An introduction to Ordinary Least Squares (OLS): Simple Linear Regression

Marcio Santetti | Fall 2023

Table of contents

After some quick review of statistical concepts, we will begin our content with the basic—and most important—technique in Econometrics: **linear regression**. You may have been introduced to it in your Stats courses, but we need to go deeper within its concept, in order to adapt it to real-world uses. The key argument is that the world that surrounds us produces *diffuse* and *fuzzy* data, and we are far from estimating empirical models free from **error**. We need to live with this fact, and the best way to treat this inherent error is by **minimizing** it.[1] The most popular method for that is **Ordinary Least Squares** (OLS), and we introduce this technique in the context of *simple linear regression*, that is, a model where we try to explain a variable of interest's behavior in terms of only *one single* explanatory factor.

[1] By *minimizing* something, keep in mind that we will use its mathematical meaning; that is, we try to make something (in this case, our error) not zero, but *as small as possible*.

The modern interpretation of *regression*

For our purposes, the term *regression* is concerned with the study of the dependence of one variable (dependent) on one—or more—other variable (explanatory/independent). We do this in order to predict the **population** mean or average effect of the independent variable(s) on some other variable of interest.[2]

In other words, a statistical regression compresses a **simple model** through which we wish to analyze how a change in one or more independent variable affects a dependent variable of interest. For example, in case we want to estimate the average effect of *education* on *wages*, the simplest possible model would be a simple regression, with *wages* as the dependent variable, and *education* as the independent variable. But how do we do this in practice? Keep going, we'll get there.

[2] As a quick historical note, in case you are curious to know more about the historical origin of the term "regression," make sure to check out the paper by J. Stanton listed on the "Additional Readings" module on `theSpring`.

Deterministic *vs.* stochastic relationships

In statistical relationships, we generally deal with **random variables**. This term should not be something new, but a quick reminder does not hurt: a random variable is a variable whose value is unknown until it is oberved, i.e., its outcomes are not *predictable*, thus following a *probability distribution*.

As an example, consider the relationship between crop yield and other variables, such as the amount of rainfall, sunshine, and fertilizer use. Such association is *statistical* in nature: although relevant, these factors will not enable an agronomist to *exactly* predict crop yield, due to possible errors involved in measuring these variables, as well as a host of other factors that collectively affect the yield at a certain point in time. Therefore, the **random** variability in the variable "crop yield" cannot be fully explained, regardless of the amount of explanatory variables we list.[3]

[3] What *other* variables would you consider relevant to explain crop yield, in addition to those already listed? And why?

Therefore, most—if not all—relationships in Economics and other Social Sciences involve *uncertainty* by definition. When there is **no** uncertainty involved in an association between two or more variables, we call it a *deterministic* relationship. The equation of a straight line ($y = ax + b$) that you learned in high school is an example. If you look at Figure 1, I have just plotted a few x and y points and connected them with a straight line. There is no uncertainty in this model, since all the points are captured by the line. Your Math teacher may have never told you this, but you were being taught *deterministic* Mathematics.

If you consider the 10 points in this graph as **data points**, and x and y as data variables, such relationship can be perfectly explained by a simple straight line. This looks great, but when we look at the real world, **uncertainty** is everywhere, so we are surrounded by *stochastic* relationships. Econometrics is concerned with these types of relations.

We cannot expect that real-world data behaves like a straight line, especially in the Social Sciences, where data are diffuse

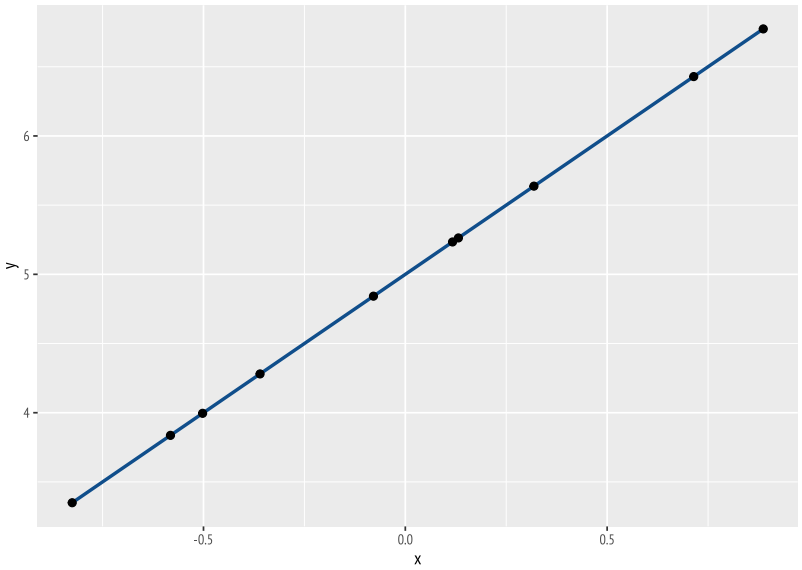

Figure 1: A simple straight line.

and fluctuate according to several different factors. A typical **scatter diagram** between two variables may look something like Figure 2.

