Adding more Variables to your Model: Multiple Linear Regression

Marcio Santetti | Spring 2023

Table of contents

Introduction

Sometimes, a regression model with only one control variable is enough for our analysis. However, we can give at least **two** reasons for including more independent variables in our model: *first*, we were introduced in the last lecture to the *stochastic error term*, which includes all other variables and factors that are not explicitly considered on the regression's right-hand side. If an omitted variable is important to explain variations is the dependent variable, this brings many problems, which we will investigate later on. *Secondly*, in the social sciences, events change due to a myriad of other events, hardly so just from one single event. Take the *wage-education* relationship, for instance. Variables such as *gender*, *experience*, *race*, *tenure*, and many others, must at least be considered to be included in our models, in order to reduce the inherent amount of **error** that a regression analysis comprises.

The natural evolution of a simple regression are **multiple regression models**, which we will cover in this lecture. In addition, we will discuss how to better assess the *quality* of any regression model (regardless of having one or more independent variables), as well as working with *measurement unit* transformations and *functional forms* that can help us extract more information from our OLS estimators.

Recall that whatever a regression model *does not* explicitly include will be part of the stochastic error term.

How to interpret multiple slope coefficients?

A multivariate regression model with k independent variables can be represented by the following equation:

$$y_i = \beta_0+\beta_1 x_{1i}+\beta_3 x_{2i}+\beta_3 x_{3i}+...+\beta_k x_{ki}+u_i \qquad \forall i = 1, 2, 3, ..., n$$

Notice that we are including two indexes (subscripts) next to each independent variable. One indexes the number of

each slope coefficient—and the order in which you include your slope variables does not matter here—, and the other indexes the variable to its respective *individual* (*observation*), denoted by i. Thus, if our sample size is n, we will have one observation of each variable $(x_1, x_2, ..., x_k)$ for each i individual (from individual 1 until the n^{th} individual). Lastly, the $\forall$ symbol next to the above equation is read as "for all." Therefore, it should be read as "for all i, ranging from the first until the n^{th} observation."

If you are still struggling with this notation, consider your data set as a spreadsheet. Each **column** represents a variable, while each **row** brings individual information for the corresponding variable. Thus, columns are the y and x variables, and rows bring data on each i individual contained in your data set. Using subscripts just compresses information, so we do not need to write down n different regression equations. Notation is important and it is meant to simplify our lives, and thus I want to demystify many mathematical issues that you may have come across in your life that could have been made much simpler. Feel free to shoot me an email if any part of our mathematical notation is not clear.

After we estimate $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$, how do we interpret these coefficients, now that the model is a little more complex than before? The answer is really simple: in the same way! These coefficients still indicate the **change** in the dependent variable associated with a **1-unit increase** in the respective independent variable, *holding constant the other independent variables in the equation*. For example, if we want to interpret the $\hat{\beta}_2$ coefficient, it represents the change in y associated with a 1-unit increase in x_2, holding $x_1, x_3, ..., x_k$ constant. This *ceteris paribus* (all else constant) assumption is a *partial equilibrium* interpretation, and you must have been introduced to this idea elsewhere. However, a very important *warning*: this *all else constant* assumption **does not** apply to any variables that might have been **omitted** from the model, therefore lying in the residual term u_i.

The intercept coefficient, $\hat{\beta}_0$, still could be interpreted in the

same say as in the simple regression case, but it is not very useful anymore. The role played by the intercept term in multiple regression analyses is more mathematical than statistical, and we usually do not even bother about its numerical value. However, it is still really important to include it in our models.

Let us look at an example with $k = 2$ independent variables:

$$\widehat{CB}_i = 37.4 - 0.88P_i + 11.9Y_i$$

where CB_i is beef consumption for individual i (in pounds), P_i is the price of beef (in dollars) paid by individual i, and Y_i is the i^{th} individual's disposable income (in thousands of dollars).

The estimated value of $\hat{\beta}_1$ is -0.88. This means that, holding an individual's disposable income constant, a one-dollar increase in the price of beef *decreases* beef consumption by .88 pounds, on average. Likewise, $\hat{\beta}_2$ is 11.9, meaning that, *ceteris paribus* (i.e., holding the price of beef constant), if an individual's disposable income increases by one thousand dollars, her beef consumption will increase, on average, by 11.9 pounds. Notice that we have to respect the **measurement units** by which the variables are defined here, adapting the 1-unit change to the respective way in which they are measured.

