Omitted Variables Bias (OVB)

EC 339

Marcio Santetti
Fall 2022

Motivation

Well-specified models

Recall **CLRM Assumption I**:

> *"The regression model is linear, correctly specified, and has an additive stochastic error term."*

The hardest part regarding this assumption is to have a **well-specified model**.

Suppose we have the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

- How can we evaluate whether this is a well-specified model?
- Does it have the appropriate functional form?
- Is this model in accordance with economic theory?

Well-specified models

In fact, we can never know for sure if we have the most appropriate model.

Theory is always (and will always be) the best guide.

In addition, we must always **visualize** our data, knowing it better in order to define the model's functional form.

- **A different functional form may also be an omitted variable!**
- For instance, if the 'true' model contains a squared term, in case we omit it from our sample regression model, it will be **misspecified**.

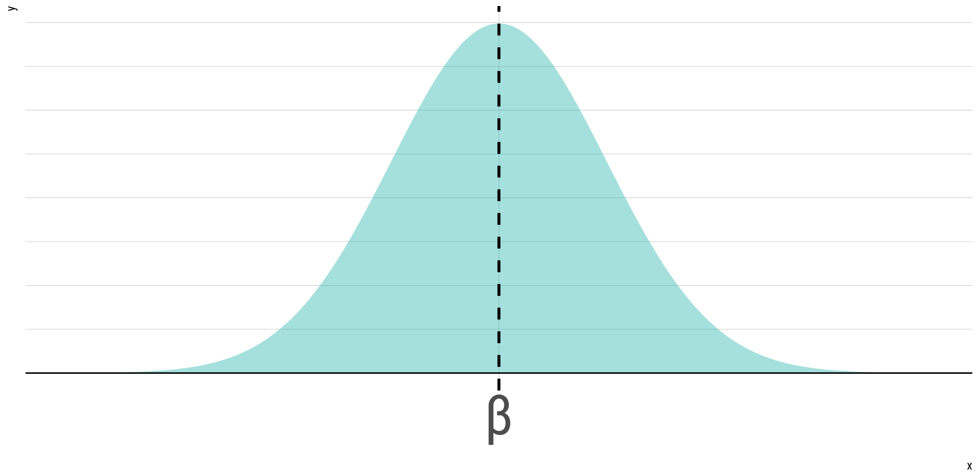
The nature of the problem

Recalling bias

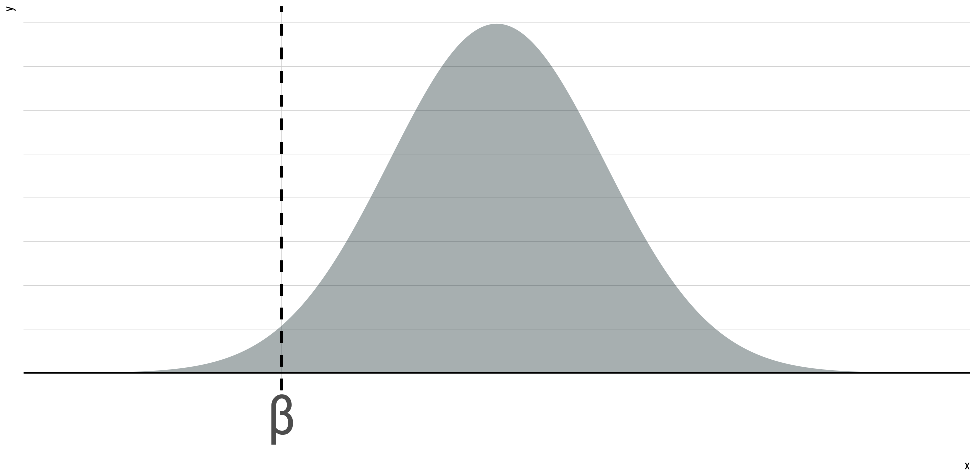
An estimator is **biased** if its expected value is different from the *true* population parameter.

When considering our slope coefficients ($\hat{\beta}_i$), we expect that they, on average, are close to the "true" population parameter, β_{pop}.

Unbiased: $\mathbb{E}\left[\hat{\beta}_{OLS}\right] = \beta_{pop}$



Biased: $\mathbb{E}\left[\hat{\beta}_{OLS}\right] \neq \beta_{pop}$



Omitting a relevant variable

- Assume we know the **true** population model:

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

- And we estimate the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i^*$$

with

$$u_i^* = u_i + \beta_2 x_{2i}$$

- Assuming that x_1 and x_2 (the omitted variable) share some degree of correlation (which is usually the case), the error term is no longer **independent** of an explanatory variable, as per CLRM Assumption III.

Omitting a relevant variable

- Consider a simple demand model:

$$log(qchicken_i) = \beta_0 + \beta_1 pchicken_i + \beta_2 pbeef_i + \beta_3 dispinc_i + \beta_4 log(xchicken_i) + u_i$$

- And we estimate it:

$$\begin{aligned} log(\widehat{qchicken_i}) = {} & 2.95 - 0.23\, pchicken_i + 0.18\, pbeef_i + \\ & + 0.000036\, dispinc_i + 0.75\, log(xchicken_i) \end{aligned}$$

Omitting a relevant variable

- And now we omit `dispinc` from the model:

$$log(\widehat{qchicken_i}) = 3.49 - 0.30\, pchicken_i + 0.25\, pbeef_i + 1.65\, log(xchicken_i)$$

- This model's residual term contains `dispinc`.

