Week 7 Lab

Omitted Variable Bias and F-Tests

Micaela Wood
02/17/2022

Today

- Omitted variable bias with simulated data
- Omitted variable bias with real data
- Irrelevent Variable with real data
- F Tests and adjusted R^2

Omitted Variable Bias

When are we concered about this?

- If a variable effects y but has no correlation with x...
  - We have no concerns
- If a variable effects y and has correlation with x...
  - A3: Exogeneity of the independent variables is violated
  - We should be concerned about bias

Omitted Variable Bias Example

Suppose you are interested in studying the effect of high school GPA on SAT scores. What are some variables that could lead to omitted variable bias if we ran this regression:

$$SAT = \beta_0 + \beta_1 GPA + \varepsilon$$

- Parent's Income
- Public vs. Private School
- Parent's Job
- Where you live(city vs. country)
- Quality of Teachers
- Tutoring

Data Example

Let's see what omitted variable bias looks like with some simulated data.

```r
#library(tidyverse, huxtable, broom)
# Choose an observation count (number of rows)
n <- 1000
set.seed(1245)
# Generate data in a tibble
data_sim = tibble(
  e1 = rnorm(n, sd = 30),
  e2 = rnorm(n, sd = 20),
  e3 = rnorm(n, sd = 10),
  x = runif(n, min = 0, max = 10),
  y = runif(n, min = 10, max = 20),
  z = 20 - 0.3*y + 3*x + e1,
  a = 6 + 2*x -1.5*y + e2,
  b = 10 - .5*y + 4*z + e3
)
```

Data Example

- First we will run regression for a.

- Which of these models will have omitted variable bias?

    1. $a = \beta_0 + \beta_1 x$

    2. $a = \beta_0 + \beta_1 y$

    3. $a = \beta_0 + \beta_1 x + \beta_2 y$

- **None** be x and y are not correlated

Lets run the three models and see what happens.

```
lm1 = lm(data = data_sim, a ~ x)
lm2 = lm(data = data_sim, a ~ y)
lm3 = lm(data = data_sim, a ~ x + y)
```

```
huxreg(tidy(lm1), tidy(lm2), tidy(lm3))
```

	(1)	(2)	(3)
(Intercept)	-17.605 ***	11.053 **	1.059
	(1.293)	(3.538)	(3.558)
x	2.056 ***		2.069 ***
	(0.225)		(0.221)
y		-1.223 ***	-1.245 ***
		(0.231)	(0.222)
N	0	0	0

Now let's think about models for b. Will any of these have omitted variable bias?

1. $b = \beta_0 + \beta_1 y$
2. $b = \beta_0 + \beta_2 z$
3. $b = \beta_0 + \beta_1 y + \beta_2 z$

Model 1 and 2 may be **Biased** because y and z are correlated with each other as well as with b

Let's run these regression and see what happens

```
lm4 = lm(data = data_sim, b ~ y)
lm5 = lm(data = data_sim, b ~ z)
lm6 = lm(data = data_sim, b ~ y + z)
```

```
huxreg(tidy(lm4), tidy(lm5), tidy(lm6))
```

	(1)	(2)	(3)
(Intercept)	110.816 ***	2.238 ***	9.134 ***
	(21.000)	(0.451)	(1.725)
y	0.837		-0.461 ***
	(1.371)		(0.111)
z		3.997 ***	3.999 ***
		(0.010)	(0.010)
N	0	0	0

Now let's a new data set that includes variables for earnings and demographics

```r
library(haven)
nsw <- read_dta("nsw.dta")
```

This data looks at how a training program for low income individuals effected peoples earnings. People with low enough income were but into the study. Some individuals were randomly assigned to receive treatment while others did not.

What should we model to see the effect of the program on earnings?

$$earnings = \beta_0 + \beta_1 training$$

Lets run this model

```
lm7 = lm(data = nsw, re78 ~ treat)
huxreg(tidy(lm7))
```

	(1)
(Intercept)	5090.048 ***
	(302.783)
treat	886.304
	(472.086)
N	0

*** p < 0.001; ** p < 0.01; * p < 0.05.

Are there any variables that may cause concerns of omitted variable bias?

Age, education, race, and earning before the program may be correlated with earnings and if the person qualified for the program.

Let's run a new regression with these controls

```
lm8 = lm(data = nsw,
         re78 ~ treat + re75 + age +
         education + black + hispanic)
```

Irrelevent Variables

- To avoid omitted variable bias you may be tempted to add every variable you can think of
- Not only would this be very time consuming, but it can cause us to lose all significance.