In the above interpretation, however, other factors, such as the *price of chicken*, for instance, **cannot be held constant**, since it is included in the error term, in case you believe this variable is important to explain variations in beef consumption, and it is omitted from the model.[1]

The key detail from multiple regression models is that we have one slope for each independent variable, as we can see in Figure 1. This model, therefore, allows for a *negative* slope (with respect to P_i), and for a *positive* slope (associated with Y_i).

[1] Recall from your Micro classes that the price of *substitutes*, in addition to other factors, is relevant to explain the demand for one good.

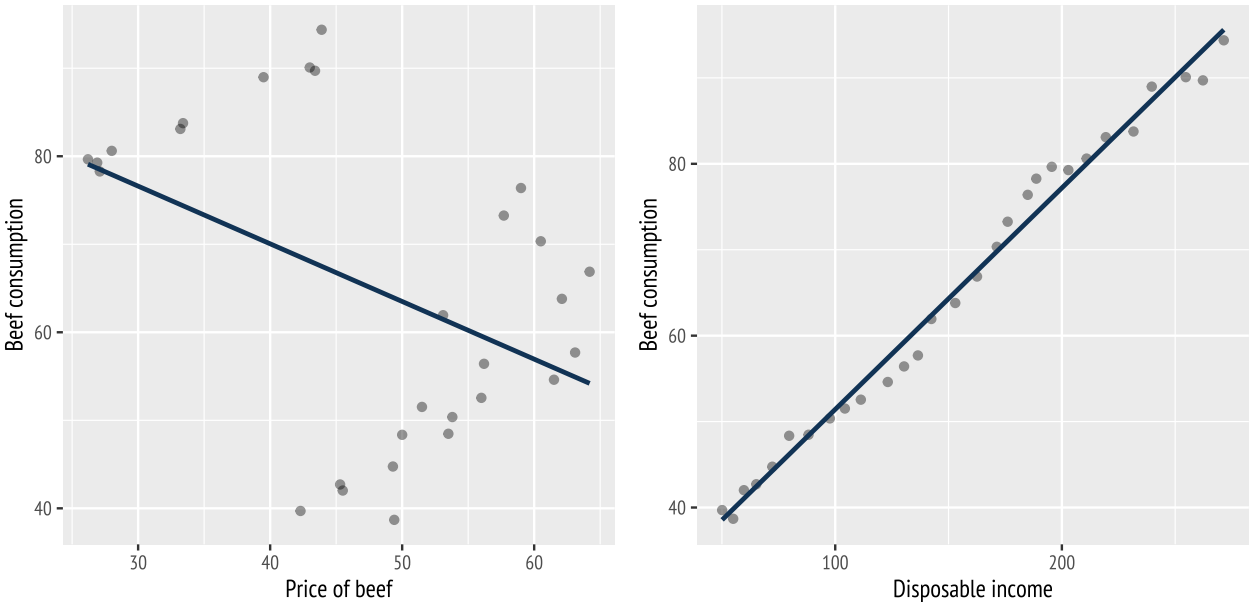

Figure 1: Individual slopes.

Total, Explained, and Residual Sum of Squares

After our regression is estimated, we would like to assess *how well* our model fits the data. You may have been introduced to the *coefficient of determination*, R^2, elsewhere, and it assesses the **variation** in y caused by **variations** in our independent variable(s). And you have probably learned that it is calculated by squaring the correlation coefficient, usually known as r, thus ranging between 0 and 1, or 0 and 100%.

We will now derive the R^2 from a *regression perspective*: the squared deviations of y around its mean are a measure of the amount of variation to be explained by the regression model. These are called **Total Sum of Squares** (TSS):

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where $\bar{y}$ denotes the mean of the dependent variable. The above

formula simply calculates the deviation of each observation of the dependent variable (y_i) from its mean ($\bar{y}$). We add up this squared difference—avoiding negative values—for our entire sample, whose size is n.

