The Classical Linear Regression Model (CLRM)

Marcio Santetti | Spring 2023

Table of contents

Introduction

Last week, we studied the basic regression model. We started from its simple version, with only one regressor, and later we made it more complex by adding a larger number of control variables, so that we mitigate the influence of the error term in our estimations, as well as allowing for empirical analyses that are more consistent with theoretical priors.

Our baseline method for estimating the coefficients of interest is **Ordinary Least Squares** (OLS). It is a powerful—and incredibly simple—technique that allows for diverse applications. However, as you are probably familiar with, all economic models are built upon a set of theoretical assumptions that are necessary for its functioning. Some of these assumptions are hard to wrap our heads around, and that's natural. But this fact does not imply that we should throw away the techniques and only use those 100% consistent with our views. I used to think like that as an early graduate studuent, but reality is not that simple; instead, it is preferable to be aware of the pros and cons of every model, so that we can use it for our purposes, but at the same time being able to criticize it and discuss its weaknesses whenever necessary. As we have been trying to do in the lectures, the most important thing to pay attention is the **intuition** behind models, procedures, and theories. The assumptions and math behind these become secondary as soon as we concentrate on the intuitive properties of our models.

Econometric theory is no different than Macro or Micro models. It is built upon several assumptions, and we will investigate those pertaining mostly to OLS this week. Some of these assumptions are straightforward, and others are easily broken. The good part is that we will learn how to deal with these failures later on in the course, more specifically after the Midterm exam. And we can fix some of these problems still using OLS as our main method.

But before we mess around with some classical assumptions, we must know each one of them. Books usually differ with respect to the number of assumptions, given that some authors prefer to split one assumption into two or three separate ones. Here, we will compress the classical assumptions into 7.

The classical assumptions

The term *classical* refers to a set of assumptions required for OLS to hold, in order to be the *"best"*[1] estimator available for regression models. One of the most important tasks in Econometrics is to decide whether these assumptions hold for a model or not. Let us investigate these further.

- ***Assumption I***: *The regression model is linear, correctly specified, and has an additive stochastic error term.*

Consider the following model with k independent variables:

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

This OLS model is **linear**, since all parameters (that is, the β's) are all linear, not assuming any functional form other than its original level form.

A model like this, for example,

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

is still linear, since the $\log(\cdot)$ functional form is only applied to some of the model's **variables**. A model that violates the linearity assumption looks something like this:

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

[1] You will see why the term *best* lies around quotation marks in a moment, hold on.

In this particular case, β_1 is log-transformed, and β_2 is squared. That is, these two parameters are no longer **linear**, and OLS can no longer estimate such model. Of course, this model can still be estimated—and may be useful in some contexts—, but it will no longer be so through OLS, which is our standard technique for Econometrics.

Assumption I still has two other parts: one concerns the model being correctly specified, and the other concerns its residual. Starting with the latter, it states that the **error term** must be included in the model through *addition*. This simply means that the residual cannot be multiplied/divided by other variable(s), and must appear only by itself in the model.

The basic definition of a well-specified model is that it has **no omitted variables** and **no incorrect functional form**. These two are hard to determine precisely, since we do not have access to the "true" underlying population model for our problem at hand. What we can do, however, is estimate our sample model according to *theory*, since it tries to map the "true" model as precisely as possible.

- ***Assumption II***: *The stochastic error term has a zero population mean*.

This was already introduced last week, but we need to formalize the expected value of the stochastic error term as a classical assumption. The *specific value* of the error term u_i for each observation is determined purely *by chance*, since it is a random variable that follows a given probability distribution. What OLS assumes is that each observation of u is being drawn from a distribution whose mean (expected value) is **zero**.

The graph below illustrates the distribution of a given residual term. Notice that it is centered around a mean of zero. Thus, from this assumtpion the $\mathbb{E}(u) = 0$ from last week is derived.

