Violations of Classical Assumptions II: Multicollinearity

Marcio Santetti | Spring 2023

Table of contents

Introduction

In this second lecture covering possible violations of OLS assumptions, we will study **multicollinearity**. CLRM **Assumption VI** states that *no independent variable is a perfect linear function of one or more other covariates*. Even though a *perfect* relationship between x_i variables is almost impossible, cases of *imperfect* (i.e., stochastic) linear associations are not uncommon when setting up econometric models. Even though Assumption VI *does not* cover the latter case, we have to be prepared to deal with it, since it may bring substantial problems to our estimated models.

The real meaning of multicollinearity is that, the more highly correlated two (or more) independent variables are, the more difficult it becomes for OLS to *accurately* estimate the coefficients close to the "true" regression model. If, for instance, x_1 and x_2 move *identically*, and both are present in a regression model, how can OLS clearly *disentangle* the impact of each regressor on the dependent variable? If the correlation coefficient between x_1 and x_2 is low, we can still be fairly accurate; however, as it increases, it is almost impossible to distinguish between these two variables with respect to their effects on the variable of interest.

We start studying the type of multicollinearity considered in Assumption VI, known as **perfect multicollinearity**. Then, we move on to its **imperfect** version, which is more commonly seen in practice. Later, we will look at its major *consequences* for OLS estimation, ways to *detect*, and *treat* this problem within the range of Ordinary Least Squares.

Perfect multicollinearity

Cases of *perfect* multicollinearity directly violate CLRM Assumption VI, since there is a **perfect linear relationship** between two

or more independent variables. As an example, consider the definition of an independent variable x_1:

$$x_{1i} = \alpha_0 + \alpha_1 x_{2i}$$

Notice that there is no **stochastic** term in x_1's definition. This variable, then, shares a **deterministic** relationship with x_2, with no *uncertainty* involved. In other words, movements in x_1 can be *completely* explained by movements in x_2. The next figure illustrates an example of such relationship, with $\alpha_0 = 3$ and $\alpha_1 = 1$.

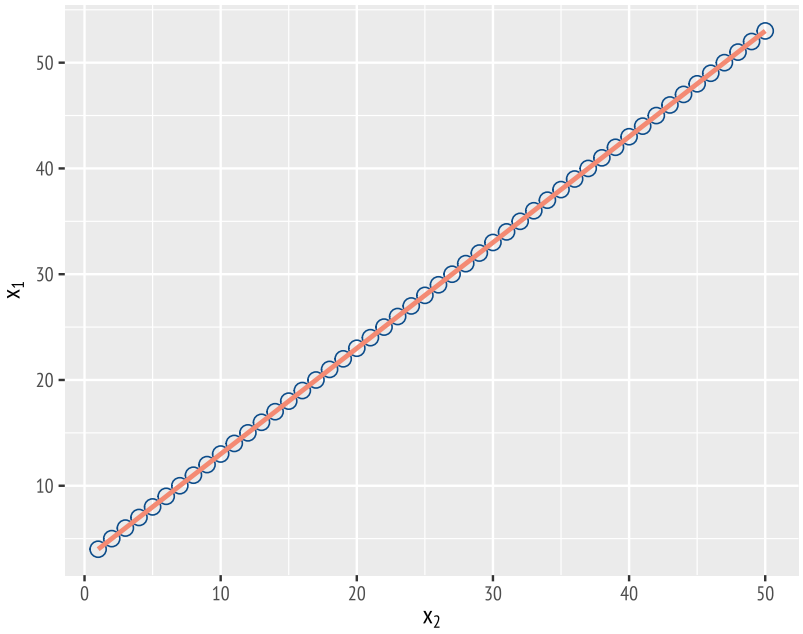

Figure 1: A perfect linear relationship.

If both variables are included in a regression model, such as the one below,

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

it will suffer from perfect multicollinearity.

