Omitted Variable Bias and Multicollinearity

AECN 396/896-002

What variables to include or not

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 concepts you need to be aware of:

- Multicollinearity
- Omitted Variable Bias

Multicollinearity and Omitted Variable Bias

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,

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

Your interest is in estimating the impact of x_1 on y.

Objectives:

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

Questions we tackle to answer

The model:

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

Question 1:

What happens if $\beta_2 = 0$, but include x_2 that is not correlated with x_1?

Question 2:

What happens if $\beta_2 = 0$, but include x_2 that is highly correlated with x_1?

Question 3:

What happens if $\beta_2 \neq 0$, but omit x_2 that is not correlated with x_1?

Question 4:

What happens if $\beta_2 \neq 0$, but omit x_2 that is highly correlated with x_1?

Key consequences of interest

- Is $\hat{\beta}_1$ unbiased, that is $E[\hat{\beta}_1] = \beta_1$?
- $Var(\hat{\beta}_1)$? (how accurate the estimation of $\hat{\beta}_1$ is)

Case 1

Case 1

True Model:

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

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

Example:

$\text{corn yield} = \beta_0 + \beta_1 \times N + \beta_2 \text{farmers' height} + u$

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?

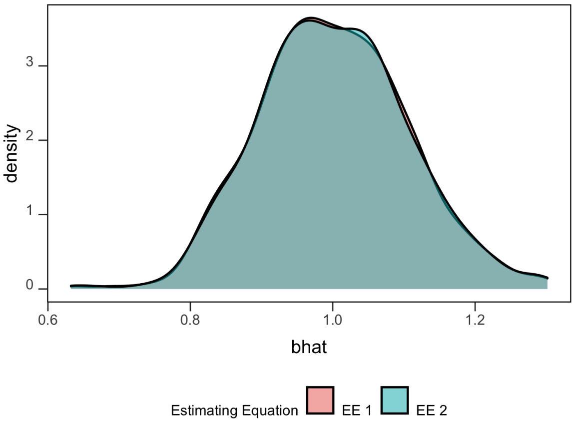
- $E[\hat{\beta}_1] = \beta_1$ in EE_1? (omitted variable bias?)
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Monte Carlo Simulation

```r
#* load packages
# library(fixest)
# library(data.table)

#-------------------------
# Monte Carlo Simulation
#-------------------------
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.frame(y = y, x1 = x1, x2 = x2)

  #--- OLS ---#
  beta_ee1 <- feols(y ~ x1, data = data)$coefficient["x1"] # OLS with EE1
  beta_ee2 <- feols(y ~ x1 + x2, data = data)$coefficient["x1"] # OLS with EE2

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

#-------------------------
# Visualize the results
#-------------------------
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 = bhat, fill = type), alpha = 0.5) +
  scale_fill_discrete(name = "Estimating Equation") +
  theme(legend.position = "bottom")
```

MC Results

```
g_case_1
```



Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Yes, because x_1 is not correlated with either of x_2 and u.

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Yes, because x_1 and x_2 are not correlated with u (by assumption).

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

0 because there are no other variables included in the model.

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

0 on average because $cor(x_1, x_2) = 0$

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

They are the same because $\beta_2 = 0$, meaning $u = v$.

Summary

- If you include an irrelevant variable that has no explanatory power beyond x_1 and is not correlated with x_1 (EE2), then the variance of the OLS estimator on x_1 will be the same as when you do not include x_2 as a covariate (EE1)

- If you omit an irrelevant variable that has no explanatory power beyond x_1 (EE1) and is not correlated with x_1, then the the OLS estimator on x_1 is still unbiased

Case 2

Case 2

True Model:

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

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

Example:

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

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?

