More on functional forms

Marcio Santetti

EC 339 | Fall 2022

Table of contents

Introduction

When we studied the Classical Assumptions of OLS, we established that our regression models are *linear* whenever the linearity in *parameters* is preserved. This is why we are able to incorporate nonlinearities, such as applying log-transformations to dependent and independent variables. But there's more we can do with our variables. In case we expect or detect possible nonlinear behaviors when plotting a scatter diagram of two variables, we can model those nonlinearities in many ways. Here, we will look at the most popular functional forms, so you can add these to your arsenal.

Regression though the origin

There are occasions when the **population regression model** assumes the following form:

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

Note that the regression is estimated *without an intercept*. In these cases, when $x_1 = 0$, $\mathbb{E}(y) = 0$. Although a **rare** case, there are certain relationships for which this is reasonable.

As an example, consider income tax revenues. When income (x) is *zero*, tax revenues (y) will also be *zero*, and it is reasonable to assume that these will not go below zero, only taking values over the positive domain of income. In case we assume a progressive taxation regime, we can illustrate it with Figure 1.

Unless recommended by **theory**, estimating a regression without an intercept is not usually recommended. Such practice is also more common in *simple regression* models, where the intercept having a value of 0 tends to make more practical sense.

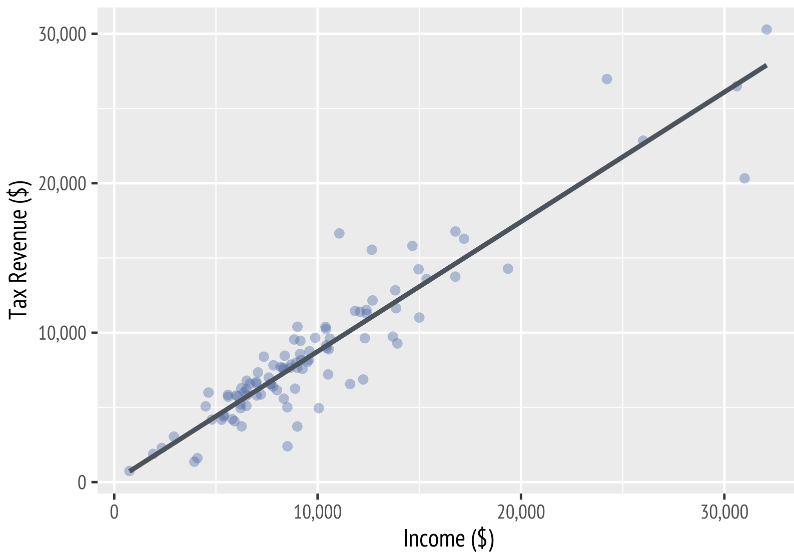

Figure 1: A progressive taxation regime.

Using squared terms

In some cases, the slopes of a regression model are expected to depend also on the **level** of the independent variable itself. For such cases, *polynomial* functional forms may be adequate. Consider the following *quadratic* model:

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

Before we move on, you have probably already noticed that interpreting slope coefficients is nothing but *computing the partial derivative* of y_i with respect to the desired variable, x_i. So, if we want to compute the effect on y of a one-unit increase in, say, x_2, we are basically calculating a partial derivative:

$$\frac{\partial y}{\partial x_2} = \beta_3$$

where the "∂" symbol denotes a partial derivative.

Now, what if we want to compute the effect on y of a one-unit increase in x_1? We do the same thing:

$$\frac{\partial y}{\partial x_1} = \beta_1 + 2 \cdot (\beta_2 \cdot x_1)$$

Since x_1 appears in the model both in levels and squared, we have to calculate the partial effect accordingly. Thus, we see that the effect of x_1 on y also depends on the **level** of x_1: as it changes, the effect on y will also change. This is something that is not captured in models with a lower polynominal order. Let us look at another example.