The left panel illustrates the relationship between x and y, while the right panel fits a straight line to these points. Unlike the high school case, we cannot capture all the points only with a straight line. This is just a simple representation of how working with data and trying to explain it through statistical models is not an easy task. But we will do our best here.

Regression *vs.* correlation

Since in this course we will be mostly interested in statistical relationships, it is important to remark one point: a statistical relationship, in itself, cannot *logically* imply **causation**. The only thing regression can do is testing whether a significant relationship/association exists, as well as giving a quantitative

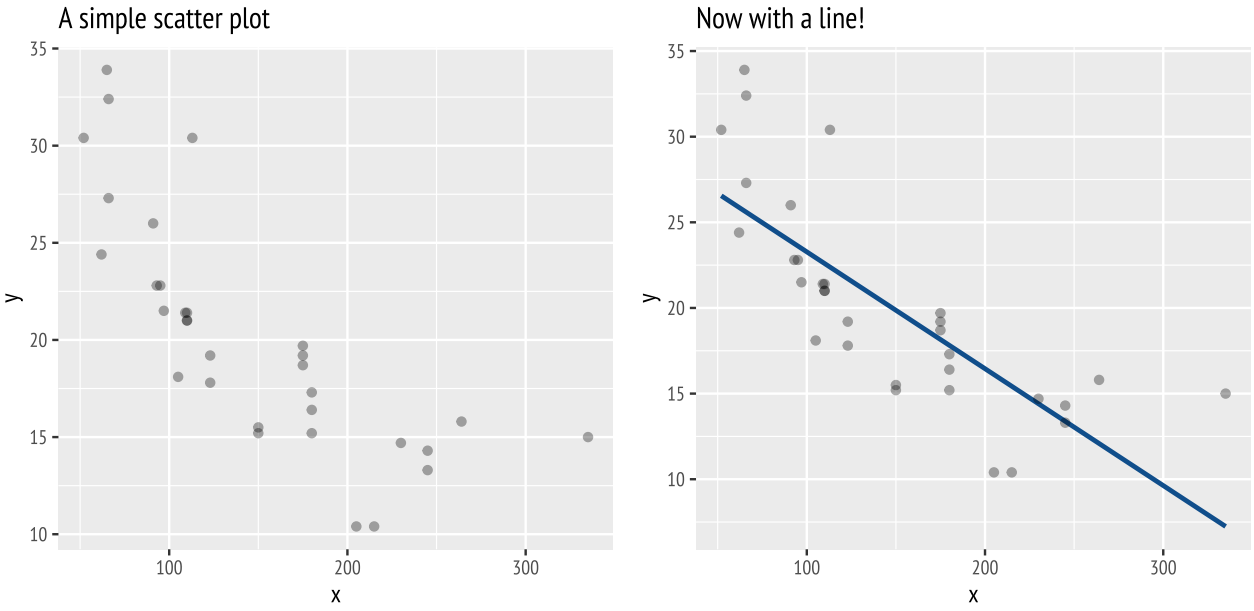

Figure 2: A stochastic relationship.

assessment on how the variables are related. However, no causality direction can be inferred from this methodology.

Furthermore, when we jump to regression analysis, the term *correlation* may be a temptation, but you should actually discard it when interpreting regression results. If you recall from your Stats classes, correlation analysis does not imply setting dependent or independent variables, i.e., x and y are treated symmetrically. When we study regression, we assume an inherent **asymmetry** between x and y, since we will be trying to explain changes in y in terms of changes in x. Even though these variables may be correlated (that is, having a *linear* relationship), our interests go way beyond simple correlation coefficients. Therefore, keep correlation out of your future regression interpretations.

Correlation implies a *linear* association between two variables, without an explicit call for which one is dependent, which one is independent. A more general measure of association is **covariance**, which computes how two variables are associated, *regardless of the shape*.

A note on terminology

The table below brings a few synonyms for both dependent and independent variables. These terms can be used interchangeably, with no loss of generality or meaning.