For OLS models, the TSS has **two** components: one variation that can be explained by the model, and one that cannot:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{u}_i^2 \implies TSS = ESS + RSS$$

The above procedure is known as *decomposition of variance*, and it basically decomposes the deviations of y relative to its mean between what **is** explained by our regression model and what **is not**. The first is denoted as the **Explained Sum of Squares** (ESS) and the second, as the **Residual Sum of Squares** (RSS). The *smaller* the RSS is, relative to TSS, the *better* the model fits the data.

The coefficient of determination, R^2

The three estimates presented in the previous section can be used to derive a regression's **coefficient of determination** (R^2):

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

with R^2 lying between 0 and 1. We can also present its value in percentage units, therefore lying also between 0 and 100%.

The R^2 measures the **goodness-of-fit** of a regression model. In other words, the **variation** (%) in the dependent variable explained by our regression model. In case the regression only has one independent variable, the R^2 illustrates the variation (%) in the dependent variable explained by variations (%) in the

dependent variable. Lastly, since OLS provides the parameters that *minimize* the RSS, it provides the **largest** possible R^2, given our estimated model.

Next, we see examples of a low and a high R^2. Notice that, the more distant the data points are from the regression line, the lowest a model's goodness-of-fit.

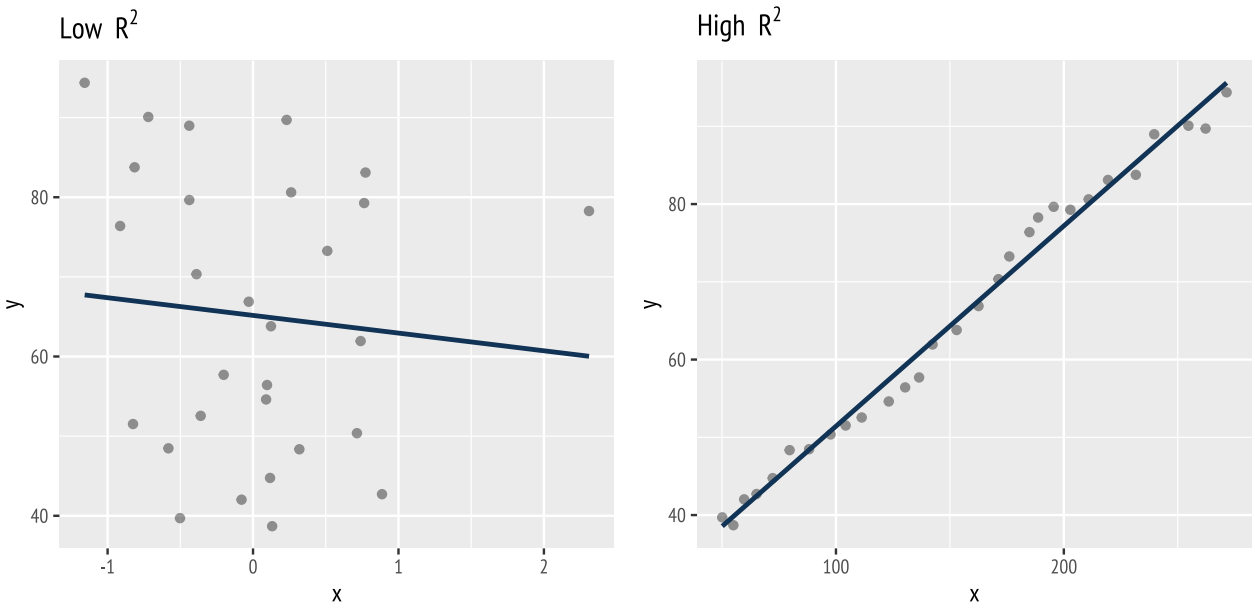

Figure 2: Low and high R-squared coefficients.

The adjusted R^2, $\bar{R}^2$

The R^2, however, is not bulletproof. If one keeps adding control variables (with many of them being, actually, unnecessary) to a regression model, this will **never** decrease the R^2. From the formula:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

If we keep adding regressors to our model, the formula's denominator remains the same, since nothing happens to y or its mean. The numerator, on the other hand, can only decrease or, at most, stay the same, since we are, in theory, "removing elements" from the error term. Thus, the R^2 does not fall. This is not good to assess a model's quality.

