# Lab 8 - Homework

Nikita Grabher-Meyer

11/5/2020

## Task 1

### Question 1

*What are the four main sources of endogeneity?* Endogeneity occurs when a predictor variable (x) in a regression model is correlated with the error term (e). The main sources of endogeneity are four: 1) Omitted variable bias, which occurs when one or more relevant variables have not been included in the model 2) Measurement errors in one or more independent variables 3) Simultaneous equations, when both the dependent variable and a regressor are simultaneously determined, such as in a market equilibrium 4) Program evaluation with selection into treatment, where individuals select themselves into a group (such as self-selection bias)

### Question 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* Selection bias occurs when individuals or groups in a study differ systematically from the population of interest leading to a systematic error in an association or outcome. In other words, selection bias can arise in studies because groups of participants may differ in ways other than the interventions or exposures under investigation.

For example, we want to study the effects of working nights on the incidence of a certain health problem. We collect health information on a group of 9am-to-5om workers and a group of workers doing the same kind of work, but at night. We then measure the rates at which members of both groups reported the health problem. We might conclude that night work is associated with an increase in that problem. However the two groups we studied may have been very different to begin with. The people who worked nights may have been less skilled, with fewer employment options. In addition, their lower socioeconomic status would also be linked with more health risks—due to less healthy diets, less time and money for leisure activities and so on. So our findings may not be related to night work at all, but a reflection of the influence of other variables, such as their socioeconomic status.

### Question 3

*Derive the bias that results from omitting x2 in the model y = b1x1 + b2x2 + u* Let's assume that the true population model is Yi = b0 + b1Xi1 + b2Xi2 + ui. For example, y is log of hourly wage, x1 is education, and x2 is a measure of innate ability. Due to data

unavailability, we estimate the model by excluding x2. In particular, we decide to estimate: Yi = b0 + b1Xi1 + vi. The OLS estimate of b1 from the misspecified model is: E (b1^) = (X1'X1)-1 X1'Y = b1 + (X1'X1)-1 X1'x2b2 + (X1'X1)-1 X1'u = b1 * b2 (E(xi1 xi1))-1 E(xi1 xi2) Bias (b1^) = (X1'X1)-1 X1'X2b2

## Question 4

*a. Do the simulations that are shown from page 20 onwards to illustrate OVB. Can you replicate the bias?*

*When d=0, then b1°=b1^*

```
set.seed(1984)
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 = -0.20691, df = 998, p-value = 0.8361
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06851468  0.05546619
## sample estimates:
##          cor
## -0.006549411

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

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           2.972***  2.916***
##               (0.053)   (0.276)
## x2           5.020***             5.009***
##               (0.031)              (0.063)
## Constant     2.130***   1.659**  1.769***
##               (0.158)   (0.821)   (0.321)
## ----------------------------------------
## Observations   1,000     1,000     1,000
## R2             0.967     0.100     0.863
## Adjusted R2    0.967     0.100     0.862
## ========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

*When d!=0, then b1°!=b1^*

```
set.seed(1984)
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 = 18.314, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4536584 0.5465459
## sample estimates:
##       cor
## 0.5015462
```

```
library(stargazer)
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           2.951***  7.916***
```

```
##                (0.061)   (0.276)
## x2           5.020***            5.771***
##                (0.031)             (0.049)
## Constant     2.130***   1.659**  1.933***
##                (0.158)   (0.821)   (0.287)
## ----------------------------------------
## Observations   1,000     1,000     1,000
## R2             0.980     0.451     0.933
## Adjusted R2    0.980     0.451     0.933
## ========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

*Check that b1°=b1^ + b2^d°*

```
set.seed(1984)
y <- 2 + 3*x1 + 5 * x2 + rnorm(n = 1000, sd = 5)
out.y.full <- lm( y ~ x1 + x2)
out.y.incomp.x1 <- lm( y ~ x1 )
out.y.incomp.x2 <- lm( y ~ x2 )
out.x1.parti <- lm( x1 ~ x2 )
out.x2.parti <- lm( x2 ~ x1 )

library(stargazer)
stargazer(
out.x1.parti,
out.x2.parti,
out.y.full,
out.y.incomp.x1,
out.y.incomp.x2,
type = "text", omit.stat = c("f","ser"))

##
## ============================================================
##                           Dependent variable:
##              ---------------------------------------------
##                 x1       x2                 y
##                (1)      (2)      (3)      (4)      (5)
## ------------------------------------------------------------
## x2           0.254***          5.000***          6.187***
##              (0.014)           (0.000)           (0.065)
##
## x1                    0.989*** 4.667*** 9.611***
##                       (0.054)  (0.000)  (0.270)
##
## Constant      -0.067   -0.094  2.000***  1.531*  1.689***
##              (0.081)  (0.160)  (0.000)  (0.802)  (0.380)
##
## ------------------------------------------------------------
## Observations  1,000    1,000    1,000    1,000    1,000
## R2            0.252    0.252    1.000    0.559    0.901
## Adjusted R2   0.251    0.251    1.000    0.559    0.901
```

```
## ================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

*b. Instead of simply copying everything, try to center x2 around 1.33x1. What bias do you get? Can you predict the bias you will get now?*

```
set.seed(1984)
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 = 24.427, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5713587 0.6490660
## sample estimates:
##       cor
## 0.6116857

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            2.945***  9.566***
##                (0.067)   (0.276)
## x2            5.020***             5.856***
##                (0.031)              (0.042)
## Constant      2.130***   1.659**   1.985***
##                (0.158)   (0.821)   (0.270)
## ----------------------------------------
## Observations   1,000     1,000     1,000
## R2             0.983     0.546     0.951
## Adjusted R2    0.983     0.545     0.951
## ========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

*Bias when we omit x2*

```
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
## 6.62084
```

*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
## 0.8354589
```

## Question 5

*Follow all the other steps in the slide deck. Example 1: What type of endogeneity does this example highlight? Example 2: What type of endogeneity does this example highlight?* The first type of endogeneity is due to omitted variable bias. The second is due to simultaneity.

## Question 6

*Which source of Endogeneity do you consider to be the largest problem for your work. Explain why* My research project aims at assessing the effectiveness of an integrity/ethics training on corrupt attitudes and behaviour among young graduates. A possible source of endogeneity might be the omitted variable bias (students who might want to enroll on such a course might be inherently different from those who would not want to follow the course). Ideally we'll be trying to assign the treatment (training) randomly.