y X Z   
## Min. : 20.89 Min. : 3.791 Min. : 0.00   
## 1st Qu.: 83.09 1st Qu.:10.211 1st Qu.:10.00   
## Median :153.47 Median :17.110 Median :22.00   
## Mean :154.43 Mean :17.244 Mean :22.07   
## 3rd Qu.:226.20 3rd Qu.:24.449 3rd Qu.:34.00   
## Max. :287.38 Max. :30.675 Max. :44.00

#2a Estimate a model of Y on X using OLS. Report your coefficient.

OLS2 <- lm(Y~X, data=df1)  
OLS2

##   
## Call:  
## lm(formula = Y ~ X, data = df1)  
##   
## Coefficients:  
## (Intercept) X   
## -17.913 9.994

summary(OLS2)

##   
## Call:  
## lm(formula = Y ~ X, data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.52287 -0.31966 -0.00368 0.31033 1.49582   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.913152 0.036661 -488.6 <2e-16 \*\*\*  
## X 9.994434 0.001931 5175.3 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4849 on 998 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 2.678e+07 on 1 and 998 DF, p-value: < 2.2e-16

#Coefficients have changes. B0 is now ——- and B1 is now ——

#2a(i) What should it have been?

#estimates should be the same as in 1. We want B0 to be equal to 2 and B1 = 5

#2a (ii) What is the bias?

#Bias B1 = 5- 9.996728 = #Bias B0 - 2- -17.954363 =

#Bias due to Endogeneity #The missing variable is now appearing in the error term and the X variable is correlated with the error term

#2(b) Estimate a model of Y on X but include your confounder. Is the bias eliminated?

OLS3 <- lm(Y~X + Z, data=df1)  
OLS3

##   
## Call:  
## lm(formula = Y ~ X + Z, data = df1)  
##   
## Coefficients:  
## (Intercept) X Z   
## 1.930 5.019 2.988

summary(OLS3)

##   
## Call:  
## lm(formula = Y ~ X + Z, data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.27958 -0.06545 0.00301 0.06751 0.35340   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.93018 0.12987 14.86 <2e-16 \*\*\*  
## X 5.01940 0.03251 154.39 <2e-16 \*\*\*  
## Z 2.98802 0.01952 153.04 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.09803 on 997 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 3.277e+08 on 2 and 997 DF, p-value: < 2.2e-16

#The Bias is eliminated. #The estimates are close to 5 but are not exactly equal.

#Why is the coefficient not exactly equal to 5? #Becaue of the error term.

# 2c. Provide an example of an omitted variable that could cause bias in a regression.

Omitted variable is comes when we have a variable that has a significant impact on Y and correlated with X is left out of the equation. For example, A variable A causes Y and A also depends on X.

#2(d). Provide an example of selection bias that could cause bias in a regression. Is this the same as omitted variable bias?

Selection bias happens when the sample is not random such that the subset is not representative of the target population. Examples include  
-when participants from a specified population are more likely to be selected to participate in a trial than others -when the method used to assign subjects to the study groups is inadequate and produces systematic differences between the participants in the study groups. when we have Loss to follow-up and attrition.

This not the same as omitted variable bias.

##################.#######################################################################################

#2(e). Provide an example of non-random measurement error in an X variable that could lead to bias. Is this the same as omitted variable bias?

Happens when a variable is not perfectly measured

#3.Alter your program from step 1 to include a collider (a variable K that is caused by both X and Y)

X <- floor(runif(1000, min=0, max=45))  
B0 <- 2  
B1 <- 5  
Y <- B0 + B1\*X + rnorm(1000, mean=0, sd=1)  
K <- 0.60\*X +0.7\*Y + rnorm(1000, mean=0, sd=1)  
  
  
df2 <- data.frame(Y, X, K)  
head(df2)

## Y X K  
## 1 86.09493 17 68.769140  
## 2 217.95039 43 178.203810  
## 3 122.31479 24 99.374231  
## 4 186.85353 37 152.468991  
## 5 33.14685 6 25.974088  
## 6 10.50439 2 7.855109

summary(df2)

## Y X K   
## Min. : -0.2308 Min. : 0.00 Min. : -0.8302   
## 1st Qu.: 55.1551 1st Qu.:11.00 1st Qu.: 45.2972   
## Median :112.7207 Median :22.00 Median : 93.0269   
## Mean :112.5511 Mean :22.11 Mean : 92.0157   
## 3rd Qu.:167.5053 3rd Qu.:33.00 3rd Qu.:137.3127   
## Max. :224.5277 Max. :44.00 Max. :184.9322

#3(a). Estimate a model of Y on X using OLS. Report your coefficient. #(i). Is your estimate biased or not?

OLS4 <- lm(Y~X, data=df2)  
OLS4

##   
## Call:  
## lm(formula = Y ~ X, data = df2)  
##   
## Coefficients:  
## (Intercept) X   
## 1.886 5.005

summary(OLS4)

##   
## Call:  
## lm(formula = Y ~ X, data = df2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0203 -0.6060 0.0138 0.6653 3.3002   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.885790 0.058449 32.26 <2e-16 \*\*\*  
## X 5.004536 0.002273 2201.80 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9434 on 998 degrees of freedom  
## Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998   
## F-statistic: 4.848e+06 on 1 and 998 DF, p-value: < 2.2e-16

Estimates are not biased.

# 3(b). Estimate a model of Y on X but include your collider. Is your estimate of 𝛽𝑋 biased or unbiased?

OLS5 <- lm(Y~X+K, data=df2)  
OLS5

##   
## Call:  
## lm(formula = Y ~ X + K, data = df2)  
##   
## Coefficients:  
## (Intercept) X K   
## 1.373 3.276 0.421

summary(OLS5)

##   
## Call:  
## lm(formula = Y ~ X + K, data = df2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3487 -0.5642 0.0152 0.5567 2.2696   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.37314 0.05662 24.25 <2e-16 \*\*\*  
## X 3.27593 0.08992 36.43 <2e-16 \*\*\*  
## K 0.42098 0.02189 19.23 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8062 on 997 degrees of freedom  
## Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998   
## F-statistic: 3.32e+06 on 2 and 997 DF, p-value: < 2.2e-16

Estimate of B1 is now biased.

#3(C)Provide an example of a collider that could cause bias in a regression.