The addition of unnecessary variables to a model not only inflates it with useless regressors, but also requires the estimation of additional β coefficients. This fact **decreases** the **degrees-of-freedom** (DOF), i.e., the excess of observations (n) relative to the number of estimated coefficients ($k + 1$). Given that we denote the intercept with a "0" subscript, the total number of coefficients a regression estimates is denoted by $k + 1$; thus, k only refers to the *slope* coefficients of our model. Pay attention to this fact, it will be important in a moment.

The act of *adding* another control variable to a model must be compared to the *decrease* in degrees-of-freedom before a decision can be made with respect to its statistical impact. To address this problem, we compute a version of the R^2 measure that *adjusts for degrees-of-freedom*. We call it the **adjusted** R^2, denoted as $\bar{R}^2$:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2/(n-k-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

Notice that the formula is basically the same as the one for the "standard" R^2. The adjustment appears by adding denominators to the RSS ($n - k - 1$) and the TSS ($n - 1$). We are basically *normalizing* the RSS and TSS by their respective degrees-of-freedom. If you carefully look at the right-hand side of the formula above (ignoring the "1 -" part for now), you will notice that the denominator is nothing but the **variance** of y, the dependent variable. But what is the numerator? It is nothing but the **variance of the estimated error term**, $\hat{u}$. The denominator of its variance is $n - k - 1$ because this is

the number of degrees-of-freedom a regression requires to be estimated: from all n observations, we estimate k slope coefficients and 1 intercept. For y's variance, our number of degrees-of-freedom is $n - 1$ simply because we are just losing 1 observation to calculate its sample mean, $\bar{y}$, which is part of its variance formula.

Thus, what the adjusted R^2 calculates is the variance of our model's error term, *weighted* by (or *relative* to) our dependent variable's variance. Then, we subtract it from 1 and obtain the **variation (%) in the dependent variable, around its mean, that is explained by the estimated regression model, adjusted for degrees-of-freedom**. Notice how its interpretation is similar to the "standard" R^2, only adding the terms "around y's mean" and "adjusted for degrees-of-freedom." **Do not** forget these terms, since these radically change your interpretation.

Assessing the quality of a regression equation

Now that we were introduced to the two main goodness-of-fit measures, we are able to ask ourselves: How to assess the *validity* of a regression's estimates? Be aware that the statistical package you are using to estimate your models accepts *anything*: it does the "dirty job" of doing the math for you, but it does not bother about how good or how bad it is. Moreover, it does not *interpret* anything. This is *our* job.

As stated in our first week's lecture, where the main features of Econometrics were introduced, as well as the "classical" workflow for an applied work, the **key step** is to spend time thinking about what to expect **before** any estimation starts. This way, we are more prepared to get any results from our estimations.

Some aspects to ponder:

1. Is our model supported by *theory*?

2. Does it *fit* well the data we have at hand?
3. Is our *sample size* good enough for our purposes?
4. Is OLS the best *method* to answer our research question?
5. Are *all* important (relevant) variables included in our model?
6. Is the *functional form* appropriate to answer our questions?

Step 1 must always be your **starting point**. Your study will never be relevant if solely based on *personal conviction/beliefs*. These must be backed up by theory, and we have a pretty good arsenal of theories in our discipline, don't we? This step is intimately connected with Step 5, since theory will inform you what variables are worth including in your model, without unnecessarily wasting DOF.

Step 2 is quickly assessed after the regression's estimation. Now that you are aware of the limitations of the "standard" R^2, always pay more attention to the **adjusted** R^2, playing around with different specifications of your model to evaluate how $\bar{R}^2$ reacts; usually, more *parsimonious* (that is, simpler) models do way better than models with too many regressors. Simplicity is always preferred, relative to complexity, in regression analysis.

For the sample size, a minimum of $n = 30$ observations is the usual procedure, since it is in accordance with the *Central Limit Theorem* (CLT). But the more data you can gather, the better the properties of OLS will develop in your estimation. Regarding Step 4, OLS will not always be the most appropriate estimator for our purposes. For example, when our dependent variable is *binary* (i.e., either taking on a value of 1 or 0, according to a certain criterion), OLS does not work well. We will come back to this issue in future weeks.

Lastly, Step 6: *functional forms*. This is the subject of the next section.

Units of measurement & functional forms

For this section, we address two questions:

- How does changing the *units of measurement* of y and/or x affect OLS estimates?
- How to incorporate popular *functional forms* used in Economics into regression analysis?