With perfect multicollinearity, OLS estimates tend to be *indeterminate*, since it cannot distinguish effects coming from x_1 or x_2 with enough precision, given that these basically describe the same process. In other words, the *ceteris paribus* assumption no longer holds, since it is not possible to assume x_1 constant to describe the partial effect of x_2 on y, for example.

Fortunately, perfect multicollinearity is *rare* to occur in practice, since theory tends to prevent it and such redundancies are relatively easy to detect prior to any estimation.[1] Given this fact, either x_1 or x_2 should be dropped from the regression model—or one could generate a third variable, derived from a combination of these two—, thus avoiding this violation.

[1] Also, statistical packages like R and Stata will not let you estimate a model with perfect multicollinearity.

Imperfect multicollinearity

The *imperfect* version of multicollinearity is defined as a linear functional relationship between two or more independent variables, with the difference of not being a *perfect* linear association. (That is, the correlation coefficient is less than 100%.) However, depending on the strength of the relationship, it can *significantly* affect the estimation of β coefficients if the related variables are all included in the same model.

Consider again two linearly related variables x_1 and x_2:

$$x_{1i} = \alpha_0 + \alpha_1 x_{2i} + \epsilon_i$$

Notice that now x_1 is not *fully* explained by x_2, since we have included a *stochastic* term ϵ_i that addresses some *uncertainty* to this relation. In other words, x_1 is determined by other factors (included in ϵ_i), and not only by x_2. The next figure illustrates two example of such relationships. In the left panel, x_1 and x_2 have a correlation coefficient of 0.87, while in the right panel, the correlation is 0.36. The *more* scattered the data points are around the straight line, the *less* correlated the variables are.

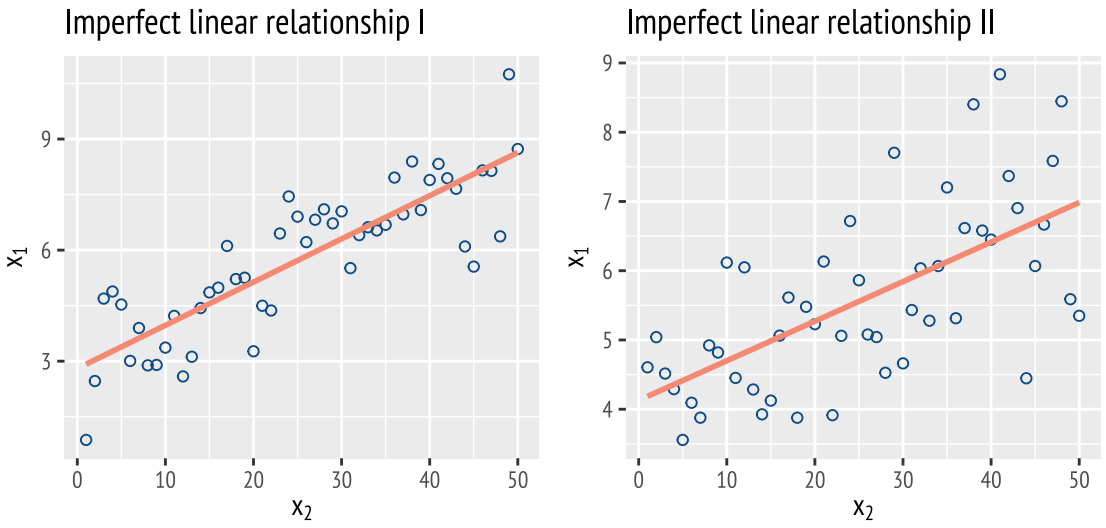

Figure 2: Two imperfect relationships.

Even though imperfect multicollinearity **does not** violate CLRM Assumption VI, it can also bring several problems to the quality of a regression model. We look at these consequences in the next section.

Consequences of multicollinearity

Given that our model suffers from multicollinearity (either perfect or imperfect), what happens to our β estimates?

