ECO-R002 Lab 8

Task 1:

Use slide deck ExtraBonusLect03\_Causality\_Correlation\_MGM\_2017\_Primer for this Home Assignment. Note that the content on Selection Bias will prime you for the session on Wednesday, so you might as well take a quick read (if you have not read the section in Mostly Harmless).

**1.) What are the 4 main sources of endogeneity.**

**Omitted variables**

**Simultaneity**

**Selection bias**: Cannot determine the effect of treatment on a subject if sampling is biased

**Measurement error**

2.) Provide a formal definition of Selection Bias and show how it arises when (naively) comparing outcomes from two groups that are drawn from different populations.

Definition of Selection Bias: Selection bias occurs due to the heterogeneity between treatment state and the state is not observed in the treatment, rather than the treatment state, which results in alternative effect.

According to using slides, we let , otherwise, let be the health status of patient indexed by data determines whether he went to the hospital such that . However, we cannot observe the outcomes from and together, which are and , because a person could only choose go to the hospital or not. Therefore, we could learn about the effects of hospitalization by comparing the average health of those who were and were not hospitalized.

Causal Effect Selection Bias

3.) Derive the bias that results from omitting x2 in the model y = b1x1 + b2x2 + u?

=

## 4）Setup and require packages

setwd("C:/Users/22700/Desktop")  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

## Simulate

set.seed(0)  
ssize <- 1000  
x1 <- rnorm( n = ssize , sd = 3 )  
x2 <- rnorm( n = ssize , sd = 5 )  
y <- 2 + 3\*x1 + 5 \* x2 + rnorm(n = ssize, sd = 5)  
out.y.full <- lm( y ~ x1 + x2)  
out.y.x1.om <- lm( y ~ x1)  
out.y.x2.om <- lm( y ~ x2 )  
cor.test(x = x1, y = x2)

##   
## Pearson's product-moment correlation  
##   
## data: x1 and x2  
## t = -0.39878, df = 998, p-value = 0.6901  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.07455685 0.04940964  
## sample estimates:  
## cor   
## -0.01262211

## Output

stargazer(out.y.full, out.y.x1.om, out.y.x2.om,  
type = 'text', omit.stat = c('f','ser'), no.space=T)

##   
## ==========================================  
## Dependent variable:   
## -----------------------------  
## y   
## (1) (2) (3)   
## ------------------------------------------  
## x1 3.020\*\*\* 2.910\*\*\*   
## (0.050) (0.279)   
## x2 5.019\*\*\* 4.997\*\*\*   
## (0.029) (0.063)   
## Constant 2.344\*\*\* 1.717\*\* 2.198\*\*\*   
## (0.151) (0.835) (0.323)   
## ------------------------------------------  
## Observations 1,000 1,000 1,000   
## R2 0.971 0.098 0.865   
## Adjusted R2 0.970 0.097 0.864   
## ==========================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## #– Simulate x variables

set.seed(1)  
ssize <- 1000  
x1 <- rnorm( n = ssize , sd = 3 )  
x2 <- rnorm( n = ssize , mean = x1, sd = 5 )  
y <- 2 + 3\*x1 + 5 \* x2 + rnorm(n = ssize, sd = 5)  
out.y.full <- lm( y ~ x1 + x2)  
out.y.x1.om <- lm( y ~ x1)  
out.y.x2.om <- lm( y ~ x2 )  
cor.test(x = x1, y = x2)

##   
## Pearson's product-moment correlation  
##   
## data: x1 and x2  
## t = 19.065, df = 998, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4697453 0.5607242  
## sample estimates:  
## cor   
## 0.5166918

#– Output

stargazer(out.y.full, out.y.x1.om, out.y.x2.om,  
 type = 'text', omit.stat = c('f','ser'), no.space =T)

##   
## ==========================================  
## Dependent variable:   
## -----------------------------  
## y   
## (1) (2) (3)   
## ------------------------------------------  
## x1 3.060\*\*\* 8.136\*\*\*   
## (0.061) (0.271)   
## x2 5.022\*\*\* 5.830\*\*\*   
## (0.031) (0.050)   
## Constant 2.081\*\*\* 1.675\*\* 2.068\*\*\*   
## (0.163) (0.842) (0.305)   
## ------------------------------------------  
## Observations 1,000 1,000 1,000   
## R2 0.980 0.474 0.931   
## Adjusted R2 0.980 0.473 0.931   
## ==========================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## x2 centered around x1

set.seed(1)  
ssize <- 1000  
x1 <- rnorm( n = ssize , sd = 3 )  
x2 <- rnorm( n = ssize , mean = 1.33\*x1, sd = 5 )  
y <- 2 + 3\*x1 + 5 \* x2 + rnorm(n = ssize, sd = 5)  
out.y.full <- lm( y ~ x1 + x2)  
out.y.x1.om <- lm( y ~ x1)  
out.y.x2.om <- lm( y ~ x2 )  
cor.test(x = x1, y = x2)

##   
## Pearson's product-moment correlation  
##   
## data: x1 and x2  
## t = 25.29, df = 998, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5856438 0.6613196  
## sample estimates:  
## cor   
## 0.6249476

## Output

stargazer(out.y.full, out.y.x1.om, out.y.x2.om,  
type = 'text', omit.stat = c('f','ser'), no.space =T)

