Violations of Classical Assumptions I: Omitted Variables Bias

Marcio Santetti | Spring 2023

Table of contents

Introduction 2

Omitted variables bias (OVB) 2

Consequences of OVB 3

Correcting for OVB 5

Including irrelevant regressors 6

Four important specification criteria 7

The RESET test for functional form misspecification 8

Introduction

Now that we know how to run, interpret, and test hypotheses from our regression models, it is time to go a little beyond its assumptions. What if one (or more) **Classical Assumptions** are violated? Two things about it. First, OLS will no longer be *BLUE*, that is, the *best linear unbiased estimator*, and our estimated coefficients will not be valid. Second, there is no need to worry. We can fix several mistakes with simple tests and procedures. We will learn the most common violations and solutions.

The first OLS problem we investigate regards **omitting** important independent variables from a regression model, known as **Omitted Variables Bias** (OVB). As the name already anticipates, such issue causes **bias** in our estimated coefficients, the $\hat{\beta}$'s. Thus, the *expected value* of our slope coefficients are no longer equal to the "true" population parameters.

However, since day 1 you are aware that we almost *never* have access to the "true" model. We can never be sure whether we are close to the population specification or not, and this is a serious challenge. As said before, **theory** is our best guide, and based on that we try to estimate the best model possible. Since in practice this problem is very common and hard to deal with, we first need to understand its theoretical details, and then evaluate what to do from there.

Omitted variables bias (OVB)

Before any regression estimation, our model must be **well specified**. This is part of CLRM **Assumption I**, and this means that the model must:

- Have the correct covariates (x_i);
- Have the correct functional form (whether or not to use logs, quadratic terms, interactions, etc.);

- Have the correct form of the stochastic error term (which must be additive).

If **any** of these requirements is not met, the regression model is **misspecified**. The last item is easy to address; the other two are more complicated.

Deciding whether a variable belongs in an equation should be based on **theory**. If it supports its inclusion, then the variable should be explicitly on the right-hand side of the regression. However, if there is a theoretical ambiguity, a *dilemma* arises. *Leaving* a relevant variable out of a model will likely **bias** OLS estimates, while *including* unnecessary regressors tends to **inflate** estimates' variances and standard errors, harming *inference* from our models.

Let us start with the first case. Suppose a *relevant* independent variable is left out of an econometric model, either because you have forgotten it, or perhaps there is no available data. This situation is known as an *omitted variable* case. The most serious problem associated with this fact is the **bias** such omission causes in the estimated coefficients, our $\hat{\beta}'s$. Analytically, this means

$$\mathbb{E}(\hat{\beta}_i) \neq \beta_i^{\text{true}}$$

That is, the expected value of the estimated coefficient *deviates* from the "true" value of the population parameter.

Consequences of OVB

What happens if a relevant variable is omitted from a model? For theoretical and presentation purposes, we will assume knowledge of the "true" model, at least for now. Say the population model is

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

Now suppose we *omit* x_2 in our sample regression model (regardless of the reason). Then, it becomes

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

Now, we have the error term denoted by u_i^*. Given the "true" and the estimated models, we can represent u_i^* as

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

The impact of x_2 then goes to u_i^*, so x_2 and u_i^* will be *correlated*. If x_2 and x_1 have some type of correlation—which is usually the case—, x_1 and u_i^* will change as x_2 changes. Therefore, the error term is no longer *independent* of the explanatory variable, as stated by CRLM **Assumption III**. As a consequence, the Gauss-Markov theorem is violated and OLS is no longer *BLUE*, since

$$\mathbb{E}(\hat{\beta_1}) \neq \beta_1^{\text{true}}$$

The estimated coefficient of β_1 will **compensate** for the fact that x_2 is missing from the equation. If x_1 and x_2 are correlated, the estimated model will attribute to x_1 variations in y actually caused by x_2, denoting a **bias** in x_1's coefficient.

Let us look at an example:

$$\hat{Y}_i = 27.7 - 0.11 PC_i + .03 PB_i + .23\ YD_i$$
$$n = 29 \qquad \bar{R}^2 = .99$$

This is a standard demand model for chicken (Y_i), controlling for its price (PC_i), the price of a substitute—in this case beef (PB_i)—, and per capita disposable income (YD_i). Notice that the model's adjusted R^2 is excellent (0.99), and it follows what standard microeconomic theory recommends for normal goods, such as chicken.