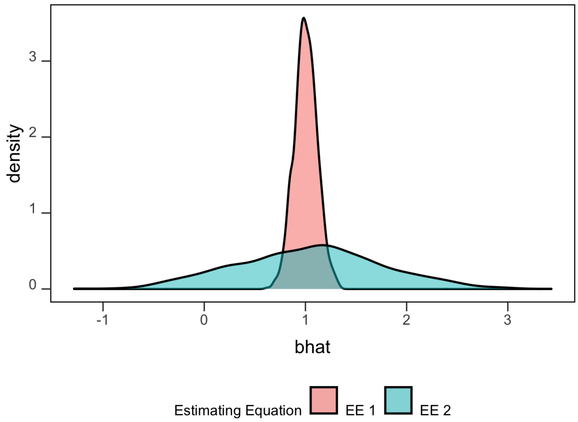
- $E[\hat{\beta}_1] = \beta_1$ in EE_1? (omitted variable bias?)
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Monte Carlo Simulation

```r
#--------------------------
# Monte Carlo Simulation
#--------------------------
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 ---#
  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.frame(y = y, x1 = x1, x2 = x2)

  #--- OLS ---#
  beta_ee1 <- feols(y ~ x1, data = data)$coefficient["x1"] # OLS with EE1
  beta_ee2 <- feols(y ~ x1 + x2, data = data)$coefficient["x1"] # OLS with EE2

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

#--------------------------
# Visualize the results
#--------------------------
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 = bhat, fill = type), alpha = 0.5) +
  scale_fill_discrete(name = "Estimating Equation") +
  theme(legend.position = "bottom")
```

MC Results

```
g_case_2
```



Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Yes, because $\beta_2 = 0$, meaning that x_2 is actually not part of the error term (u).

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Yes, because x_1 and x_2 are not correlated with u (by assumption).

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

0 because there are no other variables included in the model.

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Very high because x_1 and x_2 are highly correlated!

So, the estimation accuracy of β_1 in EE_2 is much lower that in EE_1!.

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

They are the same because $\beta_2 = 0$, meaning $u = v$.

Summary

- If you include an irrelevant variable that has no explanatory power beyond x_1, but is highly correlated with x_1 (EE2), then the variance of the OLS estimator on x_1 is larger compared to when you do not include x_2 (EE1)
- If you omit an irrelevant variable that has no explanatory power beyond x_1 (EE1), but is highly correlated with x_1, then the the OLS estimator on x_1 is still unbiased

Case 3

Case 3

True Model:

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

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

Example: Randomized N trial

$\text{corn yield} = \beta_0 + \beta_1 \times N + \beta_2 \times \text{organic matter} + u$

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?

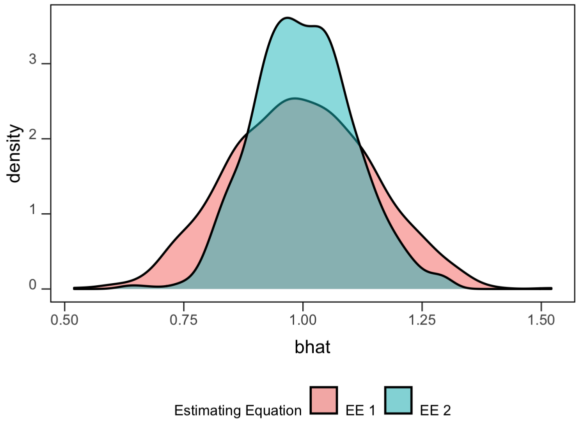
- $E[\hat{\beta}_1] = \beta_1$ in EE_1? (omitted variable bias?)
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Monte Carlo Simulation

```
#-------------------------
# Monte Carlo Simulation
#-------------------------
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 + x2 + u # dependent variable
  data <- data.frame(y = y, x1 = x1, x2 = x2)

  #--- OLS ---#
  beta_ee1 <- feols(y ~ x1, data = data)$coefficient["x1"] # OLS with EE1
  beta_ee2 <- feols(y ~ x1 + x2, data = data)$coefficient["x1"] # OLS with EE2

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

#-------------------------
# Visualize the results
#-------------------------
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 = bhat, fill = type), alpha = 0.5) +
  scale_fill_discrete(name = "Estimating Equation") +
  theme(legend.position = "bottom")
```

MC Results

```
g_case_3
```



Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Yes, because x_1 and x_2 are not correlated.