Changing the dependent variable's measurement

Consider the following model:

$$\text{salary}_i = \beta_0 + \beta_1 \text{roe}_i + u_i$$

where *salary* is annual salary, in thousands of dollars, for individual i, and *roe* is the return on equity for the CEO's firm for the previous 3 years. The latter is a profitability measure, defined as net income as a percentage of common equity.

After we estimate the model, we have:

$$\widehat{\text{salary}}_i = 963.191 + 18.501 \text{ roe}_i$$

Now, suppose we decide to change the measurement of *salary*. Let *salardol* be the salary measured in dollars. Thus, $\text{salardol}_i = 1,000 \cdot \text{salary}_i$ for all i individuals in the sample. Since we changed the dependent variable's measurement, we altered the left-hand side of our regression. In order to keep both sides equal, we simply perform the same operation on the right-hand side. Therefore, if we multiplied the right-hand side by 1,000, we do the same on the right-hand side. Then, the regression output becomes:

$$\widehat{\text{salardol}}_i = 963,191 + 18,501\ \text{roe}_i$$

The interpretation stays the same as before. We only have to adjust it for the new coefficients and measurement units.

Generally, if the **dependent variable** (y) is multiplied by a constant c, the β coefficients are also multiplied by c.

Changing an independent variable's measurement

Let us go back to the original regression, with *salary* as the dependent variable. Now, let $roedec_i$ be defined as the decimal equivalent of roe_i. Thus, $\text{roedec}_i = \text{roe}_i/100$. Given this latter definition, cross-multiplying both sides gives $\text{roe}_i = 100 \cdot \text{roedec}_i$. Then, the estimated original coefficient on $\text{roe}_i = 18.501$ becomes $\text{roedec}_i = 18.501 \cdot 100 = 1,850.1$.

Then, the regression output becomes:

$$\widehat{\text{salary}}_i = 963.191 + 1,850.1\ \text{roedec}_i$$

For our example, the intercept coefficient (β_0) remains the same, since both $\text{roedec}_i = 0$ and $\text{roe}_i = 0$ mean the *exact same thing*. Lastly, the **goodness-of-fit** measures are **unchanged** by these transformations (either to dependent or independent variables), since the variables are just being normalized differently.

If an **independent variable** (x_i) is multiplied (divided) by some *nonzero* constant c, then the corresponding OLS **slope coefficient** is divided (multiplied) by c.

Incorporating nonlinearities in a regression model

Some economic relationships are **nonlinear** by nature. Take the example of wages *vs.* years of experience. Consider the two cases in Figure 3. Which one do you think better illustrates this relationship: the first, where wage increases *monotonically* with experience; or the second, in which we also observe an increasing association, but with *marginally decreasing* increments?

If you chose "Case 2," you live in the real world. In case we lived in the world of "Case 1," there would be no uncertainty:

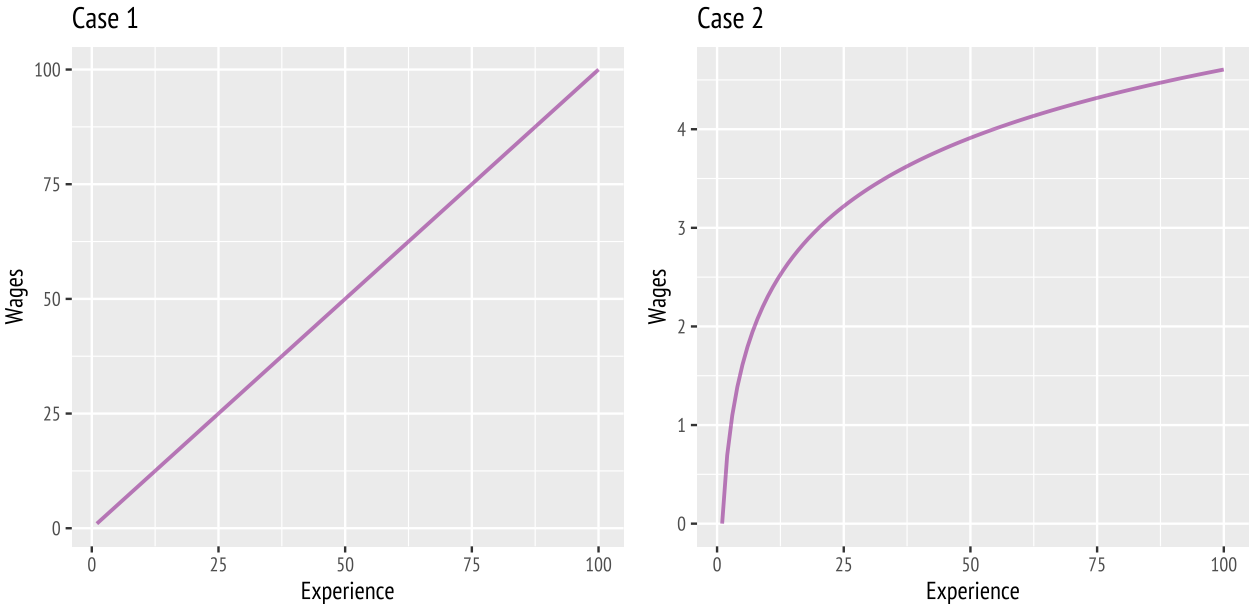

Figure 3: Linear and nonlinear relationships.

just stay on the job for the rest of your life, and your wage will increase no matter what. We know that this is basically impossible.

We will start playing around with different *functional forms* slowly. But be aware: we are able to model situations as in "Case 2" *without* compromising OLS properties. So far, we have only dealt with models where the dependent and independent variables are taken in *levels*, that is, without changing their functional definitions. For now, we will study **log-level**, **log-log**, and **level-log** models, and more functional forms are yet to come.

Log-level models

Consider now the *salary-education* relationship. The following regression model assumes a **level-level** functional form, that is, with all variables in their non-transformed *levels*:

$$\widehat{salary_i} = 7.5 + .54\ educ_i$$

The slope coefficient for $educ_i$ is .54. This means that, for every additional year of education, one's salary will increase by 54 cents. It does not matter whether one is going from the 10^{th} to the 11^{th} year of education, or from the 17^{th} to the 18^{th}. This increment of 54 cents is **constant**. The sign makes sense, but its interpretation does not correspond to one's theoretical expectations.

What we expect is that one's salary at least increases by a **constant percentage**, the more educated they become. Graphically, it looks somewhat like the graph below, relating two variables, x and y.

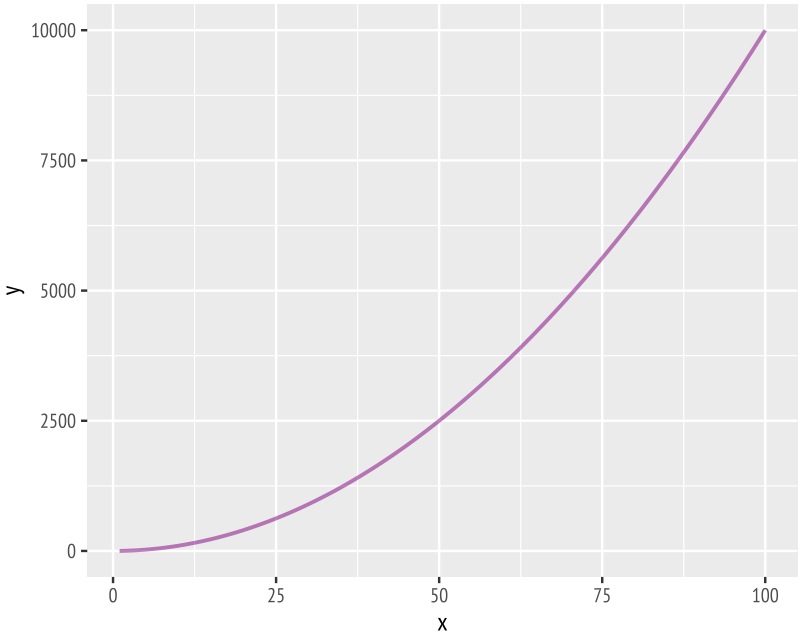

Figure 4: A nonlinear relationship.

We can still model this situation with OLS. We only have to change the *functional form* of one of our variables to allow for such interpretation: instead of using our dependent variable in levels, we use its *natural logarithm*, $\ln(y)$.[2] This allows us

[2] I will use $\ln(\cdot)$ and $\log(\cdot)$ interchangeably, since in Econometrics, when one talks about "logs," they are referring to the natural logarithm.

to interpret our slope coefficients as **constant percentage (%) changes** to the dependent variable.