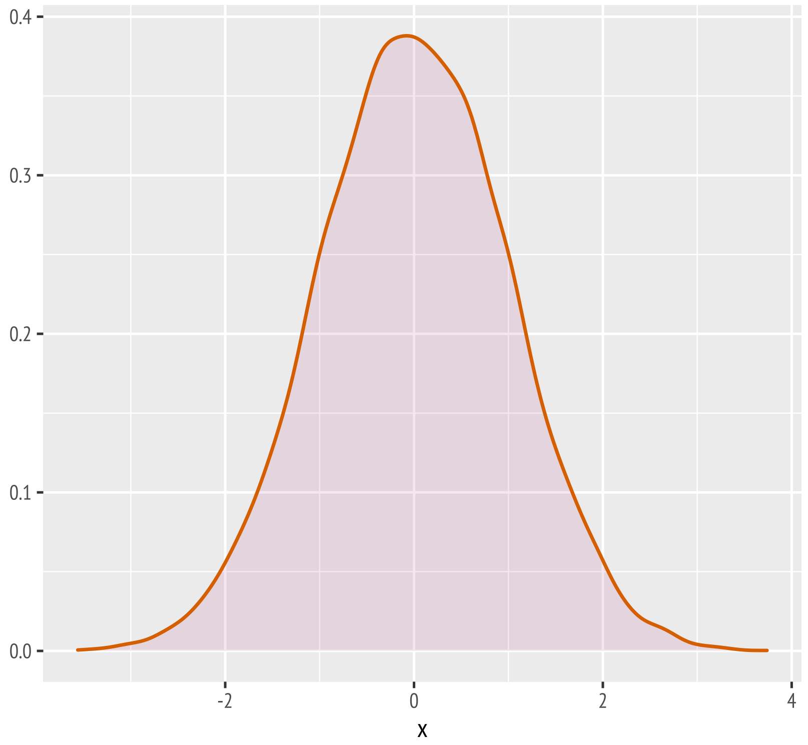

Figure 1: The residual's distribution.

- ***Assumption III***: All explanatory variables (x_i) are uncorrelated with the error term u_i.

This assumption states that observed values of control variables are determined *independently* of the values contained in the error term. In case there is correlation between one independent variable and the error term—that is, $\text{cor}(x_i, u_i) \neq 0$—, OLS estimates would likely attribute to x_i some of the variation in y that actually pertains to u_i.

What this means, in practice, is that the slope coefficient referring to x_i will likely be *over* or *underestimating* the effect of x_i on y_i, when part of this change is actually due to the error term. Such problem will likely happen whenever a variable that is correlated with x_i has been **omitted** from the model, therefore lying in u_i. This can also refer to *functional forms* of x_i, as we will see later.

Therefore, this assumption once again reinforces the importance of having a *well-specified model*, so that problems like this one are avoided.

- ***Assumption IV***: *Observations of the error term are uncorrelated with each other*.

Given that we assume that observations of the error term are drawn *independently* from each other, if a correlation exists between one observation of u_i and another, it will be difficult for OLS to get *accurate* estimates of the standard errors (SEs) of coefficients, harming inference capabilities.[2]

In a nutshell, this assumption only reinforces what was presented in **Assumption II**, adding that all draws from the error's distribution are **independent**.

- ***Assumption V***: *The error term has a constant variance*.

[2] We will study in more detail what standard errors mean. For now, you may consider these the same way as what *standard deviation* means to sample statistics. Standard errors refer to the *precision* of our β estimates.

This assumption is also known as the **homoskedasticity** assumption. In analytical terms, this assumption implies

$$\mathrm{Var}(u_i) = \sigma^2$$

The above means that the residual from our regression has an *equal spread across observations*, since σ^2 is just a constant value. In case we violate this assumption, we have a **heteroskedastic error term**. Analytically, an example of the latter may be

$$\mathrm{Var}(u_i) = \sigma^2 \cdot x_i$$

If x_i is not a constant (which is hard to imagine), then u_i's variance is a function of an independent variable, thus not being constant anymore. Such violation is problematic for OLS, and we will deal with that in a few weeks. Especially in cross-sectional data, where in many cases the variables tend to vary a lot across individuals, the homoskedasticity assumption will likely be violated. Figure 2 illustrates an example of residuals with non-constant variance. As you already know, the distance between each data point and the OLS regression line denotes the residual for each observation i. Notice that this distance increases as the values of x and y increase; thus, the *pattern* of spread of these residuals is changing across observations, meaning that the variance is not constant. Of course, we could represent this situation in a number of different ways, but this should be enough for now.

- ***Assumption VI***: *No explanatory variable is a perfect linear function of any other explanatory variable*.

This assumption, also known as the *no perfect multicollinearity assumption*, is better exposed through a simple example. Consider, for example, the definition of an independent variable we denote by x_1:

