Statistics 101

Proudly built with R, Quarto and GitHub Copilot

Daniel Sánchez Pazmiño

Table of contents

Regression

When we have data on various variables, we can use a statistical technique called regression to understand how one variable changes when another variable changes. For example, we can use regression to understand how changes in grades are related to changes in stress levels, or how changes in wages are related to years of education. In regression terminology, it is important to keep in mind the names of the variables:

- The outcome variable is also called the dependent variable. In our example, the outcome variable is the grade in a class. We also call it the response, the explained variable, the predicted variable, or the regressand.
- The predictor variable is also called the independent variable. In our example, the predictor variable is the stress level. We also call it the explanatory variable, the regressor, or the covariate.

So, when we use regression, we try and estimate the effect of the predictor variable on the outcome variable. In our example, we try and estimate the effect of stress on grades. Stress is the predictor variable, and grades is the outcome variable.

Simple linear regression

Regression starts with only two variables, the outcome and the predictor. When we have only two variables, we call the regression a simple linear regression. The word linear refers to the fact that we assume that the outcome variable changes in a straight line as the predictor variable changes. This linearity assumption is important as it is the key behind linear regression: when we use this statistical technique, we assume that the relationship between the outcome and the predictor is linear (constant slope). In other words, we assume that the outcome variable changes by a constant amount for every one-unit change in the predictor variable. For instance, we would assume that every one-point increase in a stress level test decreases the grade by a constant amount, say 0.5 points. This is the key assumption behind linear regression. Though this might sound like a strong assumption, it is actually a very useful one, as we will see later. There are ways to relax this assumption, but we will not cover them, yet they are easy to implement with statistical software.

All linear regression models can be written in the following form:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where y is the outcome variable, x is the predictor variable. The β terms are the parameters of linear regression. β_0 is the intercept, β_1 is the slope. The intercept is the value of the outcome variable when the predictor variable equals zero. The slope is the amount by which the outcome variable changes when the predictor variable increases by one unit. β_1 is usually what we are most interested about in regression analysis. ϵ is an error term: the part of the outcome variable that is not explained by the predictor variable. We will discuss the error term in more detail later.

If we keep working with the stress and grades example, the x variable is stress, and the y variable is grades. The intercept is the grade when stress equals zero. This is not a very meaningful value, as it is not possible (or not too likely) to have a stress level of zero. Many times, the slope is there because it is mathematically needed, nothing more. The slope is the amount by which grades change when stress increases by one unit. For example, if the slope is -0.5, then every one-point increase in stress decreases the grade by 0.5 points. The error term is the part of the grade that is not explained by stress. It would carry all the other factors that affect grades, but that are not included in the model. For example, the error term would include the effect of intelligence on grades, the effect of motivation on grades, the effect of family support on grades, and so on.

Parameter estimation

The equation that was given to you above was something that is called the *linear regression model*. It is called a *model* because we are trying to model reality, meaning that we are defining a mathematical understanding of the process which in reality generates the data that we have. That thing that we want to model, *reality*, is usually called the DGP (data generating process). You must use your domain knowledge, meaning, your understanding of reality, to understand the DGP. Based on that knowledge, you will be able to define a model that is a good approximation of the DGP. This model, in this case, is the linear regression model.

Great! We have a model for reality. Now, how do we use it for reality? We probably want to know the values of the parameters of the model, so that we can use the model to make predictions. There is an initial problem for this: the error term. Because the error term is not observed, we cannot use the model to make predictions for the error, they will vary. However, that is not an issue for now. We will not estimate every single value of the outcome, we will estimate the average or mean value of the outcome, given the value of x that we feed it. So, we write this.

$$E(y) = \beta_0 + \beta_1 x$$

We will model the expected value of the outcome, given the value of the predictor. This is the average value of the outcome, given the value of the predictor. This means that the regression model will help us have a prediction of the *expected* value of x. There will be some deviation, but on average, we expect our prediction to be good. Now, how can we *estimate* these two parameters using a sample of data that we actually have? There are multiple ways to do this, but the simplest one is the *least squares estimator* or the *ordinary least squares estimator* (OLS). This estimator is the one that is used by default in most statistical software. The idea behind this estimator is to find the values of the parameters that minimize the sum of squared errors. We will not explore this estimator here too much, but know that under a given set of assumptions, the OLS estimator is the best linear unbiased estimator (BLUE). This means that it is the best estimator that we can use to estimate the parameters of the linear regression model.