Let us look at an example. First, the econometric model becomes:

$$\log(salary_i) = \beta_0 + \beta_1 education_i + u_i$$

Then we estimate it:

$$\widehat{\log(salary_i)} = .584 + 0.083\ education_i$$

Now, how do we interpret the slope coefficient on education? It's simple: just **multiply the coefficient by 100%**! In other words, on average, for every additional year of education, salaries increase by $[100 \cdot 0.083 =]$ 8.3%, all else constant. Now, for every additional year of education an individual achieves, their salary will increase by a **constant percentage**, always on top of the previous salary. This may be not fully correspond to reality, but it performs way better than the level-level version of this model.

Be **careful**, though: this does not mean that level-level models are useless. The functional form you will use depends on the research question you have at hand, as well as the interpretation that will work better for your problem/model.

Here's the *recipe* for interpreting **log-level models**:

$$\%\Delta y = (100 \cdot \beta_i)\Delta x_i$$

where Δy and Δx_i denote the changes in the dependent and in the independent variable, respectively.

Lastly, notice that when you get to interpret the coefficients in a log-level specification, you are not interpreting the dependent variable in terms of *logs*, but in terms of its *level*. The *log* role has already been passed to the mathematical calculation of the

slope's interpretation. So, you are interpreting the effect of *education* on *salary*, and not on $\log(salary)$. However, when interpreting the **goodness-of-fit** measures (R^2 and $\bar{R}^2$), then you are interpreting the variations in $\log(salary)$ explained by your model. I know, it is a bit boring, but make sure to pay attention to these details, so your interpretation is sharp.

Log-log models

Depending on our interpretation purposes, we can also log-transform the independent variables of our model. If we decide to keep y in *log*-form, then we have a **log-log** functional form. This setting is known as *constant elasticity models*. Let's look at an example:

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + u_i$$

Since both the dependent and independent variables are in *logs*, we can interpret $\hat{\beta}_1$ as the **elasticity of salary with respect to sales**. You must recall the concept of *elasticity* from your Microeconomics classes; if not, make sure to give it a quick review.

Next, we interpret the estimated model:

$$\log(salary_i) = 4.822 + .257 \log(sales_i)$$

For log-log models, we **need not** multiply the slope coefficient by 100. We only take the value as a percentage. Thus, for every 1% increase in sales, one's salary increases by .257%, on average.

Here's the recipe:

$$\%\Delta y = \%\Delta x_i$$

Now, the "elasticity" meaning must be clear enough.

Level-log models

Level-log models are *less* common in our field, but it is worth presenting them, in case it ever appears. In this case, the percentage only applies to changes in the *independent variable*, while the dependent variable remains in levels. As an example,

$$\text{Bread}_i = \beta_0 + \beta_1 \log(\text{price}_i) + u_i$$

This is a simple demand function for bread, controlling for its price. Let us look at its estimated version:

$$\widehat{\text{Bread}}_i = 7.5 - 15.2 \log(\text{price}_i) + u_i$$

Now, the interpretation is the following: for every 1% increase in the price of bread, its quantity demanded will decrease by [15.2/100 =] 0.152 units, on average. Here's the recipe for **level-log models**:

$$\Delta y = (\beta_i/100)\%\Delta x_i$$

The meaning of "linear regression"

You have probably noticed that, by using natural logarithms, our models become **nonlinear**. This is correct, but, as we will see next week, the meaning of **linear** in *linear regression* does not apply to **variables**, but to **parameters**. Put simply, as long as our β coefficients are in linear form, we can transform our variables in any desired way, thus preserving OLS properties. Notice that the functional forms we have seen so far only apply to variables, with the coefficients remaining **linear**. That is

why we can keep using OLS for these apparently nonlinear models.

Lastly, it is important to remark that a multiple regression model can (and should) contain *mixed* functional forms. For example, we can have a model like this:

$$\log(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 \log(x_{2i}) + \beta_3 x_{3i} + \beta_4 \log(x_{4i}) + u_i$$

Here, β_1 and β_3 will be interpreted in a **log-level** way, whereas β_2 and β_4 will be interpreted in a **log-log** setting. If the intercept needs to be interpreted, it will always be a **log-level** interpretation.