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 (\text{age}_i)^2 + u_i$$

As a worker gets older, the *difference* between age and age^2 increases dramatically. So, age would be *more important* at lower values than it would be at higher ones. In other words, the earning gains tend to decrease over time, as an employee gets older. This does not mean that wages will necessarily fall; but the *increase* in those gains tend to fall over time. In case we want to model for such behavior, we should use quadratic terms in our regression model.

As you are probably aware, this functional form produces *parabolas*, as illustrated in Figure 2. The panel on the left shows a *convex* function, where $\beta_1 < 0$ and $\beta_2 > 0$, whereas in the right panel, $\beta_1 > 0$ and $\beta_2 < 0$, generating a *concave* function. The fitted curves are shown in red, and for comparison we plot regression lines in blue for both situations where the quadratic term is not included. Notice how the red curve fits better the data than the blue straight line. This is the gain in explanatory power we obtain by improving our functional form.

As another example, consider the following model for housing prices (in logs):

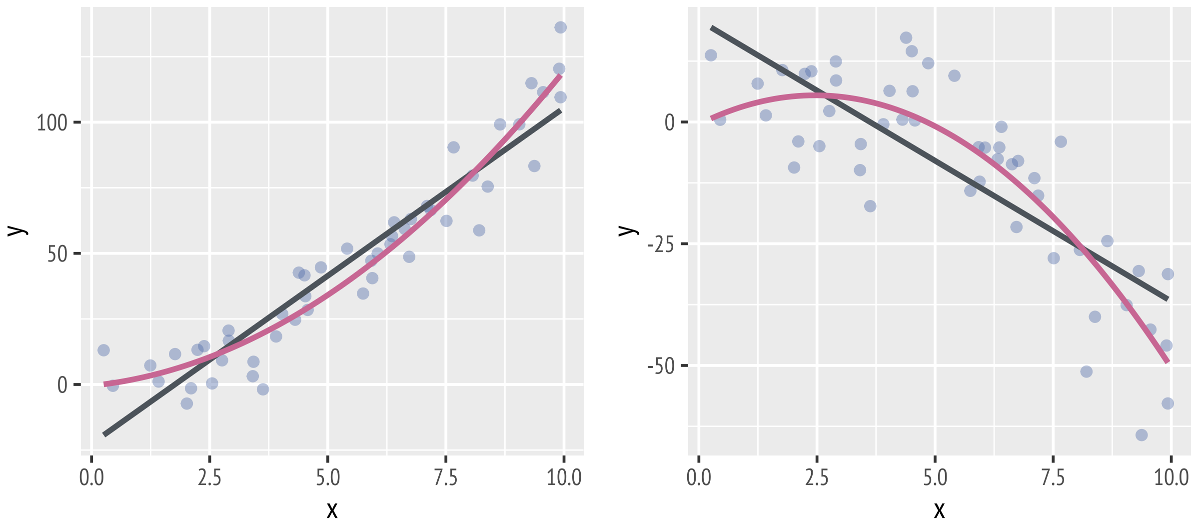

Figure 2: Quadratic relationships.

$$\widehat{\log(\text{price}_i)} = 11.26 + 0.23 \log(\text{dist}_i) - 0.82\, \text{rooms}_i + 0.089\, \text{rooms}_i^2$$

$$n = 506 \qquad \bar{R}^2 = .5$$

where $\log(\text{dist}_i)$ is the weighted distance between house i and downtown (in logs), and rooms_i is the average number of rooms per house.

Let us interpret the effect of *rooms* on *price*:

$$\frac{\partial\, \text{price}}{\partial\, \text{rooms}} = [\hat{\beta}_2 + 2 \cdot (\hat{\beta}_3 \cdot \text{rooms})] \times 100$$

Recall that, since this interpretation is in a *log-level* setting, we have to multiply the partial effect by 100. We already have the estimated coefficients for β_2 and β_3. But what to do with the *rooms* term that remains after the partial derivative calculation? Just plug in some value for it!

