Multicollinearity and Omitted Variable Bias

Taro Mieno

AECN 896-003: Applied Econometrics

What variables to include or not include

You often

- face the decision of whether you should be including a particular variable or not: how do you make a right decision?
- miss a variable that you know is important because it is not simply available: what are the consequences?

Two important (intertwined) concepts you need to be aware of

- Multicollinearity
- Omitted Variable Bias

Multicollinearity

Multicollinearity

A phenomenon where two or more variables are highly correlated (negatively or positively) with each other (consequences?)

Omitted Variable Bias

Bias caused by not including (omitting) important variables in the model

Multicollinearity and Omitted Variable Bias

Consider the following model,

The model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

(your interest is in the impact of x_1 on y)

Multicollinearity and Omitted Variable Bias

Consider the following model,

The model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

(your interest is in the impact of x_1 on y)

Objectives

Using this simple model, we discuss what happens to the coefficient estimate on x_1 if you include/omit x_2

Multicollinearity and Omitted Variable Bias

Questions

1. What happens if $\beta_2 = 0$, but include x_2 that is not correlated with x_1?
2. What happens if $\beta_2 = 0$, but include x_2 that is highly correlated with x_1?
3. What happens if $\beta_2 \neq 0$, but omit x_2 that is not correlated with x_1?
4. What happens if $\beta_2 \neq 0$, but omit x_2 that is highly correlated with x_1?

Multicollinearity and Omitted Variable Bias

Questions

1. What happens if $\beta_2 = 0$, but include x_2 that is not correlated with x_1?
2. What happens if $\beta_2 = 0$, but include x_2 that is highly correlated with x_1?
3. What happens if $\beta_2 \neq 0$, but omit x_2 that is not correlated with x_1?
4. What happens if $\beta_2 \neq 0$, but omit x_2 that is highly correlated with x_1?

Key consequences of interest

- Is $\hat{\beta}_1$ still unbiased ($E[\hat{\beta}_1] = \beta_1$)?
- What happens to $Var(\hat{\beta}_1)$?

Case 1

Case 1

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Case 1

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

An example: randomized N trial

$$\text{Corn yield} = \beta_0 + \beta_1 N + \beta_2 \text{farmer's height} + u \quad (1)$$

Case 1

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Two estimating equations (EE)

$$(EE_1): \; y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$$
$$(EE_2): \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

What do you think is gonna happen? Any guess?

- $E[\hat{\beta}_1] = \beta_1$ in EE_1? (omitted variable bias?)
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Case 1: MC simulations

R code: Case 1

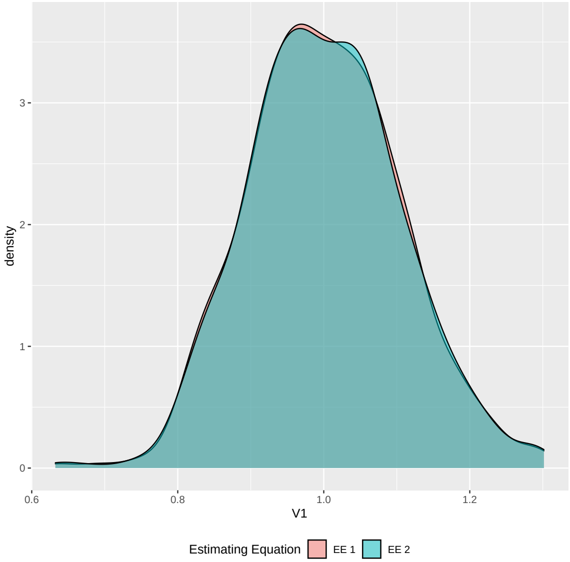
```
#--- preparation ---#
set.seed(37834)
N <- 100 # sample size
B <- 1000 # the number of iterations
estiamtes_strage <- matrix(0,B,2)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  x1 <- rnorm(N) # independent variable
  x2 <- rnorm(N) # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + 0*x2+ u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_ee1 <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1
  beta_ee2 <- lm(y~x1+x2,data=data)$coef['x1'] # OLS with EE2

  #--- store estimates ---#
  estiamtes_strage[i,1] <- beta_ee1
  estiamtes_strage[i,2] <- beta_ee2
}
```