Firstly, multicollinearity, by itself, **does not cause bias**. It is possible that a model suffering from multicollinearity also has some omitted variable, thus causing bias. But the latter problem is not caused by multicollinearity.

Secondly, despite not causing bias, multicollinearity affects the **precision** of β estimates. Although still unbiased, the $\hat{\beta}$'s will come from distributions with much *larger variances*. As a consequence, the standard errors (SEs) tend to increase, given the uncertainty regarding the respective effects of the collinear variables.

The next figure illustrates two unbiased $\hat{\beta}$ coefficients (both are centered around the "true" value, set to 1 for this example), but with different variances, and thus different standard errors. Assume that the coefficient β_j (in blue) is estimated in a model containing multicollinearity, while β_i (in red) is the same slope coefficient, estimated after removing a collinear variable. The latter becomes more *reliable* than the former, since its distribution is more tightly concentrated around the "true" value of 1.

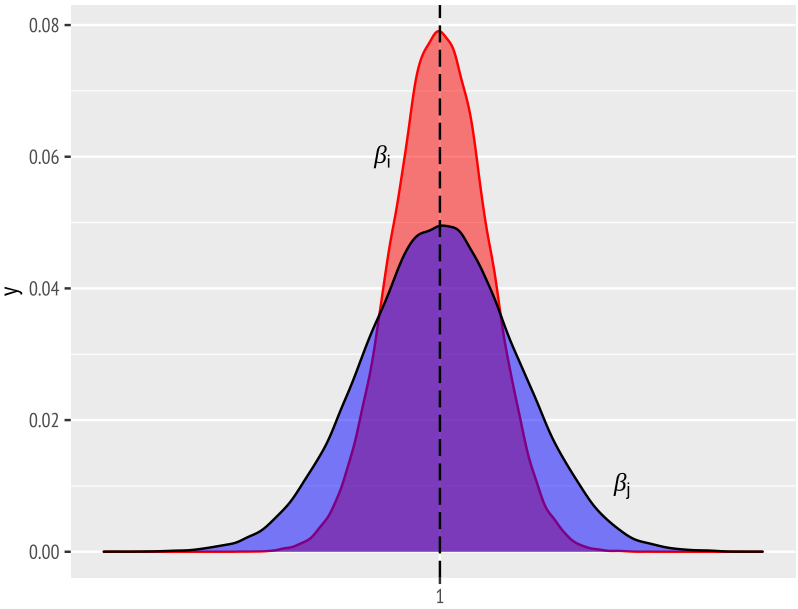

Figure 3: Two density curves.

Thirdly, *t-scores* are likely to *fall*. Recall its formula:

$$t_k = \frac{\hat{\beta}_k - \beta_{H_0}}{SE(\hat{\beta}_k)}$$

If multicollinearity increases the standard errors of β coefficients, the above formula's denominator will *increase*, while the numerator remains constant, since there is no bias. As a consequence, *t-statistics* will fall, and this may harm **inference** from our model, both regarding statistical significance and any

other individual test we may perform to our coefficients.

Lastly, the adjusted R^2 of a model with multicollinearity is *not* heavily affected, relative to a model without it. Therefore, looking at this goodness-of-fit measure does not help when trying to detect this problem.

Consider the following example (standard errors in parentheses):

$$\hat{C}_i = -367.83 + \underset{(1.0307)}{.5113}\, YD_i - \underset{(.0942)}{.0427} LA_i$$

$$n = 45 \qquad \bar{R}^2 = .835$$

where

- C_i: consumption expenditures of the i^{th} student;
- YD_i: annual disposable income of the i^{th} student;
- LA_i: liquid assets (savings) of the i^{th} student.

However, savings are a *function* of disposable income. It is likely, though, that there are more factors affecting savings than the level of disposable income alone. Therefore, this is a clear case of **imperfect** multicollinearity.