##   
## ==========================================  
## Dependent variable:   
## -----------------------------  
## y   
## (1) (2) (3)   
## ------------------------------------------  
## x1 3.052\*\*\* 9.786\*\*\*   
## (0.067) (0.271)   
## x2 5.022\*\*\* 5.911\*\*\*   
## (0.031) (0.043)   
## Constant 2.081\*\*\* 1.675\*\* 2.088\*\*\*   
## (0.163) (0.842) (0.285)   
## ------------------------------------------  
## Observations 1,000 1,000 1,000   
## R2 0.984 0.566 0.950   
## Adjusted R2 0.984 0.565 0.950   
## ==========================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## Compute the bias

out.y.full <- lm ( y ~ x1 + x2)  
coeffs.full <- coefficients(out.y.full)  
b2\_hat <- coeffs.full[3]  
b1\_hat <- coeffs.full[2]  
  
out.part.x2 <- lm ( x2 ~ x1)  
coeffs.part <- coefficients(out.part.x2)  
delta <- coeffs.part[2]  
  
bias <- delta\*b2\_hat  
bias

## x1   
## 6.733123

## Bias when we omit x1

out.part.x1 <- lm ( x1 ~ x2)  
coeffs.part <- coefficients(out.part.x1)  
delta <- coeffs.part[2]  
  
bias <- delta\*b1\_hat  
bias

## x2   
## 0.8892003

## 5)Follow all the other steps in the slide deck.

library(data.table)  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(sandwich)  
library(ggplot2)

dataset <- read.csv(  
 file = 'stork\_population.dsv' ,sep =';')  
names(dataset)

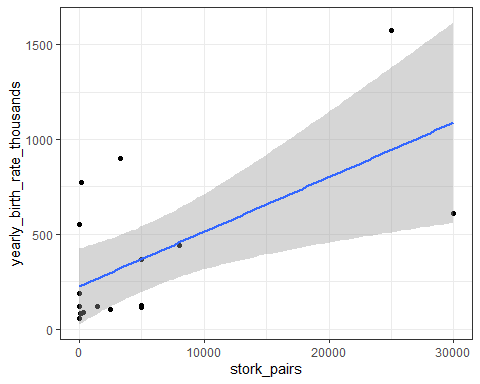
## [1] "country" "area\_km2"   
## [3] "stork\_pairs" "human\_population\_millions"   
## [5] "yearly\_birth\_rate\_thousands"

dataset[1:8,c('country','stork\_pairs','yearly\_birth\_rate\_thousands')]

## country stork\_pairs yearly\_birth\_rate\_thousands  
## 1 Albania 100 83  
## 2 Austria 300 87  
## 3 Belgium 1 118  
## 4 Bulgaria 5000 117  
## 5 Denmark 9 59  
## 6 France 140 774  
## 7 Germany 3300 901  
## 8 Greece 2500 106

qplot( data = dataset,x = stork\_pairs, y = yearly\_birth\_rate\_thousands , geom = 'point') + stat\_smooth(method='lm') + theme\_bw()

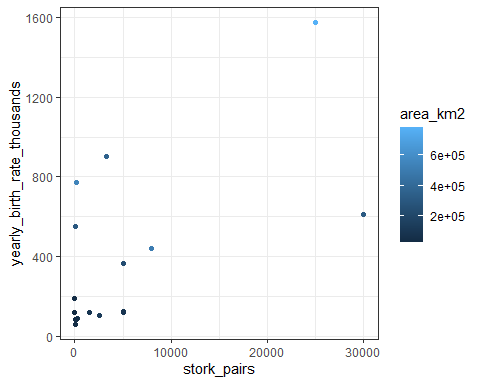
## `geom\_smooth()` using formula 'y ~ x'



out.lm <- lm( yearly\_birth\_rate\_thousands ~ stork\_pairs, data  
= dataset)  
stargazer(out.lm, type = 'text')

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## yearly\_birth\_rate\_thousands  
## -----------------------------------------------  
## stork\_pairs 0.029\*\*\*   
## (0.009)   
##   
## Constant 225.029\*\*   
## (93.561)   
##   
## -----------------------------------------------  
## Observations 17   
## R2 0.385   
## Adjusted R2 0.344   
## Residual Std. Error 332.185 (df = 15)   
## F Statistic 9.380\*\*\* (df = 1; 15)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

qplot( data = dataset,x = stork\_pairs  
 , y = yearly\_birth\_rate\_thousands  
 , geom = 'point', color = area\_km2) +  
 theme\_bw()



out.lm <- lm( yearly\_birth\_rate\_thousands ~ stork\_pairs +  
 area\_km2, data = dataset)  
stargazer(out.lm, type = 'text')

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## yearly\_birth\_rate\_thousands  
## -----------------------------------------------  
## stork\_pairs 0.006   
## (0.006)   
##   
## area\_km2 0.002\*\*\*   
## (0.0002)   
##   
## Constant -7.412   
## (56.702)   
##   
## -----------------------------------------------  
## Observations 17   
## R2 0.862   
## Adjusted R2 0.842   
## Residual Std. Error 162.744 (df = 14)   
## F Statistic 43.787\*\*\* (df = 2; 14)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

x <- dataset[,"stork\_pairs"]  
y <- dataset[,"area\_km2"]  
cor.test(x,y)

##   
## Pearson's product-moment correlation  
##   
## data: x and y  
## t = 2.7528, df = 15, p-value = 0.0148  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1367869 0.8291138  
## sample estimates:  
## cor   
## 0.5793423

## This is the omitted variable bias, because our first model does not include areas that should be in the model.