Once we use the OLS estimator, we will have values for both β_0 and β_1. These will be numerical values based on the data that we have. Normally, they are written like $\hat{\beta}_0$ and $\hat{\beta}_1$. The hat symbol means that these are estimates of the true values of the parameters. However, it is also common to see them as b_0 and b_1 or even $\hat{b}_0$ and $\hat{b}_1$. The important thing to remember is that these are estimates of the true values of the parameters. It is like $\hat{x}$ and μ: a very important difference to consider.

How to calculate these estimates? Well, we need to use the data that we have, and the following formulas:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b\bar{x}$$

where $\bar{x}$ is the mean of the predictor variable, and $\bar{y}$ is the mean of the outcome variable. n is the number of observations in the sample. x_i is the value of the predictor variable for observation i, and y_i is the value of the outcome variable for observation i. Notice that for b_0, we need to first calculate b_1.

These formulas are acquired by applying the *philosophy* or the *criterion* of least squares estimation, which defines that the values of the parameters must be those that minimize the sum of squared deviations of the outcome variable from the predicted values. That is written as follows:

$$\min \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

This is called an optimisation problem, and it is solved using a little bit of differential calculus. We will not go into the details of this; it is enough to know that the equations for b_1 and b_0 give the values that always make the sum of squared deviations as small as possible.

There are then some concepts inherent to regression that emerge once we know how to calculate the estimates of the parameters. The first one is the predicted value of y, which is denoted as $\hat{y}$. We find it by applying the regression line formula:

$$\hat{y} = b_0 + b_1 x$$

Why is it called a line? Well, it follows the equation of a line, which is $y = mx + b$, which you may remember from high school. In this case, m is the slope, and b is the intercept.

Next, we need to define the residual. The residual is the difference between the actual value of y and the predicted value of y. It is denoted as e and it is calculated as follows:

$$e = y - \hat{y} = y - (b_0 + b_1 x)$$

Residuals are our estimates of the error term ϵ. OLS is the best estimator of the parameters, but it is not perfect, which is why errors still exist when we compare the predictions of y against the real y. However, OLS makes the sum of squared residuals as small as it can be. Note that residuals are **not** the errors, as the errors are something we can never observe. Residuals are our estimates of the errors, given the sample that we have.

An example

Let's use an example to illustrate all of this. We will use define a sample for 100 students, which reported their grade in a math exam (in points) and their stress level in a scale from 1-10. We will use this data to estimate the relationship between stress and grades. Below, I present some summary statistics of the sample.

Table 1: Summary statistics of the sample

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
stress	100	0	8.2	1.1	5.4	8.0	12.2
score	100	0	68.2	2.3	60.5	68.4	72.8

We now want to use regression analysis to estimate the relationship between stress and grades. Since we have the mean for both variables, we would proceed as we have done other times for the variance and covariance calculations: we arrange the data in a table, calculate the deviations, square them, multiply them and then sum them. As you can see, if you have big sample sizes, this can be a very tedious task. Luckily, we have computers to do this for us. SPSS, for instance, will run this in less than a second, along with a bunch of other statistics we will cover.

I used a computer to calculate the slope parameter b_1, and it was -1.38, rounded to two decimals. Because the formula for the intercept b_0 is easy, we can use it to calculate it by hand.

$$b_0 = \bar{y} - b_1\bar{x} = 86.5 - 14.99(5.8) = -0.07$$

The intercept is -0.07. A negative number means that the regression line crosses the y-axis below 0, and that supposedly when there is zero stress, there will be a negative score. This is not possible, so we need to be careful with the interpretation of the intercept.