Case 1: MC simulations

R code: Case 1 (continued)

```
b_ee1 <- data.table(bhat <- estiamtes_strage[,1],type='EE 1')
b_ee2 <- data.table(bhat <- estiamtes_strage[,2],type='EE 2')
plot_data <- rbind(b_ee1,b_ee2)
g_case_1 <- ggplot(data=plot_data) +
  geom_density(aes(x=V1,fill=type),alpha=0.5)+
  scale_fill_discrete(name='Estimating Equation')+
  theme(
    legend.position='bottom'
    )
```

Case 1: MC simulations



Case 1: Theoretical Investigation (Unbiasedness)

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + u_i(\beta_2 x_{2,i} + v_i)$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Case 1: Theoretical Investigation (Unbiasedness)

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + u_i (\beta_2 x_{2,i} + v_i)$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i | x_{1,i}] = 0?$

Case 1: Theoretical Investigation (Unbiasedness)

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + u_i(\beta_2 x_{2,i} + v_i)$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

Case 1: Theoretical Investigation (Unbiasedness)

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + u_i(\beta_2 x_{2,i} + v_i)$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i|x_{1,i}, x_{2,i}] = 0?$

Case 1: Theoretical Investigation (Unbiasedness)

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + u_i(\beta_2 x_{2,i} + v_i)$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i|x_{1,i}, x_{2,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with u.

Case 1: Theoretical Investigation (Variance)

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

where R_j^2 is the R^2 when you regress x_j on all the other covariates.

Case 1: Theoretical Investigation (Variance)

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

where R_j^2 is the R^2 when you regress x_j on all the other covariates.

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

R_j^2?

Case 1: Theoretical Investigation (Variance)

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

where R_j^2 is the R^2 when you regress x_j on all the other covariates.

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

Case 1: Theoretical Investigation (Variance)

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

where R_j^2 is the R^2 when you regress x_j on all the other covariates.

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$R_j^2?$ $\Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$R_j^2?$

Case 1: Theoretical Investigation (Variance)

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) = 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

where R_j^2 is the R^2 when you regress x_j on all the other covariates.

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$R_j^2? \Rightarrow$ 0 on average because $cor(x_1, x_2) = 0$

Case 3: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) = 0$, $\beta_2\ ne0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

σ^2

$Var(v_i)\ ?\ Var(u_i) \Rightarrow Var(v_i) = Var(u_i)$ because x_2 has no explanatory power

Summary

- If you include an irrelevant variable that has no explanatory power beyond x_1 and is not correlated with x_1 (EE_2), then the variance of the OLS estimator on x_1 will be the same as when you do not include x_2 as a covaraite (EE_1)
- If you omit an irrelevant variable that has no explanatory power beyond x_1 (EE_1) and is not correlated with x_1, then the the OLS estimator on x_1 is still unbiased

Case 2

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

An example: income in a job sector where only age matters

$$\text{Income} = \beta_0 + \beta_1 Age + \beta_2 \# \text{ of wrinkles} + u \tag{1}$$

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Two estimating equations (EE)

$$(EE_1): \; y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$$

$$(EE_2): \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

What do you think is gonna happen? Any guess?

- $E[\hat{\beta}_1] = \beta_1$ in EE_1?
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Case 2: MC simulations