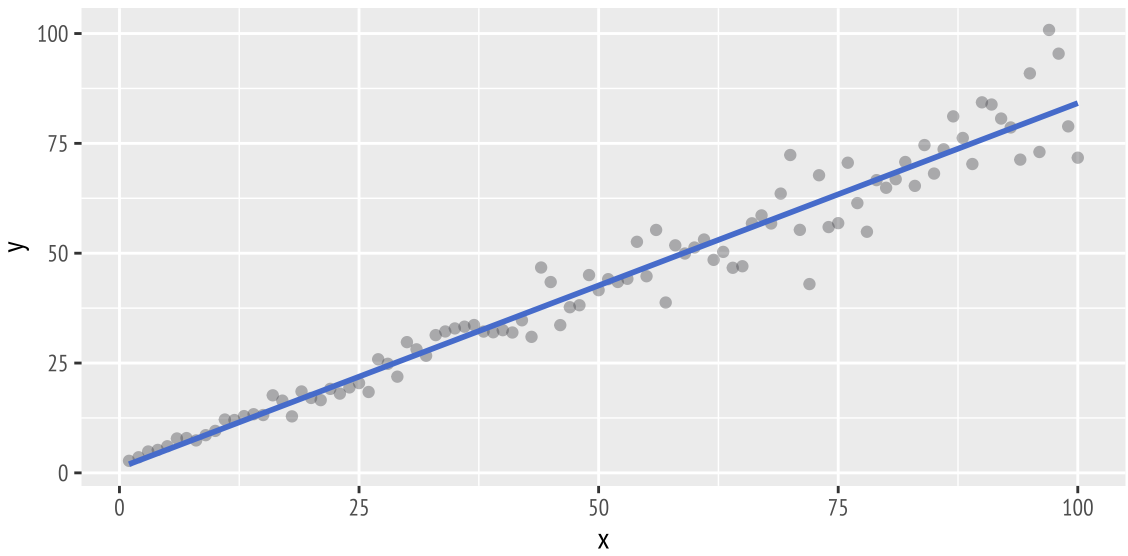

Figure 2: Heteroskedastic residuals.

$$x_1 = 50 + x_2$$

or, even,

$$x_1 = 10 \cdot x_2$$

In both examples, x_1 is a *linear, deterministic function* of another independent variable, x_2. For modeling purposes, x_1 and x_2 basically represent the *same variable*. In other words, relative movements of one will be precisely matched by the other control variable, even though their absolute magnitudes might differ. If this is the case, how can OLS distinguish between x_1 and x_2 to explain changes in the dependent variable?

It can't. Thus, we can no longer rely on OLS to have a precise model. Of course, it is wrong to assume that our explanatory variables would not have some sort of relationship. What this assumption rules out are **deterministic** associations between them. That is why we call it the *perfect multicollinearity* assumption. We will deal with such problem in detail later on.

- ***Assumption VII***: *The error term is normally distributed*.

This last assumption is **not** strongly required for OLS estimation, but of major importance for one of our next topics, *hypothesis testing*. Significance tests, such as *t-* and *F-tests* are heavily based on the normal distribution, and assuming so for the error term facilitates inference for our models.

Thus, joining **Assumptions II** and **V**, we summarize **VII** with the following statement:

$$u_i \sim \mathcal{N}(0, \sigma^2)$$

This is a simple way of compressing information about a random variable's distribution. In case it is normally distributed, we use the *mean* and the *variance* to define it. Thus, the statement above should be read as: *"the error term* u_i *is normally* ($\mathcal{N}$) *distributed* (~)*, with a mean of 0, and variance* σ^2*."*

These are the theoretical assumptions that define the **Classical Linear Regression Model** (CLRM). Some of them are easy to obey, but some others require serious work around them. We will analyze each one of them in detail after the Midterm exam.

The sampling distribution of $\hat{\beta}$

Similar to the error term, the estimates of our regression coefficients, the $\hat{\beta}$'s, also follow *probability distributions*. Such distributions, when evaluated across different samples of the same size, are known as **sampling distributions**.

Given that **Assumption VII** defines the error term following a Normal distribution, it logically follows that our estimated coefficients will also be **normally** distributed. To illustrate the latter point and what a sampling distribution means in practice, suppose we want to estimate the following model:

$$salary_i = \beta_0 + \beta_1 GPA_i + u_i$$

where $salary_i$ is the starting salary of the i^{th} graduate from last year, and GPA_i is their GPA from high school. We are interested in $\hat{\beta}_1$, that is, the coefficient capturing how an additional GPA point changes a graduate's salary.

Suppose we collect information from a sample $n = 35$, and use OLS to run our regression. After running the model, OLS gives $\hat{\beta}_1 = 8.6$. But what if we select a *second sample* of 35 individuals? Will $\hat{\beta}_1$ be the same as the first model's?

If your answer was **no**, then you haven't slept throughout your Stats classes (or at least not too much). Given that our samples are *random*, we surely do not expect equal estimates for β_1. These will *depend on the sample* we collect. Since we have different samples, we have different students, with different GPAs. Therefore, we will have different estimates for this second sample, and a third, a fourth, and so on.