Let's work on this last sentence a bit more. In theory, we can plug in *any* value for *rooms*, and we will obtain a final partial effect to interpret. However, we should use a *valid* number of *rooms*, in order to have a consistent analysis. One interesting value to use is the *average* number of rooms in the sample. It can also be the *median*, or the *mode*, or any reasonable value. What matters is that the value you choose is consistent with the used *sample* and with the *problem* at hand. For now, we'll stick with the mean. From this sample, the average number of rooms is 6.28. So, we use $rooms = 6$:

$$\frac{\partial \, price}{\partial \, rooms} = [-0.82 + 2 \cdot (0.089 \cdot 6)] \times 100 = 24.8$$

Therefore, all else constant, and starting from a house with 6 rooms, one additional room in a house increases the price of a house, on average, by 24.8%, based on our sample. Thus, in our model we have included the actual number of *rooms* into the interpretation of its effect on housing prices by including a quadratic term. Nice, isn't it?

Inverse form

The next functional form is the *inverse form*. It is used whenever the impact of a particular independent variable is expected to approach *zero* as the variable approaches infinity. Note: the *effect* (that is, the associated β coefficient) approaches zero, not the variable *itself*.

To model this effect, we use the *reciprocal* (or inverse) of one or more of the control variables. Let us look at an example:

$$y_i = \beta_0 + \beta_1 \left(\frac{1}{x_{1i}} \right) + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

Here, we are assuming that the effect of x_1 on y approaches zero as x_1 increases. Depending on the sign of its associated coefficient, in this case β_1, we have different curves. In Figure 3, we represent the fit of the model to both situations: when $\beta_1 > 0$, in red, and when $\beta_1 < 0$, in blue.

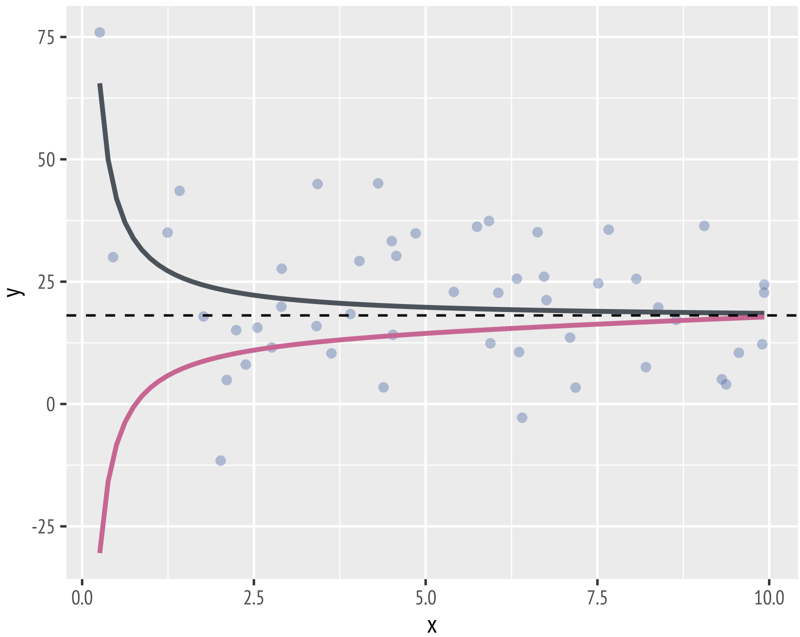

Figure 3: Inverse form.

To calculate the partial effect of x_1 on y, we once again use a partial derivative, appealing to the quotient rule:

$$\frac{\partial y}{\partial x_1} = \frac{-\beta_1}{x_1^2}$$

Lastly, an example. If we suppose the unemployment rate's (u_t) effect on wages (w_t), after certain levels, tends to be zero, we can model this situation as follows:

$$\hat{w}_t = .00679 + .1842\left(\frac{1}{u_t}\right)$$

Assuming an unemployment rate of 5%, the partial effect will be

$$\frac{\partial w_t}{\partial u_t} = \frac{-\hat{\beta}_1}{u_t^2} = \frac{-(0.1842)}{(0.05)^2} = -73.68$$

Thus, all else constant, if the unemployment rate increases by 1 percentage point, wages will, on average, decrease by 73.68 dollars.