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Yes, because x_1 and x_2 are not correlated with u (by assumption).

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

0 because there are no other variables included in the model.

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Very high because x_1 and x_2 are highly correlated!

0 on average because x_1 and x_2 are note correlated.

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

$Var(v_i) > Var(u_i)$ because $\beta_2 x_2$ (non-zero) is part of v_i on top of u_i.

So, the estimation of β_1 is more efficient in EE_2 than in EE_1.

Summary

- If you include a variable that has some explanatory power beyond x_1, but is not correlated with x_1 (EE2), then the variance of the OLS estimator on x_1 is smaller compared to when you do not include x_2 (EE1)
- If you omit an variable that has some explanatory power beyond x_1 (EE1), but is not correlated with x_1, then the the OLS estimator on x_1 is still unbiased

Case 4

Case 4

True Model:

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

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

Example

$\text{income} = \beta_0 + \beta_1 \times education + \beta_2 \times \text{ability} + u$

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?

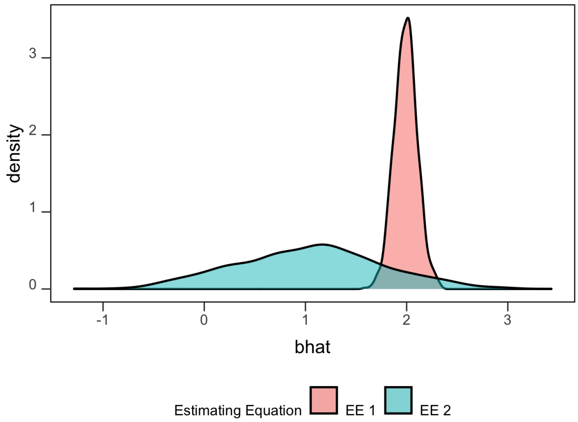
- $E[\hat{\beta}_1] = \beta_1$ in EE_1? (omitted variable bias?)
- How does $Var(\hat{\beta}_1)$ in EE_2 compared to its counterpart in EE_1?

Monte Carlo Simulation

```
#--------------------------
# Monte Carlo Simulation
#--------------------------
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 ---#
  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
  data <- data.frame(y = y, x1 = x1, x2 = x2)

  #--- OLS ---#
  beta_ee1 <- feols(y ~ x1, data = data)$coefficient["x1"] # OLS with EE1
  beta_ee2 <- feols(y ~ x1 + x2, data = data)$coefficient["x1"] # OLS with EE2

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

#--------------------------
# Visualize the results
#--------------------------
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 = bhat, fill = type), alpha = 0.5) +
  scale_fill_discrete(name = "Estimating Equation") +
  theme(legend.position = "bottom")
```

MC Results

```
g_case_4
```



Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

No, because x_1 and x_2 are correlated.

So, the estimation of β_1 in EE_1 is biased!

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Theoretical Insights: Bias

True Model:

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

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

The estimated model

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

Question:

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

Yes, because x_1 and x_2 are not correlated with u (by assumption).

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

0 because there are no other variables included in the model.

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

The estimated model

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

Question:

R_j^2?

Very high because x_1 and x_2 are highly correlated!

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

Theoretical Insights: Variance of $\hat{\beta}_1$

True Model:

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

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

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.

Two models:

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

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

Question:

Which in EE_1 and EE_2 is σ^2 larger?

$Var(v_i) > Var(u_i)$ because $\beta_2 x_2$ (non-zero) is part of v_i on top of u_i.

Estimation efficiency

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.

Summarizing the results about the components of $Var(\hat{\beta}_j)$,

- R_j^2 is very high for EE_2 because x_1 and x_2 are highly correlated, while it is 0 for EE_1.
- $Var(v_i) > Var(u_i)$ because $\beta_2 x_2$ (non-zero) is part of v_i on top of u_i.