Now, suppose we leave PB_i out of the model:

$$\hat{Y}_i = 30.68 - 0.08PC_i + .25\ YD_i$$
$$n = 29 \qquad \bar{R}^2 = .98$$

Let us compare the $\hat{\beta}$ coefficients from both models:

- $\hat{\beta}_{PC} \implies$ from -0.11 to -0.08
- $\hat{\beta}_{YD} \implies$ from 0.23 to 0.25

Notice that the coefficient for PC_i becomes *biased upward* (i.e., it gets *less negative*), and the same happens with YD's. Furthermore, despite still having a good adjusted R^2, it is lower than before. Thus, we were better off with the first model.

This is a clear case of OVB. Even though we cannot be 100% sure that the first model is the "true" one, it was estimated based on what microeconomic theory recommends,[1] and once one of these recommended variables was removed, the coefficients became biased and the goodness-of-fit of the model was also affected.

[1] Could the model be improved by adding a variable representing *preferences*? Food for thought.

Correcting for OVB

As said before, it is hard to precisely detect OVB, since it is impossible to assess the "true" population model. Furthermore, the best indications come from **theory**, guiding us with respect to the following questions:

- What variables **must** be included?
- What **signs** do we expect?
- What is the **range** of acceptable values for the $\hat{\beta}$'s?

As a *practical* recommendation, we should always invest time *thinking* about our model, before any data collection and estimation begin. This way, we know what the literature says and recommends, and we are less likely to be surprised with any results. In addition, if you know that your model suffers from OVB, it is always preferable to have a *parsimonious* model (i.e., preferring simpler specifications, as opposed to complex models), and let theory guide your next steps before crowding the model with any new variable(s).

Including irrelevant regressors

Adding variables in models where these do not belong **does not** necessarily cause bias, but tends to *inflate* the variances and standard errors of estimated coefficients.

Let us assume once again we know the "true" model governing y:

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

Now suppose we get excited with Econometrics and *include* x_2 in our sample regression model. Then, it becomes

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

And we have a residual term denoted by u_i^{**}. Now, u_i^{**} is

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

Now, our regression's error term is *underrepresented*, since it now has a term that should not be in the model at all. This inclusion will not cause bias if the "true" coefficient of the additional (irrelevant) variable is *zero* ($\beta_2^{true} = 0$). Then, $\hat{\beta}_1$ will be *unbiased*. But, since we don't know the "true" value of β_2, why take the risk?

What is almost impossible to avoid is that the inclusion of this irrelevant variable will likely *increase* the variance of the estimated coefficients, whose main consequence is a *decrease* in the absolute magnitude of *t-scores*, thus affecting inference. Lastly, the adjusted R^2 may also fall.

Let us look at an example:

$$\hat{Y}_i = 27.6 - 0.58PC_i + .012PB_i + .24\ YD_i - .14R_t$$
$$n = 29 \qquad \bar{R}^2 = .98$$

The model above adds the interest rate (R_t) to our chicken demand model. Why would someone add the interest rate to a demand model for chicken? Unless a consumer is considering taking a loan to buy dinner, the inclusion of R_t is very questionable. In addition to a drop in $\bar{R}^2$, the coefficients have slightly changed, relative to the ideal model. Thus, we have a proof that inflating our models with irrelevant variables only brings more problems than necessary.

Four important specification criteria

Based on what we have seen so far in this lecture, as well as previous contents, we can consider four model specification criteria that are necessary to think about when doing Econometrics in practice, as well as comparing different models:

1. *Theory*: is it theoretically recommended to add a variable to the model?
2. *Statistical significance*: are our coefficients ($\hat{\beta}_k$) statistically significant? Are their signs as expected? Did adding a new independent variable change the statistical significant of other control variables?
3. *Goodness-of-fit*: when a new variable is added (removed), is the overall fit (measured by the adjusted R^2) improved?
4. *Bias*: how do the coefficients behave when adding new variables to a model?

The RESET test for functional form misspecification

Suppose the true model representing a dependent variable is known as

$$y_i^{\text{true}} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{2i})^2 + \beta_4 x_{1i} x_{2i} + u_i$$

And, instead, we estimate

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

You already know that the error term from the latter model, u_i^*, will include $\beta_3 (x_{2i})^2$ and $\beta_4 x_{1i} x_{2i}$, which are part of the "true" specification. But can **functional forms** of independent variables also be considered *potential omitted variables*?