- Let us check out the correlation coefficient between `dispinc` and other variables:

corr_y_pchicken	corr_y_pbeef	corr_y_x
-0.8552982	-0.6940004	NA

Omitting a relevant variable

'True' model

term	estimate	std.error	statistic	p.value
(Intercept)	2.9575599	0.0951466	31.084255	0.0000000
p	-0.2342880	0.0176617	-13.265322	0.0000000
pb	0.1814819	0.0509694	3.560608	0.0008732
lexpts	0.7526487	0.1404342	5.359440	0.0000026
y	0.0000361	0.0000052	6.986129	0.0000000

Biased model

term	estimate	std.error	statistic	p.value
(Intercept)	3.4926329	0.0801754	43.562414	0.0000000
p	-0.3045472	0.0206204	-14.769222	0.0000000
pb	0.2551898	0.0708221	3.603253	0.0007563
lexpts	1.6504674	0.0804149	20.524400	0.0000000

Including irrelevant variables

Including irrelevant variables

- Now assume that the **true** model is:

$$y_i^{true} = \beta_0 + \beta_1 x_{1i} + u_i$$

- And, instead, we estimate

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

with

$$u_i^* = u_i - \beta_2 x_{2i}$$

Including irrelevant variables

- Suppose we add `popgro`, a variable measuring population growth, to our original model:

$$
\begin{aligned}
log(\widehat{qchicken_i}) = &\ \mathbf{2.89} - \mathbf{0.23}\, pchicken_i + \mathbf{0.19}\, pbeef_i + \\
& + \mathbf{0.000038}\, dispinc_i + \mathbf{0.69}\, log(xchicken_i) + \\
& + \mathbf{0.017}\, popgro_t
\end{aligned}
$$

Including irrelevant variables

'True' model

term	estimate	std.error	statistic	p.value
(Intercept)	2.9575599	0.0951466	31.084255	0.0000000
p	-0.2342880	0.0176617	-13.265322	0.0000000
pb	0.1814819	0.0509694	3.560608	0.0008732
lexpts	0.7526487	0.1404342	5.359440	0.0000026
y	0.0000361	0.0000052	6.986129	0.0000000

Model with irrelevant variable

term	estimate	std.error	statistic	p.value
(Intercept)	2.8951497	0.1353082	21.3967020	0.0000000
p	-0.2369439	0.0211080	-11.2253171	0.0000000
pb	0.1914541	0.0537460	3.5622008	0.0008984
lexpts	0.6996547	0.1722889	4.0609386	0.0001978
y	0.0000385	0.0000065	5.9044418	0.0000005
popgro	0.0177147	0.0300050	0.5903904	0.5579493

The `RESET` test

The RESET test

Knowing for sure whether our models suffer from Omitted Variables Bias (OVB) is hard.

However, the RESET test for functional form misspecification can help us.

It consists of running an **F-test** on **functional forms** of the **fitted values** of the dependent variable $(\hat{y})$.

These functional forms $(\hat{y}^2, \hat{y}^3, etc.)$ serve as ***proxies*** for potentially omitted variables.

Recall that functional forms of already included independent variables can also be omitted variables!

The RESET test

The **recipe** 👩‍🍳 👨‍🍳:

1. Estimate the regression model via OLS;

2. Store the regression's fitted values $(\hat{y}_i)$;

3. Use functional forms of $\hat{y}_i$ (squared, cubic terms, etc.) as **independent variables** in a new model;

4. Compare the fits of models from step **1** and **3** through an *F-test*;

5. In case these additional terms are **not** jointly significant, we do not suspect of omitted variables.

6. In case these terms are *jointly significant*, we should consider adding new regressors to the original model.

The RESET test

```
Estimate the regression model via OLS
```

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

Store the regression's fitted values $(\hat{y}_i)$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

Use functional forms of $\hat{y}_i$ (squared, cubic terms, etc.) as **independent variables** in a new model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 \hat{y}_i^2 + \beta_4 \hat{y}_i^3 + u_i$$

Compare the fits of models from step **1** and **3** through an *F-test*

- $H_0 : \hat{\beta}_3 = \hat{\beta}_4 = 0$
- $H_a : H_0$ is not true

The RESET test

- In case the **null hypothesis** is **rejected**, then we have evidence of omitted variables.
- In case we **do not reject** H_0, then we can stick with the original model.

In R...

```r
resettest(model_true, power = 2:4)
```

```
#> 
#>      RESET test
#> 
#> data:  model_true
#> RESET = 1.6352, df1 = 3, df2 = 43, p-value = 0.1953
```

What do we conclude?

The RESET test

In Stata…

```
estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of lq

H0: Model has no omitted variables

F(3, 43) =    1.64
Prob > F = 0.1953
```

What do we conclude?

Next time: OVB in practice