Interaction terms

Sometimes, it is natural for the partial effect, elasticity, or semi-elasticity of the depedent variable with respect to an explanatory variable to depend on the *magnitude* of *another* independent variable.

Consider housing prices once again. A house's number of *rooms* definitely affects its price, but don't you think that such effect is also dependent on the *size* of the house? For instance, it is likely that a house with a larger square-footage will be more expensive than a smaller house, but with the same number of bedrooms, at least on average and *ceteris paribus*.

In case we want to model such situation, we use *interaction terms*, that is, we **multiply** two independent variables together. Consider the following example:

$$\mathtt{price}_i = \beta_0 + \beta_1 \mathtt{sqrft}_i + \beta_2 \mathtt{bdrms}_i + \beta_3 \mathtt{sqrft}_i \cdot \mathtt{bdrms}_i + \beta_4 \mathtt{bthrms}_i + u_i$$

where $\mathtt{sqrft}_i$ is the average square-footage, $\mathtt{bdrms}_i$ is the average number of bedrooms, and $\mathtt{bthrms}_i$ is the average number of bathrooms for each house i.

The partial effect of $\mathtt{bdrms}_i$ on $\mathtt{price}_i$ is calculated by

$$\frac{\partial \mathtt{price}}{\partial \mathtt{bdrms}_i} = \hat{\beta}_2 + \hat{\beta}_3 \cdot \mathtt{sqrft}$$

Once again, to complete the interpretation, we simply plug in a *useful* value of `sqrft`. Usually, the a measure such as the *mean* is mostly recommended. For this example, if $\hat{\beta}_3 > 0$, an additional bedroom yields a *higher* increase in prices for *larger* houses. In other words, if statistically significant, there is an *interaction* effect between a house's square-footage and the number of bedrooms.

Using *dummy* variables

Not every variable that we consider including in a regression model can be *quantitatively measured*. For example, how do we measure factors such as *gender*, *race*, *religious beliefs*, an so on? These are **qualitative** variables, which are not easily translated into numerical values. However, such covariates can aggregate several interesting features to our models, and that is the reason we are able to include these by using **binary** (or **dummy**) variables.

A **dummy** variable, by definition, takes on the values of *0* or *1*, depending on a *qualitative attribute*. For example, we could then model *gender* as taking the value of 1 if the individual is *female*, and 0 if *male*; for for *religion*, 1 if *LDS*, and 0 *otherwise*, and so on. Furthermore, we could use binary variables to model for a variable fulfilling some kind of **criterion**, such as whether an individual has attended college or not, committed felony, etc.

In our course, we will restrict our analysis of qualitative variables to the *binary* case, but be aware that it is possible to include more categories for qualitative variables.

Let us consider regression models that include binary covariates. These can appear in two forms: **intercept** and **slope** dummy variables.

Intercept dummy variables

Let us start with the simplest case for including dummy variables in a regression model. When the binary variable appears *by itself* in a model, we have an **intercept dummy variable**. Here's an example:

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

where

$$D_i = \begin{cases} 1, & \text{if the } i^{th} \text{ observation meets a particular criterion} \\ 0, & \text{otherwise} \end{cases}$$

Since we will be working with *binary* cases, we always want to use **one fewer dummy variable than the number of conditions**. Thus, if 2 conditions, 1 dummy variable. The "omitted" condition—that is, when $D_i = 0$—, forms the *basis* against which the included condition—$D_i = 1$—is compared.