R code: Case 2

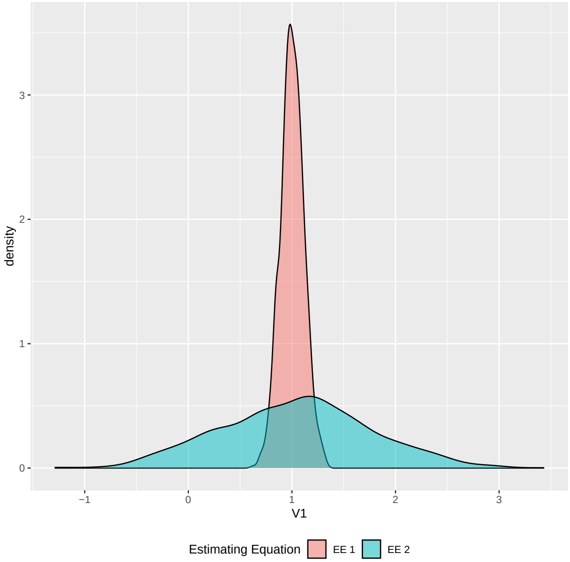
```r
#--- preparation ---#
set.seed(37834)
estiamtes_strage <- matrix(0,B,2)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  mu <- rnorm(N) # common term shared by x1 and x2
  x1 <- 0.1*rnorm(N) + 0.9*mu # independent variable
  x2 <- 0.1*rnorm(N) + 0.9*mu # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + 0*x2+ u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_ee1 <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1
  beta_ee2 <- lm(y~x1+x2,data=data)$coef['x1'] # OLS with EE2

  #--- store estimates ---#
  estiamtes_strage[i,1] <- beta_ee1
  estiamtes_strage[i,2] <- beta_ee2
}
```

Case 2: MC simulations

R code: Case 2 (continued)

```
b_ee1 <- data.table(bhat <- estiamtes_strage[,1],type='EE 1')
b_ee2 <- data.table(bhat <- estiamtes_strage[,2],type='EE 2')
plot_data <- rbind(b_ee1,b_ee2)
g_case_2 <- ggplot(data=plot_data) +
  geom_density(aes(x=V1,fill=type),alpha=0.5)+
  scale_fill_discrete(name='Estimating Equation')+
  theme(
    legend.position='bottom'
    )
```

Case 2: MC simulations



Case 2: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Case 2: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$

Case 2: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

Case 2: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$E[v_i | x_{1,i}] = 0? \Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i | x_{1,i}, x_{2,i}] = 0?$

Case 2: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i|x_{1,i}, x_{2,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with u.

Case 2: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

Case 2: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

R_j^2?

Case 2: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow 0$

Case 2: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$R_j^2?$

Case 2: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) \neq 0$, $\beta_2 = 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

R_j^2? $\Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

R_j^2? $\Rightarrow$ high because x_1 and x_2 are highly correlated.

Case 2: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) = 0$, $\beta_2\ ne 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

σ^2

$Var(v_i)\ ?\ Var(u_i) \Rightarrow Var(v_i) = Var(u_i)$ because x_2 has no explanatory power

Summary

- If you include an irrelevant variable that has no explanatory power beyond x_1, but is highly correlated with x_1 (EE_2), then the variance of the OLS estimator on x_1 is larger compared to when you do not include x_2 (EE_1)
- If you omit an irrelevant variable that has no explanatory power beyond x_1 (EE_1), but is highly correlated with x_1, then the the OLS estimator on x_1 is still unbiased

Case 3

Case 3

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Case 3

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

An example: Randomized N trial

$$yield = \beta_0 + \beta_1 N + \beta_2 Organic\ \ Matter + u \tag{1}$$

Case 3

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Two estimating equations (EE)

$$(EE_2): \; y_i = \beta_0 + \beta_1 x_{1,i} + v_i \; (\beta_2 x_{2,i} + u_i)$$

$$(EE_1): \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

What do you think is gonna happen? Any guess?

- $E[\hat{\beta}_1] = \beta_1$ in EE_2?
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Case 3: MC simulations