Irrelevent Variables

The first way to know if you are adding an irrelevant variable is to look at the adjusted R squared, $\bar{R}^2$

Let's add an irrelevant variable to our simulated data and see what happens to our R^2s

```r
data_sim = data_sim %>% mutate(
  c = rnorm(n, mean = 10, sd = 5)
)
lm9 = lm(data = data_sim, a ~ c + y)
```

F Tests

Lets go back to the model $\boldsymbol{b} = \beta_0 + \beta_1 \boldsymbol{y} + \beta_2 \boldsymbol{z}$

What if we think that $\beta_1 = \beta_2$.

--

We can make a second, restricted, model to test this hypothesis. What should the model be?

$$\beta_1 = \beta_2$$

$$b = \beta_0 + \beta_1 y + \beta_2 z$$

$$b = \beta_0 + \beta_1 y + \beta_1 z$$

$$b = \beta_0 + \beta_1 (y + z)$$

Let's make a variable equal to y + z.

```
data_sim = data_sim %>%
  mutate(plus = y + z)
```

Now we can run the new model and construct an **F-Test**.

```r
resticted = lm(data = data_sim, b ~ ???)

unrestricted = lm(data = data_sim, b ~ ???)
```

```
restricted = lm(data = data_sim, b ~ plus)

unrestricted = lm(data = data_sim, b ~ y + z)
```

The F test is

$$F_{q,n-k-1} = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k-1)}$$

So we should first calculate the RSS

```
res_r = unname(resid(???))
res_u = unname(resid(???))

res_r_sq = ???
res_u_s = ???

rss_r = ???
rss_u = ???
```

The F test is

$$F_{q,n-k-1} = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k-1)}$$

So we should first calculate the RSS

```
res_r = unname(resid(resticted))
res_u = unname(resid(unrestricted))

res_r_sq = res_r^2
res_u_sq = res_u^2

rss_r = sum(res_r_sq)
rss_u = sum(res_u_sq)
```

```r
rss_r
```

```
#> [1] 263205.6
```

```r
rss_u
```

```
#> [1] 101832.4
```

The F test is

$$F_{q,n-k-1} = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k-1)}$$

Now lets save q, n, and k

```
q = ???
n = ???
k = ???
```

Our Hypothesis is $H_0 : \beta_1 = \beta_2$

The F test is

$$F_{q,n-k-1} = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k-1)}$$

Now lets save q, n, and k

```
q = 1
#n is the same as when we made the data so we don't need to save it here
k = 2
```

Now we can find our F statistic

```
F = ((rss_r - rss_u)/q)/(rss_u/(n-k-1))
F
```

```
#> [1] 1579.94
```

Now we need to use an F Table to find our comparison.

$F_{q,n-k-1}$ = ???

$F_{1,97}$ = 3.936

Now we have everything we need.

$H_0 : \beta_1 = \beta_2$

$F_{1,97}$ = 3.936

```
F
```

```
#> [1] 1579.94
```

Do we reject or fail to reject the null hypothesis?

Reject the null because 1579.9 > 3.936

Bonus

Repeated Simulations

- When we make simulated data, we have only been doing one draw so far.
- What if we didn't actually have omitted variable bias, but a bad draw?
- We can do an iterated simulation to get many estimates for $\hat{\beta}$ to test this.

Step 1

Generate Data

We already did this at the beginning of class

```r
n <- 1000
set.seed(1245)
# Generate data in a tibble
data_sim = tibble(
  e1 = rnorm(n, sd = 30),
  e3 = rnorm(n, sd = 10),
  x = runif(n, min = 0, max = 10),
  y = runif(n, min = 10, max = 20),
  z = 20 - 0.3*y + 3*x + e1,
  b = 10 - .5*y + 4*z + e3
)
```

Step 2

Run Regressions

For this we will run 2 models:

- $b = \beta_0 + \beta_1 y$

- $b = \beta_0 + \beta_1 y + \beta_2 z$

```
lmy <- lm(data = data_sim, b ~ y)
lmyz <- lm(data = data_sim, b ~ y + z)
```

Step 3

Put results into a table

```r
bind_rows(tidy(lmy), tidy(lmyz)) %>%
    filter(term == "y") %>%
    mutate(
      model = c("without z", "with z")
      #i = iter, we will use this line in a minute
      )
```

term	estimate	std.error	statistic	p.value	model
y	-1.78	1.35	-1.31	0.19	without z
y	-0.577	0.109	-5.31	1.36e-07	with z

Step 4

Wrap steps 1, 2, and 3, into a function

```r
sim_fun <- function(iter){
  data_sim = tibble(
  e1 = rnorm(n, sd = 30),
  e3 = rnorm(n, sd = 10),
  x = runif(n, min = 0, max = 10),
  y = runif(n, min = 10, max = 20),
  z = 20 - 0.3*y + 3*x + e1,
  b = 10 - .5*y + 4*z + e3
  )

  lmy <- lm(data = data_sim, b ~ y)
  lmyz <- lm(data = data_sim, b ~ y + z)

  bind_rows(tidy(lmy), tidy(lmyz)) %>%
    filter(term == "y") %>%
    mutate(model = c("without z", "with z"), i = iter)
}
```