Dependent (y)	**Independent** (x)
Explained	Explanatory
Regressand	Regressor
Outcome	Covariate
Endogenous	Exogenous
Controlled	Control
Predictand	Predictor

Single-equation (simple) linear models

The simplest single-equation **linear regression model** can be represented by:

$$Y = \beta_0 + \beta_1 X$$

Let's break down every component of this equation. Once again, this should not be anything new up to this point, since you were probably introduced to this type of equation in your Stats courses. But let us review this topic and also implement our future notation. First, X and Y are our familiar independent and dependent variables, respectively. Moreover, we can see that this is the equation of a *line*, right? Okay, good.

The *stars* of our model—so far—are the β coefficients: β_0 and β_1. The first is called the *intercept*, or *constant*, term, and is simply the value that Y assumes when X is set to zero. The second is the *slope* coefficient, and represents the amount of Y that changes when X changes by **1 unit**. In other words, β_1

represents by how much the dependent variable changes, given a *marginal* change in the independent variable.

For those familiar with **calculus**, β_1 is simply $\beta_1 = \partial y / \partial x$.

The stochastic error term

Now, we introduce something new (hopefully): besides the variation in Y due to changes in X, it is almost certain that X cannot fully explain changes in Y. This additional variation comes in part from *omitted* independent variables (we will deal with this issue later), but, even if these omitted variables were added, do you think that only a chosen set of X covariates would be able to explain 100% of the variation in our variable of interest? If the answer is **no**, you are in the right place.

Such unexplained changes in Y may come from factors such as *measurement error*, incorrect *model specification*, or purely *random* events—that is, whose value is determined by *chance*. Does this ring a bell?

Therefore, this intrinsic and *inevitable* lack of explanation is fulfilled by the **stochastic error term** (also known as the **residual** term), in order to account for the variation in Y that is **not** explained by the independent variable(s). In simple language, the **error** captures our *ignorance* or *inability* to model the entire population model. However, its existence should not be an excuse for estimating poor models. We need to do our best in order to *minimize* this ignorance.

Let us, then, update our linear regression model by incorporating this error term, which we will denote, for now, by ε:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Now, we have a complete regression model! And it has a **deterministic** part, composed by "$\beta_0 + \beta_1 X$," as well as a **stochastic**

part, represented by the residual, "ε." Thus, the uncertain part of any statistical/econometric model is represented by an error term, which captures *everything that the explanatory variables are not capable of doing* to explain variations in Y, our dependent variable.

Since we will always be dealing with *random variables* in this course, let us apply the **expectations** operator (that is, compute the **Expected Value**, which you learned in your Stats course) to both sides of this equation:[4]

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X + 0$$

Put simply, we are just calculating the average value of Y, *given* (that is what the " | " symbol stands for) values of X. On average, the **expected value** of Y is given by the deterministic portion of our regression model,[5] while, on average, the value of our error term is *zero*. This is a crucial assumption we are making here, concerning the **distribution** of our error term. Figure 3 illustrates this latter point, simply showing that the central location of our ε friend is 0.

However, make sure you understand this point: this does not mean that, when we get to apply this concept to real data, our error term will show a value of zero. This is a statement about its *average*, assuming we can run the same regression model with several different samples of the same size for the same variables X and Y. This is related to the **sampling distribution** of the error term. If you do not recall what sampling distributions mean, make sure to also give it a quick review from Stats.

[4] Let us **demystify** what the term *Expected Value* means. It is simply the long-run average (mean) value of a *random variable*. Nothing else.

[5] If you do not recall the laws of Expected Value from your Stats class, make sure to give it a quick review, so you fully understand what is going on at this point. It is an important step to understand the basic foundations of our class.

How are ε and X related?

If ε and X are uncorrelated, then by definition these are not *linearly* related. However, the error term can be correlated with functions of X, such as X^2, for example—we will deal with these issues later. Therefore, correlation **is not** the most appropriate