Now, suppose that for our second sample, our regression model gives us $\hat{\beta}_1 = 8.1$. For a third, $\hat{\beta}_1 = 11.3$; for a fourth, $\hat{\beta}_1 = 6.9$; and for a fifth sample, $\hat{\beta}_1 = 8.5$. The **average** for $\hat{\beta}_1$ across these 5 different samples of size 35 is 8.68. This means that we can construct a **sampling distribution** out of these 5 coefficients, obtained from each different sample. And this distribution would be centered around its mean, 8.68.

The distribution of the $\hat{\beta}$ coefficients across **all possible samples** has its own mean and variance. For an adequate estimation, we would want the mean of the sampling distribution of these $\hat{\beta}$'s to be equal to its *true* population value, β_{pop}. Suppose this *true* value for our $\hat{\beta}_1$ from before is 8.4. Our mean from the 5 samples of size 35, equal to 8.68, does not match the true value, but it is likely that if **enough samples of the same size** are taken, the average $\hat{\beta}_1$ would eventually approach the true average value, 8.4.

Properties of the mean

When using OLS to estimate our regression models, we desire that the distribution of $\hat{\beta}$'s centers around a mean that equals (or at least gets as close as possible) to the *true* mean of the coefficient being estimated. Such property is called **unbiasedness**, and it should not be a new word to you.

An estimator $\hat{\beta}$ is an **unbiased** estimator if its sampling distribution has as its expected value the *true* value of β. That is,

$$\mathbb{E}(\hat{\beta}) = \beta_{\text{true}}$$

Here, *true* and *population* values are the same. If an estimator produces $\hat{\beta}$'s that are *not* centered around the *true* β, we have a **biased** estimator. In practical terms, the coefficients we are estimating are not representative of the *population* parameter we would like to obtain from our study. We will see reasons for that in a few lectures.

Properties of the variance

In addition to the mean of our estimated β coefficients, we also want their distributions to be as **narrow** (i.e., precise) as possible. We compare some distributions in Figure 3:

In the figure above, we have the distributions of two $\hat{\beta}$ coefficients: one in *orange* and one in *green.* In case we assume the *true* parameter value to be 0, we see that both sampling distributions are centered pretty close to 0, thus being **unbiased**. However, which one has the **highest variance**? The distribution with the

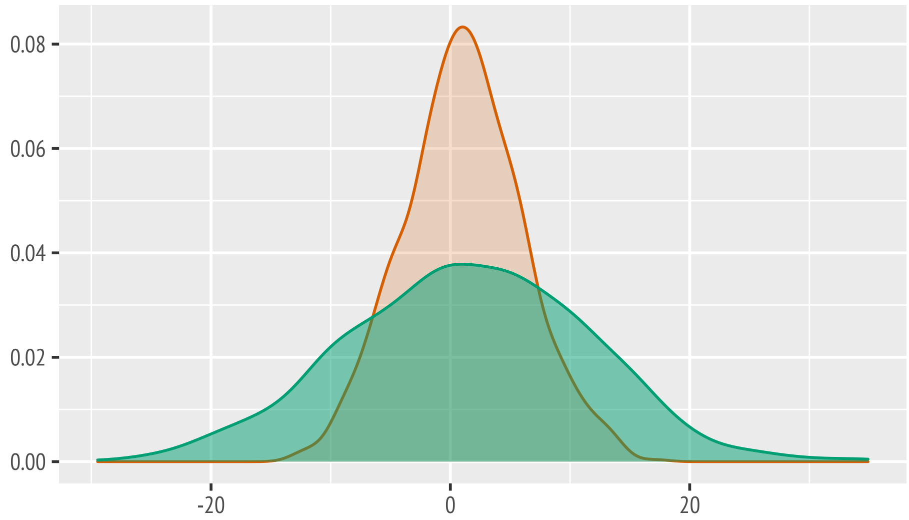

Figure 3: Distribution of two slopes.

widest density is for the $\hat{\beta}$ in *green*, meaning that it is more likely to be far from the true average value of 0, since its range comprehends a much greater area then that of $\hat{\beta}$ in *orange.* In other words, $\hat{\beta}$ in *orange* is **more precise** than $\hat{\beta}$ in *green.*

The statistical jargon for this precision feature is called **relative efficiency**. It serves to compare two unbiased estimators: the one with the smallest variance is said to be *relatively more efficient* than another with a wider sampling distribution. Now, let's mix once again the unbiasedness and efficiency properties by analyzing Figure 4.

Here, once again we assume that the *true* value of the β parameter is 0. Now, we present three sampling distributions: $\hat{\beta}$, in *orange*, $\hat{\beta}$, in *green*, and $\hat{\beta}$, in *blue.* Try to *compare* these 3 distributions in terms of unbiasedness and relative efficiency.