Step 5

Run the simulation multiple times and combine each result into one big table

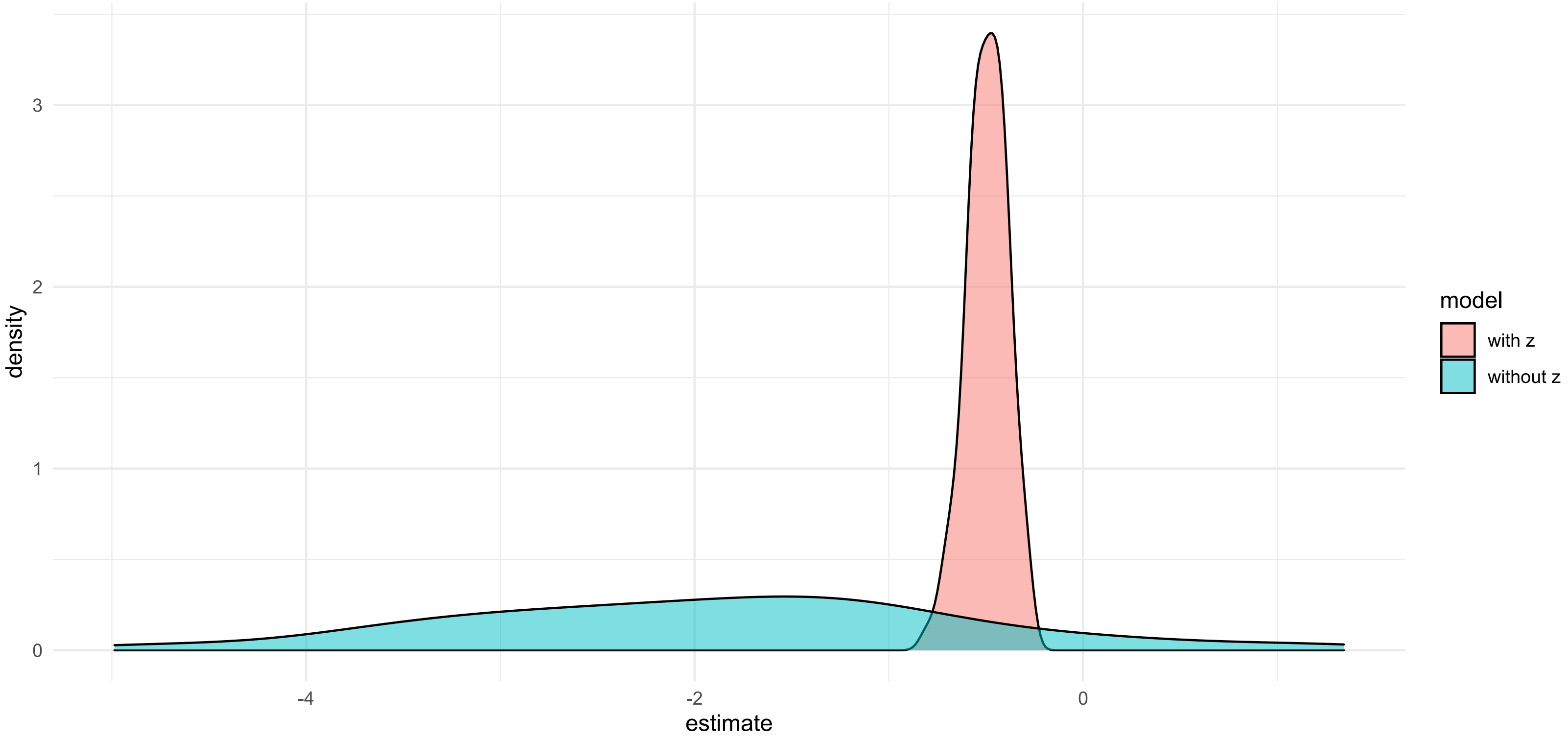
```
sim_list = map(1:100, sim_fun)
sim_df = bind_rows(sim_list)
head(sim_df)
```

term	estimate	std.error	statistic	p.value	model	i
y	-3.75	1.36	-2.75	0.00611	without z	1
y	-0.554	0.105	-5.25	1.81e-07	with z	1
y	-0.912	1.4	-0.653	0.514	without z	2
y	-0.499	0.114	-4.39	1.24e-05	with z	2
y	-2.25	1.42	-1.58	0.115	without z	3

Graph Results

Now we can use the density plot to graph all the different β

```
ggplot(data = sim_df, aes(x = estimate, fill = model))+
  geom_density(alpha = 0.5) + theme_minimal()
```



Another Example

One reason economists like to simulate results is to replicate outcomes from papers.

I just made a simulation based on how job training program could potentially hurt people who didn't get to receive the training.

This is another area there is bias, but not from omitted variables.