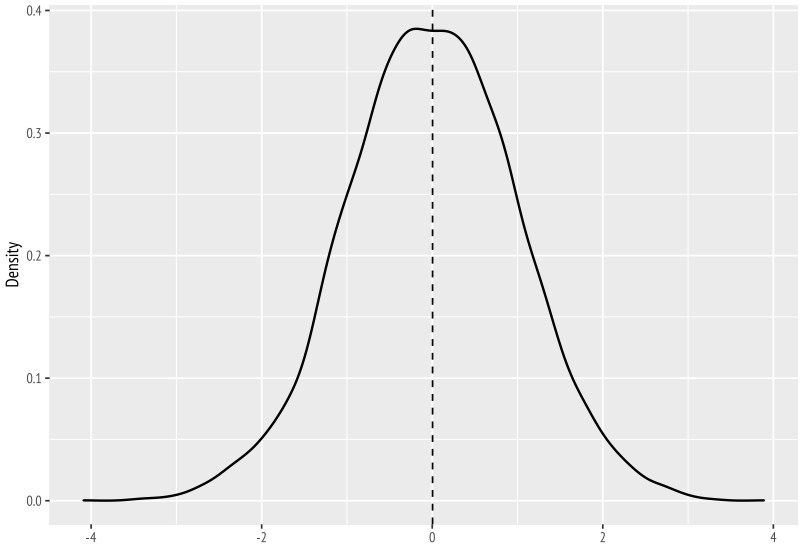

Figure 3: The residual's distribution.

approach to define this relationship. Instead, we can define the **conditional distribution** of ε, given *any* value of X. Let's see this:

$$\mathbb{E}(\varepsilon|X) = \mathbb{E}(\varepsilon)$$

The above equation simply states that, **on average, the value of** ε does not depend on X. In other words, these are **independent** of each other.

In order to make more sense of the above statement, let us look at a more realistic model, relating *wages* and *education*:

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \varepsilon$$

This is the simplest way to estimate how an individual's *wage* is affected by their *education*. But before we analyze the relationship between the residual and the independent variable, it is worth asking: since ε contains everything that is not explicitly

accounted for in the model, what is contained in there? Think about what is also relevant to explain wages.

Factors such as *years of experience*, *innate ability*, *gender*, *race*, and many other variables are included in the error term, since the only independent variable is *education*. To keep things simple, assume that ε is only *ability* (*abil*), and assume it is lying on the error term due to the impossibility of measuring it, or obtaining data.

Then, going back to our previous assumption, it requires that the average level of *ability* is the same, regardless of one's years of education. Illustrating:

$$\mathbb{E}(\text{abil}|\text{educ}) = \mathbb{E}(\text{abil})$$

If this assumption is true, then

$$\mathbb{E}(\text{abil}|\text{educ} = 8) = \mathbb{E}(\text{abil}|\text{educ} = 16) = \mathbb{E}(\text{abil})$$

This means that years of education are *independent*, **on average**, of what is contained in the error term, which is assumed to be only *ability*.

In case you believe that *ability* increases with *years of education*, congratulations, you are not a robot. However, this independence assumption is often *useful* and *theoretically necessary*. We will see this in more depth later on. If you are confused, that is a sign that you are paying attention. When we look at some applied examples, this will make sense.

The regression function (finally!)

The main goal of Statistics (and, therefore, Econometrics) is estimating **population parameters** based on **sample statistics**. In order to understand the latter, we begin with the former.

The *population* regression function

So far, we have worked with the notion of a **population** regression function, denoting Y and X as the *"true"*population values for the dependent and independent variables. Stating once again:

$$Y = \beta_0 + \beta_1 X$$

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$$

From these two equations, we know that the **average** value of Y changes with X. In case we had access to data from the entire population of interest, there would be no need for statistical techniques, such as the ones we will cover in this course. The solution is, then, working with an appropriate **sample** extracted from the overall **population**.

The *sample* regression function

We almost never have access to the *"true"* regression model defining Y. Rather, we work with **samples**. Let $\{(x_i, y_i) : i = 1, 2, 3, ..., n\}$ denote a random sample pair of size *n* from the whole population N. Then, we can write:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

for each $i = 1, 2, 3, ..., n$.

Notice a few changes in our notation. From now on, we will only deal in practice with *sample* data. For this course, lower-case letters, such as y_i and x_i will denote sample data, extracted from the overall population data represented by Y and X, respectively. You may have also noticed that the *sample error term* is denoted by u_i. We have also introduced *i* subscripts denoting each **individual** from the sample. These individuals will depend

on the research we are conducting: these may be households, houses, cats, cities, etc. Conceptually, however, all the terms of the sample regression functions are the same as before.

Additionally, we assume:

$$\mathbb{E}(u) = 0$$

$$Cov(x_i, u_i) = 0$$

That is, the **expected value** of the sample error term is zero, and the **covariance** between the independent variable(s) and the residual is zero. Recall that the *covariance* is a more general concept than *correlation*: while the latter is only concerned with *linear* relationships, the former defines relationships of *any form*.

What is the best method to estimate the sample regression function?

So far, we have only **conceptually** defined the regression function which will be the bread-and-butter of our class. But how do we calculate the β coefficients? Moreover, what is the best **straight line** that represents the relationships we want to estimate in the future, regardless of our specific research question?