R code: Case 3

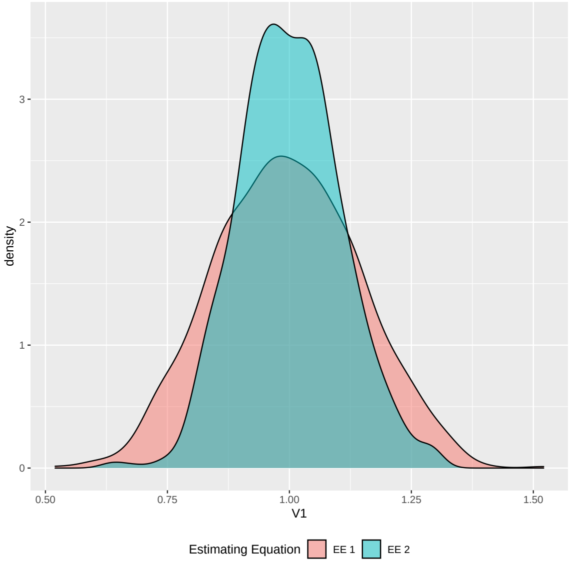
```r
#--- preparation ---#
set.seed(37834)
estiamtes_strage <- matrix(0,B,2)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  x1 <- rnorm(N) # independent variable
  x2 <- rnorm(N) # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + x2 + u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_ee1 <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1
  beta_ee2 <- lm(y~x1+x2,data=data)$coef['x1'] # OLS with EE2

  #--- store estimates ---#
  estiamtes_strage[i,1] <- beta_ee1
  estiamtes_strage[i,2] <- beta_ee2
}
```

Case 3: MC simulations

R code: Case 3 (continued)

```r
b_ee1 <- data.table(bhat <- estiamtes_strage[,1],type='EE 1')
b_ee2 <- data.table(bhat <- estiamtes_strage[,2],type='EE 2')
plot_data <- rbind(b_ee1,b_ee2)
g_case_3 <- ggplot(data=plot_data) +
  geom_density(aes(x=V1,fill=type),alpha=0.5)+
  scale_fill_discrete(name='Estimating Equation')+
  theme(
    legend.position='bottom'
    )
```

Case 3: MC simulations



Case 3: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$E[v_i | x_{1,i}] = 0$? $\Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

Case 3: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i | x_{1,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i | x_{1,i}, x_{2,i}] = 0?$

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i | x_{1,i}] = 0? \Rightarrow$ Yes, because x_1 is not correlated with either of x_2 and u.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i | x_{1,i}, x_{2,i}] = 0? \Rightarrow$ Yes, because x_1 is not correlated with u.

Case 3: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) = 0$, $\beta_2\ ne 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2\ ne 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

R_j^2?

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \; ne 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow 0$

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \; ne 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$R_j^2?$

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \ ne0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$R_j^2? \Rightarrow$ 0 on average because $cor(x_1, x_2) = 0$

Case 3: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \; ne0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

σ^2

$Var(v_i) \; ? \; Var(u_i) \Rightarrow Var(v_i) > Var(u_i)$ because x_2 has some explanatory power

Summary

- If you include a variable that has some explanatory power beyond x_1, but is not correlated with x_1 (EE_2), then the variance of the OLS estimator on x_1 is smaller compared to when you do not include x_2 (EE_1)
- If you omit an variable that has some explanatory power beyond x_1 (EE_1), but is not correlated with x_1, then the the OLS estimator on x_1 is still unbiased

Case 4

Case 4

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Case 4

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

An example: Income

$$income = \beta_0 + \beta_1 education + \beta_2 ability + u \tag{1}$$

Case 4

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Two estimating equations (EE)

$$(EE_1): \; y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$$
$$(EE_2): \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

What do you think is gonna happen? Any guess?

- $E[\hat{\beta}_1] = \beta_1$ in EE_2?
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Case 4: MC simulations

R code: Case 4

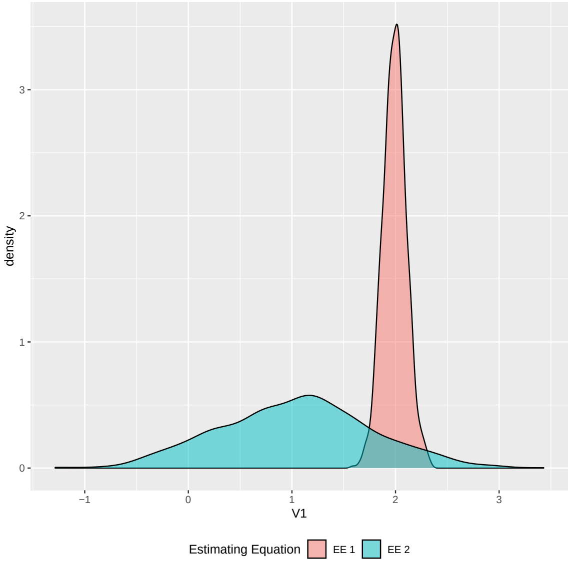
```r
#--- preparation ---#
set.seed(37834)
estiamtes_strage <- matrix(0,B,2)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  mu <- rnorm(N) # common term shared by x1 and x2
  x1 <- 0.1*rnorm(N) + 0.9*mu # independent variable
  x2 <- 0.1*rnorm(N) + 0.9*mu # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + 1*x2+ u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_ee1 <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1
  beta_ee2 <- lm(y~x1+x2,data=data)$coef['x1'] # OLS with EE2

  #--- store estimates ---#
  estiamtes_strage[i,1] <- beta_ee1
  estiamtes_strage[i,2] <- beta_ee2
}
```