The answer is **yes**. I know, this adds another layer of complexity regarding OVB, but, if that is the case, at least we do not need to look for new data; we just work with what we already have. With this in mind, it is possible to *test* for possible omitted

variables. And one of the most popular tests for OVB and model misspecification is *Ramsey's Regression Specification Error Test*, also known as the **RESET** test.

In previous sections, we have observed how the adjusted R^2 changes as we include/remove covariates from a regression model. This is a good start, but if we want to be more robust, the RESET test also checks for model misspecification. Put simply, it measures whether the *fit* of a given model can be significantly improved by the addition of squared, cubed or even higher powers of the estimated dependent variable, $\hat{y}$. As will be more evident in a moment, the basic *intuition* of this test is to include additional terms as *proxies* for any possibly omitted variable(s) or incorrect functional forms being present in a regression model.

We will look at a more analytical explanation of the test now, and in our applied lecture, we will apply this test to real data. The RESET test has basically *three* steps. The **first** is to estimate our regression model using OLS, which is what we have been doing so far. After the model is estimated, we end up with (for our example purposes, we have a regression model with only 2 regressors):

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

The **second** step is to take the fitted values for y, i.e., $\hat{y}_i$, from the above estimated model and create *new* variables, namely $\hat{y}_i^2$, $\hat{y}_i^3$, and even $\hat{y}_i^4$. Then, we use these terms as *independent variables* and re-estimate the model once again using OLS:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 \hat{y}_i^2 + \beta_4 \hat{y}_i^3 + \beta_5 \hat{y}_i^4 + u_i \tag{2}$$

Equation (2) can be considered an **auxiliary regression**. Usually, using the second and third powers is enough for RESET test purposes. The **third** and last step is to compare the fits of models (1) and (2) using an *F-test*. If the two models are *significantly*

different from each other, then model (1) may be misspecified. The RESET test, however, does not specifically inform in what variable lies the problem, or what functional may have been omitted from the model. Despite this fact, it is a great way to diagnose possibly omitted variables and/or functional forms.

Okay, these are the steps of the RESET test. But what is the logic contained in the **second** and **third** steps? A few lines ago, I said that $\hat{y}_i^2$, $\hat{y}_i^3$, and other powers of the estimated values of the dependent variable may serve as *proxies* for omitted variables. That is true. Now, look once again at equation (1). If you *square* the left-hand side, you'll get $\hat{y}_i^2$. To maintain equality between both sides, you are also *squaring* the right-hand side, obtaining $(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})^2$. I will not ask you to solve this latter term, since it looks ugly, but I am sure you have at least some idea of what it will look like. It will generate several *squared* and *interaction* terms. Likewise, the same logic applies if we raise $\hat{y}_i$ to higher powers.

The newly generated terms will act as proxies for **potentially omitted variables**, especially functional forms (such as squared and interaction terms) that may not have been included in the original model. But why do we use $\hat{y}_i^2$, $\hat{y}_i^3$ and perhaps other powers, and not the right-hand side of equation (1) raised to these powers? Ramsey was well aware of **degrees-of-freedom**, and since the LHS must equal the RHS, we are able to save several DOFs by using functional forms of the estimated dependent variable, instead of flooding the auxiliary regression with an enormous amount of additional regressors.

Then, what we do in the third step is running an F-test on the new coefficients included in the auxiliary regression. For our example, it would be

- $H_0 : \hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 0$
- H_a : H_0 is not true

Thus, the RESET test is nothing but an F-test applied to an

auxiliary regression that includes higher powers of $\hat{y}$ to identify possibly omitted variables, especially related to functional forms. The above null and alternative hypotheses can be translated for the RESET test as

- H_0 : *the model is well specified*
- H_a : *the model is not well-specified*

In case we *do not reject* H_0, our model **does not** suffer from functional form misspecification, since the "new" coefficients are not jointly significant. In case we *reject* H_0, then our model suffers from a type of omitted variables problem. A good starting point to fix this is testing different functional forms of the independent variables, such as x_1^2, $x_1 \cdot x_2$, an so on. Then, we can run this test again and see whether the problem was fixed.

This test's procedure will be made clearer as we practice this with real data. However, it is important that you capture the **intuition** behind this test, so performing it with real-world data will be really simple.