Lastly, the coefficient on D_i, $\hat{\beta}_3$, is interpreted as the effect of the included condition, relative to the omitted condition. Therefore, notice that we **do not** interpret binary variables the same way we do with "regular" variables. When interpreting dummy variables, we are comparing the category/criterion representing $D_i = 1$ to the "base" category/criterion, $D_i = 0$, and its effect on the dependent variable, and not the outcome of a 1-unit increase in the criterion/category on the dependent variable.

Let us look at a more specific example, relating participating in a committee and the number of new articles written in a semester for faculty members:

$$\hat{A}_i = .37 + .81pp_i - .38C_i$$

$$n = 25 \qquad \bar{R}^2 = .45$$

where

$$C_i = \begin{cases} 1, & \text{if the } i^{th} \text{ faculty member is part of a committee} \\ 0, & \text{otherwise} \end{cases}$$

and pp_i is the amount of papers written by faculty member i before joining the committee.

The effect of joining a committee is calculated by

$$\frac{\partial A}{\partial C} = \hat{\beta}_2 = -.38$$

This means that, all else constant, faculty members who have joined a committee write, on average, .38 papers *less* than those who do not join a committee. Thus, the negative sign indicates a relative *disadvantage* for those who commit to a faculty group, having less time to write. In case the sign of $\hat{\beta}_2$ were positive, it would be the opposite case.

The next figure illustrates how only the *intercept* changes when $C_i = 1$ and when $C_i = 0$. We plot A_i against pp_i, and depending on the value the dummy variable takes on, only the intercept changes, with the slope ($\hat{\beta}_1 = .81$) remaining the same. That is why we call the dummy variable here as an **intercept** variable.

The *blue* line depicts the effect of previous papers written on the amount of new ones when $C_i = 0$, while the *red* line illustrates when $C_i = 1$. The distance between these two lines is given by β_2, that is, the dummy variable's estimated coefficient.

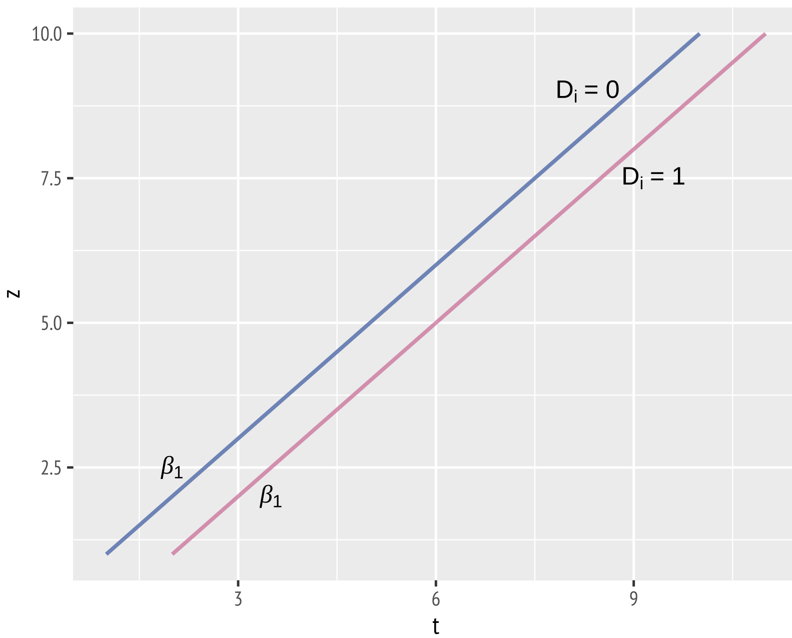

Figure 4: Intercept dummy variable.

Slope dummy variables

You have already been introduced to *interaction terms*, that is, when we multiply two independent variables together. A **slope** dummy variable is nothing but an interaction term, this time multiplying a dummy variable with another independent variable. And the latter may be a continuous, discrete, or even another dummy variable. The choice depends on our research question.