Case 4: MC simulations

R code: Case 4 (continued)

```
b_ee1 <- data.table(bhat <- estiamtes_strage[,1],type='EE 1')
b_ee2 <- data.table(bhat <- estiamtes_strage[,2],type='EE 2')
plot_data <- rbind(b_ee1,b_ee2)
g_case_4 <- ggplot(data=plot_data) +
  geom_density(aes(x=V1,fill=type),alpha=0.5)+
  scale_fill_discrete(name='Estimating Equation')+
  theme(
    legend.position='bottom'
    )
```

Case 4: MC simulations



Case 4: Theoretical Investigation

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i | x_{1,i}] = 0?$ $\Rightarrow$ No, because x_1 is correlated with x_2.

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i | x_{1,i}] = 0?$ $\Rightarrow$ No, because x_1 is correlated with x_2.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i | x_{1,i}, x_{2,i}] = 0?$

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \neq 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$

$E[v_i|x_{1,i}] = 0?$ $\Rightarrow$ No, because x_1 is correlated with x_2.

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$E[u_i|x_{1,i}, x_{2,i}] = 0?$ $\Rightarrow$ Yes, because x_1 is not correlated with u.

Case 4: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) \neq 0$, $\beta_2\ ne0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \ ne 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

R_j^2?

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \ ne 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) \neq 0$, $\beta_2 \ ne 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$R_j^2?$

Case 4: Theoretical Investigation

True Model

true model : $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

($cor(x_1, x_2) \neq 0$, $\beta_2\ ne 0$, and $E[u_i|x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \ \ (\beta_2 x_{2,i} + u_i)$

$R_j^2? \Rightarrow$ 0

EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$

$R_j^2? \Rightarrow$ high because x_1 and x_2 are highly correlated.

Case 4: Theoretical Investigation

True Model

$$\text{true model}: \; y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$$

($cor(x_1, x_2) = 0$, $\beta_2 \neq 0$, and $E[u_i | x_{1,i}, x_{2,i}] = 0$)

Variance

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

σ^2

$Var(v_i)$? $Var(u_i) \Rightarrow Var(v_i) > Var(u_i)$ because x_2 has some explanatory power beyond x_1

The Magnitude of the Omitted Variable Bias

What do you expect?

- As β_2 gets larger (the more influential $x_{2,i}$), the magnitude of the bias on the coefficient estimator on $x_{1,i}$ gets (greater or smaller)
- As $cor(x_1, x_2)$ gets larger in magnitude, the magnitude of the bias on the coefficient estimator on $x_{1,i}$ gets (greater or smaller)

Magnitude of Bias: MC simulations

R code: Low impact of x_2

```
#--- preparation ---#
set.seed(37834)
estiamtes_strage <- rep(0,B)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  mu <- rnorm(N) # common term shared by x1 and x2
  x1 <- rnorm(N) + 0.5*mu # independent variable
  x2 <- rnorm(N) + 0.5*mu # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + x2+ u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_hat <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1

  #--- store estimates ---#
  estiamtes_strage[i] <- beta_hat
}
#--- bias ---#
mean(estiamtes_strage)-1
[1] 0.2022757
```