Given that we know the true value of β, both $\hat{\beta}$ in *orange* and $\hat{\beta}$ in *green* are *unbiased*, since their distributions are centered around this true value. And $\hat{\beta}$ in *orange* is *relatively more efficient* than $\hat{\beta}$ in *green*, since it has a lower variance (i.e., a thinner density curve). With respect to $\hat{\beta}$ in *blue*, it is a *biased* estimator, since its sampling distribution centers around 10, far from the true value.

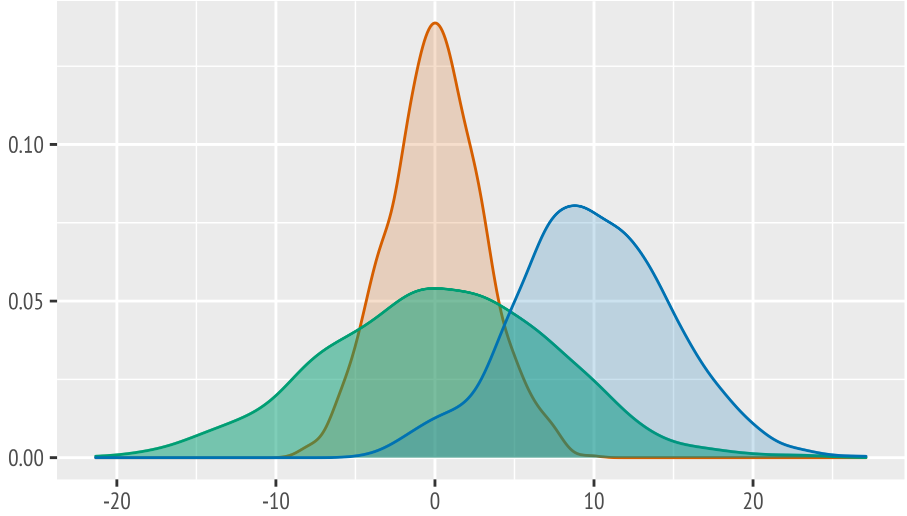

Figure 4: Distribution of three slopes.

Thinking in practical terms, the samples collected to estimate $\hat{\beta}$ in *blue*'s distribution likely come from a a problematic sampling process. And this can have several reasons, such as a poor data collection, and even a failure to capture the accurate sample that the study targets.

The Gauss-Markov theorem

Given the **Classical Assumptions I** through **VI**, the OLS estimator of a given coefficient β_k is the "*minimum variance estimator from among the set of all linear unbiased estimators of* β_k*, for* $k = 0, 1, 2, 3, ..., k$."

The **Gauss-Markov theorem** is summarized by the statement "OLS is **BLUE**." The latter term means **B**est **L**inear **U**nbiased **E**stimator. Let's break down each component of this acronym.

"Best" refers to *minimum variance*. That is, provided that **Assumptions I** to **VI** are satisfied, OLS will produce estimates with the lowest possible variance (i.e., the "thinnest" possible sampling distribution). "Linear" is just a reassessment of the linearity assumption contained in **Assumption I**. "Unbiased" is simply summarized by $\mathbb{E}(\hat{\beta}) = \beta_{true}$, discussed in detail in the last section. Lastly, you know from last week what an "Estimator" means.

In case we add **Assumption VII** to the Gauss-Markov theorem, then OLS becomes the best of *all* estimators, not just out of the linear ones. Then, *BLUE* becomes *BUE*.

However, as we will see in future lectures, it is not easy to fulfill all of these assumptions. But we will learn how to deal with the main violations of CLRM, and many of the possible corrections can be applied still using OLS. Therefore, it is a really powerful and flexible technique, which we will keep exploring until our semester is done.

Properties of OLS estimators

Summarizing what we have seen in this lecture, let us specify the four main properties of OLS estimators:

1. **Unbiased**: Our $\hat{\beta_i}$ estimates are centered around the *true* population values of β_i.
2. **Minimum variance**: The sampling distribution of $\hat{\beta_i}$ estimates are as narrowly concentrated as possible around the *true* value, given that the estimator is unbiased.
3. **Consistent**: As our sample size increases (i.e., $n \to \infty$), OLS estimates converge to the *true* population parameters.
4. **Normally distributed**: The distribution of $\hat{\beta}_i$ estimates can be summarized by

$$\hat{\beta}_i \sim \mathcal{N}[\beta_{\text{true}}\,, Var(\hat{\beta}_i)]$$

That is, our estimates are *normally distributed*, with the mean equal to its *true* value, and a given, constant variance.