When including slope dummy variables, we usually do so also including the dummy by itself in the model, thus including an intercept dummy variable as well. Let us look at an example:

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

Now, in addition to having an intercept, we also have a slope dummy variable, with the interaction between x_1 and D. We should set up our regression like this whenever we consider that the effect of an independent variable on y also depends on some *qualitative* factor.

Before we explain the latter sentence in more detail with an example, consider Figure 5, where we depict two regression lines for the above model: the one in blue when $D_i = 1$, and the one in red when $D_i = 0$. Notice that the *slopes* are now different. Where do these different slopes come from? Let us investigate the partial effect of x_1 on y:

$$\frac{\partial y}{\partial x_1} = \hat{\beta}_1 + \hat{\beta}_3 D_i$$

Nothing surprising here, right? But recall: D_i can be either 0 or 1. Thus, when $D_i = 1$,

$$\frac{\partial y}{\partial x_1} = \hat{\beta}_1 + \hat{\beta}_3$$

But when $D_i = 0$, the derivative becomes

$$\frac{\partial y}{\partial x_1} = \hat{\beta}_1$$

That is why we have different *slopes*, as illustrated in the graph.

To wrap up these notes, let us consider a model for earnings, controlling for *experience* and *gender*:

$$\text{earnings}_i = \beta_0 + \beta_1 \text{exp}_i + \beta_2 G_i + \beta_3 \text{exp}_i G_i + u_i$$

where

$$G_i = \begin{cases} 1, & \text{if the } i^{th} \text{ individual is female} \\ 0, & \text{otherwise} \end{cases}$$

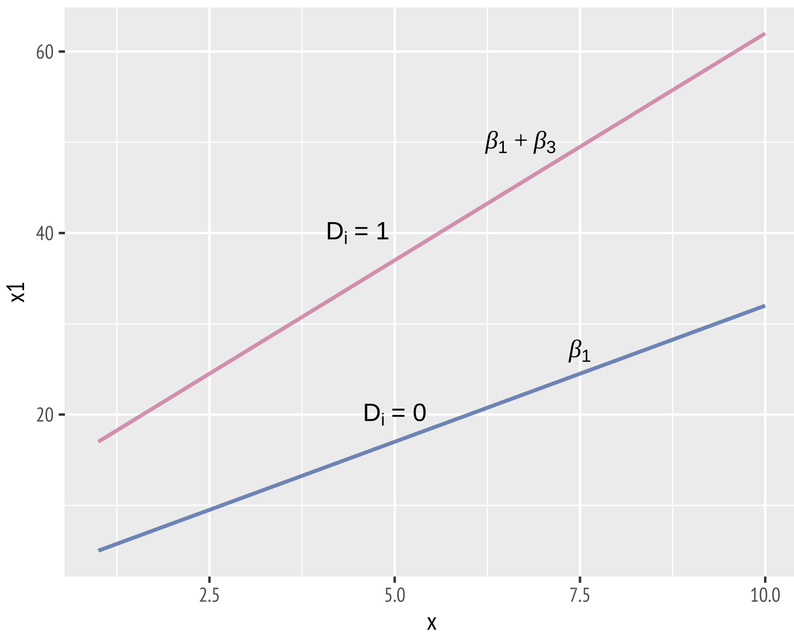

Figure 5: Slope dummy variable.

In case we consider that the effect of one additional year of *experience* on a worker's earnings is *also* dependent on gender, we should include an interaction term, denoted by the slope dummy variable with coefficient β_3 in the above model.

The β_3 coefficient captures the **differential impact** of an extra year of experience on earnings between non-female and female employees. In other words, if we select two individuals from our sample, one non-female and one female, with the *same* years of experience, is there an earnings differential between them? $\hat{\beta}_3$ will tell us that, and if it is statistically significant, then we have a gender *differential* between male and female workers, based on our model and on our sample.

As an exercise, compute the effect of *gender* on earnings, and also the effect of *experience* on earnings from the above regression.