Magnitude of Bias: MC simulations

R code: High impact of x_2

```
#--- preparation ---#
set.seed(37834)
estiamtes_strage <- rep(0,B)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  mu <- rnorm(N) # common term shared by x1 and x2
  x1 <- rnorm(N) + 0.5*mu # independent variable
  x2 <- rnorm(N) + 0.5*mu # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + 3*x2+ u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_hat <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1

  #--- store estimates ---#
  estiamtes_strage[i] <- beta_hat
}
#--- bias ---#
mean(estiamtes_strage)-1
```

```
[1] 0.6056892
```

Summary

As β_2 gets larger (the more influential $x_{2,i}$), the magnitude of the bias on the coefficient estimator on $x_{1,i}$ gets gureater

Magnitude of Bias: MC simulations

R code: Low impact of x_2

```
#--- preparation ---#
set.seed(37834)
estiamtes_strage <- rep(0,B)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  mu <- rnorm(N) # common term shared by x1 and x2
  x1 <- rnorm(N) + 0.5*mu # independent variable
  x2 <- rnorm(N) + 0.5*mu # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + x2+ u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_hat <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1

  #--- store estimates ---#
  estiamtes_strage[i] <- beta_hat
}
#--- bias ---#
mean(estiamtes_strage)-1
```

```
[1] 0.2022757
```

Magnitude of Bias: MC simulations

R code: High impact of x_2

```r
#--- preparation ---#
set.seed(37834)
estiamtes_strage <- rep(0,B)

for (i in 1:B){ # iterate the same process B times
  #--- data generation  ---#
  mu <- rnorm(N) # common term shared by x1 and x2
  x1 <- rnorm(N) + 2*mu # independent variable
  x2 <- rnorm(N) + 2*mu # independent variable
  u <- rnorm(N) # error
  y <- 1 + x1 + x2+ u # dependent variable
  data <- data.table(y=y,x1=x1,x2=x2)

  #--- OLS ---#
  beta_hat <- lm(y~x1,data=data)$coef['x1'] # OLS with EE1

  #--- store estimates ---#
  estiamtes_strage[i] <- beta_hat
}
#--- bias ---#
mean(estiamtes_strage)-1
```

```
[1] 0.8034367
```

Summary

As $cor(x_1, x_2)$ gets larger in magnitude, the magnitude of the bias on the coefficient estimator on $x_{1,i}$ gets greater

Omitted Variable Bias: Theoretical Investigation

- true model: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$
- EE1: $y_i = \beta_0 + \beta_1 x_{1,i} + v_i \;\; (\beta_2 x_{2,i} + u_i)$
  - Let $\tilde{\beta}_1$ denote the estimator of β_1
- EE2: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$
  - Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the estimator of β_1 and β_2
- x_1 on x_2: $x_{1,i} = \sigma_0 + \sigma_1 x_{2,i} + \mu_i$
  - Let $\tilde{\sigma_1}$ denote the estimator of σ_1

Bias

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \times \tilde{\sigma_1} \\ \Rightarrow E[\tilde{\beta}_1] &= E[\hat{\beta}_1] + E[\hat{\beta}_2] \times \tilde{\sigma_1} \\ &= \beta_1 + \beta_2 \tilde{\sigma_1} \;\; \text{(bias)} \end{aligned}$$

Direction of the Bias

Bias

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2\tilde{\sigma}_1 \text{ (bias)}$$

Direction of Bias

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Magnitude of Bias

Obvious

Dropping a variable

Question

Should I drop x_2 because it is highly correlated with x_1, which makes the estimation of the coefficient on x_1 very imprecise?

Answer

No!! Your coefficient estimation on x_1 would be biased unless you know that x_2 has no explanatory power on y beyond x_1

Multicolinearity between control variables

Question

Should you be concerned about multicolinearity between control variables (variables you are not interested in)?

Answer

No, because you don't care about the precise estimationf of the coefficient on control variables individually.

Summary

- Whether you should include a variable (x_2) depends crucially on how x_2 is related with x_1 and how influential x_2 is
- If x_2 are extremely highly correlated with x_1 and x_2 has big impacts on y, then you are doomed: trade-off: severe bias or extremely variable estimator
- If x_2 are extremely highly correlated with x_1, but x_2 has very small impacts on y, then you might be better off omitting x_2 (small bias, large gain in efficiency)