As Econometrics practitioners, our main goal should be explaining variations in y_i due to changes in x_i, with the role played by our "ignorance," u_i, being as small as possible. So, how do we translate these words (*i*) graphically and (*ii*) mathematically?

I will answer these questions in class. You may stop reading this notes here. The rest will make sense after we discuss these issues in person.

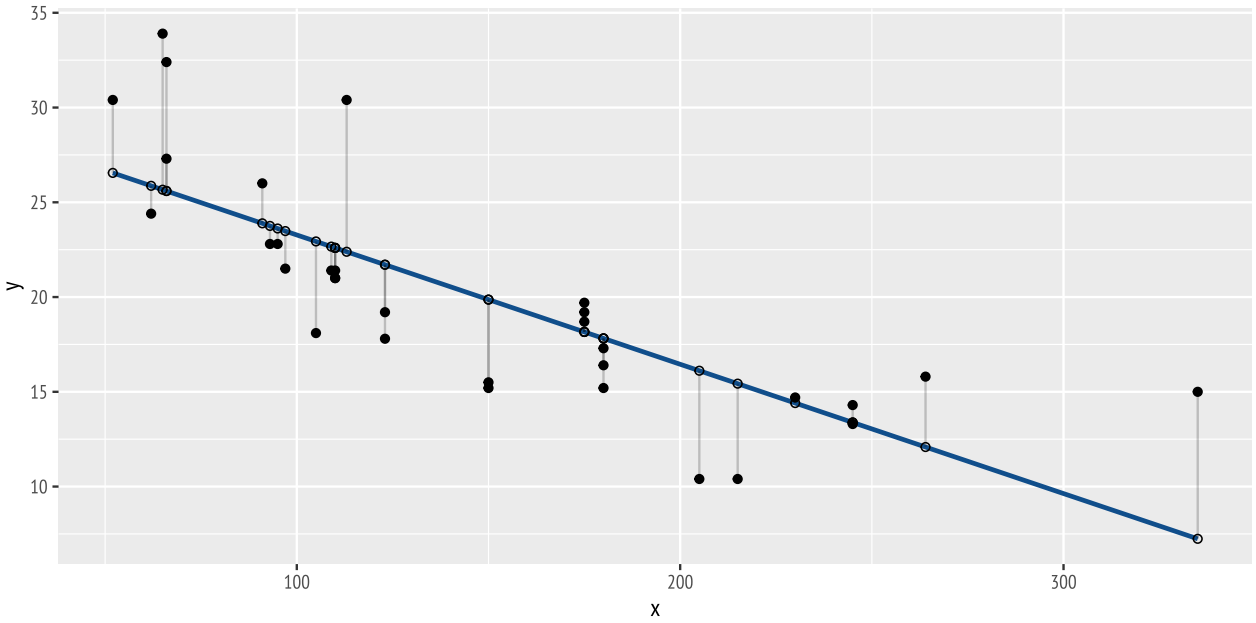

Figure 4: Residuals vs. regression line.

After this graph makes sense in your mind, you can see that **Ordinary Least Squares** (OLS) is the mathematical technique for obtaining the estimates of y_i, β_0 and β_1 that will make the residuals, u_i, as **small** as possible. In other words, it calculates the β coefficients that **minimize** the sum of the squared residuals of our model.[6]

OLS gives the *"best"* **estimator** possible for $\hat{\beta}_0$ and $\hat{\beta}_1$ under a set of assumptions, that we will explore in a couple of weeks. Also, pay attention to the *"best"* term stated in the previous sentence. We will se what this means in future lectures.

[6] **Why OLS, though**? Put simply, it is easy to compute (both manual and computationally); it is theoretically appropriate for statistical work, and has a great number of useful characteristics that we will explore little by little in our course.

Estimator: it is a mathematical technique that is applied to a sample to produce numerical estimates of the *"true"* population parameters. Do not confuse **estimator** with **estimate**: estimator is the *formula*, and an estimate is the *final numerical value* generated by this formula. To be clear, OLS is an *estimator*, as well as β_0 and β_1. On the other hand, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{u}$ are *estimates*.

Now, we are all set to start playing with real-world data. OLS will be our preferred estimator until the end of the semester. Just be aware that there are many other estimators out there, depending on the circumstances. However, the basic foundations of Econometrics are built upon OLS, and it is important to master it before moving on to more complex methods.

Keep these formulas in mind, we will use them!

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x_i, y_i)}{Var(x_i)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Last question: refresh your memory from your Stats class and **interpret the following regression model estimates**: $\hat{y}_i = 103.4 + 6.38x_i$. This interpretation routine will give you many points in future assignments and exams.