So, whether EE_1 is more efficient than EE_2 or not is ambiguous. It depends on

- the degree of the correlation between x_1 and x_2
- the magnitude of β_2

Summary

- There exists bias-variance trade-off when independent variables are both important (their coefficients are non-zero) and they are correlated
- Economists tend to opt for unbiasedness

Omitted Variable Bias

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 from this model

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

Relationship between x_1 and x_2

$$x_{1,i} = \sigma_0 + \sigma_1 x_{2,i} + \mu_i$$

Let $\tilde{\sigma_1}$ denote the estimator of σ_1

Then,

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \tilde{\sigma_1}$$

where $\beta_2 \tilde{\sigma_1}$ is the bias.

Magnitude and direction of bias

Then,

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2\tilde{\sigma_1}$$

where $\beta_2\tilde{\sigma_1}$ is the bias.

Direction of bias

- $Cor(x_1, x_2) > 0$ and $\beta_2 > 0$, then $bias > 0$
- $Cor(x_1, x_2) > 0$ and $\beta_2 < 0$, then $bias < 0$
- $Cor(x_1, x_2) < 0$ and $\beta_2 > 0$, then $bias < 0$
- $Cor(x_1, x_2) < 0$ and $\beta_2 < 0$, then $bias > 0$

Magnitude of bias

- The greater the correlation between x_1 and x_2, the greater the bias
- The greater β_1 is, the greater the bias

Direction of bias: Practice

Example 1

$$\text{corn yield} = \alpha + \beta \cdot N + (\gamma \cdot \text{soil erodability} + \mu)$$

- Famers tend to apply more nitrogen to the field that is more erodible to compensate for loss of nutrient due to erosion
- Soil erodability affects corn yield negatively $(\gamma < 0)$

What is the direction of bias on $\hat{\beta}$?

Example 2

$$\text{house price} = \alpha + \beta \cdot \text{dist to incinerators} + (\gamma \cdot \text{dist to city center} + \mu)$$

- The city planner placed incinerators in the outskirt of a city to avoid their potentially negative health effects
- Distance to city center has a negative impact on house price $(\gamma < 0)$

What is the direction of bias on $\hat{\beta}$?

Example 3

$$\text{groundwater use} = \alpha + \beta \cdot \text{precipitation} + (\gamma \cdot \text{center pivot} + \mu)$$

groundwater use: groundwater use by a farmer for irrigated production

center pivot: 1 if center pivot is used, 0 if flood irrigation (less effective) is used

- Farmers who have relatively low precipitation during the growing season tend to adopt center pivot more
- center pivot applied water more efficiently than flood irrigation $(\gamma < 0)$

What is the direction of bias on $\hat{\beta}$?

So when does it help to know the direction of bias

When the direction of the bias is the opposite of the expected coefficient on the variable of interest, you can claim that even after suffering from the bias, you are still seeing the impact of the variable interest. So, it is a strong evidence that you would have had an even stronger estimated impact.

Example 1

$$\text{groundwater use} = \alpha + \beta \cdot \text{precipitation} + (\gamma \cdot \text{center pivot} + \mu)$$

- The true β is -10 (you do not observe this)
- The bias on $\hat{\beta}$ is 5 (you do not observe this)
- $\hat{\beta}$ is -5 (you only observe this)

You believe the direction of bias is positive (you need provide reasoning behind your belief), and yet, the estimated coefficient is still negative. So, you can be quite confident that the sign of the imapct of precipitation is negative. You can say your estimate is a conservative estimate of the impact of precipitation on groudwater use.

Example 2

$$\text{house price} = \alpha + \beta \cdot \text{dist to incinerators} + (\gamma \cdot \text{dist to city center} + \mu)$$

- The true β is -10 (you do not observe this)
- The bias on $\hat{\beta}$ is -5 (you do not observe this)
- $\hat{\beta}$ is -15 (you only observe this)

You believe the direction of bias is negative, and the estimated coefficient is negative. So, unlike the case above, you cannot be confident that $\hat{\beta}$ would have been negative if it were not for the bias (by observing dist to city center and include it as a covariate). It is very much possible that the degree of bias is so large that the estimated coefficient turns negative even though the true sign of β is positive. In this case, there is nothing you can do.