What happens if we *drop* LA_i from this model?

$$\hat{C}_i = -471.43 + \underset{(.157)}{.9714}\, YD_i$$

$$n = 45 \qquad \bar{R}^2 = .861$$

$\bar{R}^2$ has slightly improved, and the standard error of $\hat{\beta}_{YD}$ has *decreased*, making it more *precise*. Notice that $\hat{\beta}_{YD}$ also has changed a lot. This does not mean that the model does not suffer from OVB, though. If we consider that consumption is not only determined by disposable income, we may find other relevant variables to include in the model. However, the model is free from multicollinearity for now.

Detecting multicollinearity

How do we realize our model suffers from multicollinearity? When working with real-world data, it is almost *impossible* to set up a model where all explanatory variables are totally uncorrelated with each other. The *severity* of this correlation may change from sample to sample, even if the variables are the same, but what matters is that the damage caused by multicollinearity is a *matter of degree*, and there is *no* widely accepted statistical test that can *prove* that a model suffers from this problem.

We can use some *tools*, though. The *first* and most *basic* thing to do is computing **pairwise correlation coefficients** for the set of independent variables. In case the correlation between two independent variables is *high* (usually, above 80% can be considered high), it may be better to drop one covariate from the model. However, the correlation coefficient is a *bivariate* measure. In larger models, it may be easy to lose sight of the pairwise measures.

Nevertheless, looking at correlations must be the starting point for detecting multicollinearity, as well as looking at *scatter diagrams* between the independent variables. After that is done, we can move on to possible remedies for it.

Remedies for multicollinearity

If a high correlation coefficient between independent vaiables is detected, *dropping* one of them may the the best thing to do. Another solution may be to *transform* collinear variables in a *single* variable, if the case allows for it.

In case theory recommends including the variables that are highly correlated in your model, this may be a *sample phenomenon*. In this case, *increasing the sample size* may be a solution, even though its feasibility is not always straighforward.

Variance Inflation Factors (VIFs)

An interesting tool that helps us to detect multicollinearity are Variance Inflation Factors (VIFs). It looks at the extent to which a given explanatory variable can be explained by all the other explanatory variables in the regression equation.

The *VIF* is an index of how much multicollinearity has increased the variance of an estimated coefficient. A high *VIF* indicates that multicollinearity has increased the variance of the estimated coefficient by quite a bit, yielding a decreased t-score. In a nutshell, the *VIF* is simply the factor by which the variance of a coefficient β_i is inflated by the presence of correlation among the independent variables of a regression model.

Suppose you want to use the *VIF* as an attempt to detect multicollinearity in an original equation with *k* independent variables:

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

Doing so requires calculating *k* different *VIF*s, one for each x_i. Calculating the *VIF* for a given x_i involves two steps:

1. Run an OLS *auxiliary regression* that has x_i as a function of all the other explanatory variables in the equation. For $i = 1$, this equation would be:

$$x_1 = \alpha_0 + \alpha_1 x_{2i} + \alpha_2 x_{3i} + ... + \alpha_{k-1} x_{k-1i} + v_i$$

where v is a classical error term.

2. Calculate the Variance Inflation Factor for $\hat{\beta}_i$:

$$\text{VIF}(\hat{\beta}_i) = \frac{1}{(1 - R_i^2)}$$

where R_i^2 is the coefficient of determination of the *auxiliary regression* in the first step. There will be a different R_i^2 and $\text{VIF}(\hat{\beta}_i)$ for each x_i. The higher the *VIF*, the more severe the effects of multicollinearity.

A common rule of thumb indicates that a *VIF* higher than 5 already indicates a high multicollinearity. If its value is 1, there is nothing to worry about. If it goes beyond 10, then the model definitely suffers from multicollinearity. It is also important to remind that VIFs are not meant to be the ultimate measure to detect multicollinearity, but indeed a **simple tool** to indicate its presence and help us in our modeling decisions.