You must use your common sense to interpret the intercept, and you need not panic if it doesn't make any sense. Regression is basically making a line that fits the data as best as it can, and sometimes the line will not make sense. This is specially so if we don't have a lot of data near zero for the predictor variable. Most of the time, we don't really care about the intercept, and we focus on the slope. However, there are ways to fix this problem. First, if you're worried about knowing what is the grade when stress is zero, you should try and get more data about students who reported zero stress to know this relationship. Second, you can use a different type of regression, called *constrained regression*, which forces the intercept to be zero. This is useful when you know that the intercept should be zero, and you want to know the slope. The problem about constrained regression is that it is not as good as OLS,

and it is not as flexible. We won't cover it here, but as mentioned, it is not nearly as useful as good ol' OLS.

Take a look at the graph below for the regression line.

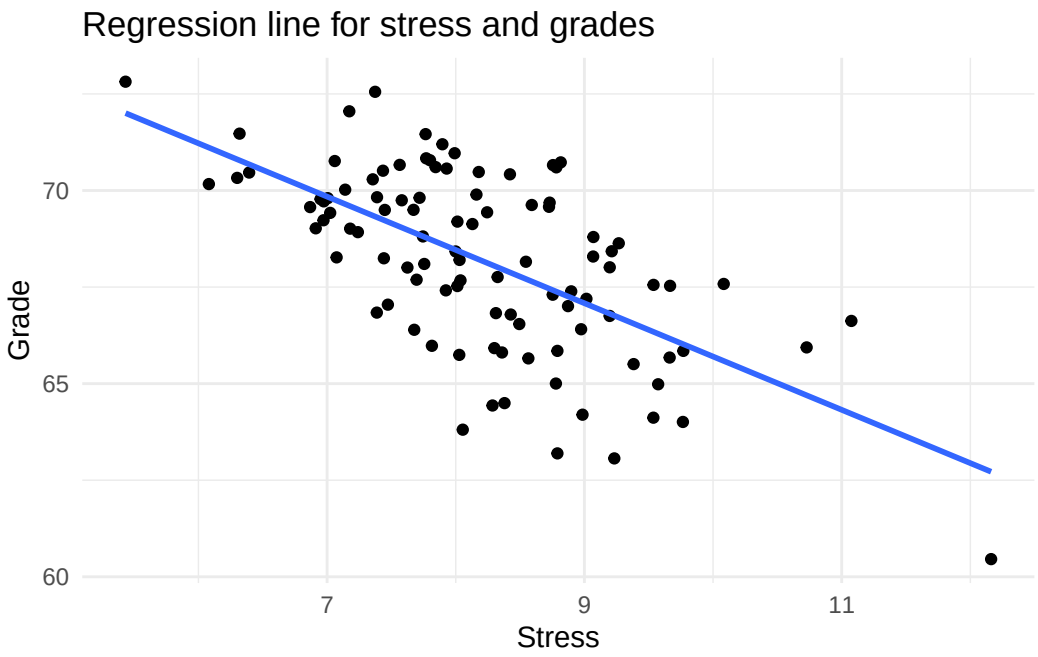

Figure 1: Regression line for stress and grades

As you can see, the points are scattered around the line. You may note that, in this particular case, the points are very close to the line. This is because I generated the data in a way that the relationship between stress and grades is very strong. In real life, this is not always the case.

When the point are very near or on top of the regression line, it means that the model is a good fit for the data. When the points are very far from the line, it means that the model is not a good fit for the data. The distance between the prediction (which lies on the line) and the real data point is the residual. So, the more distance between the points and the line, the bigger the residuals.

Goodness of fit

In the example above, we saw how the points were very close to the line. This is not always the case. Sometimes, the points are very far from the line, and the model is not a good fit for the data. How do we know if the model is a good fit for the data? We use the sum of squared residuals.

The sum of squared residuals is the sum of the squared distances between the points and the line. The smaller the sum of squared residuals, the better the fit of the model. The bigger the sum of squared residuals, the worse the fit of the model.

The coefficient of determination, called R^2, is a measure of how good the model is. It is calculated as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Where SSR is the sum of squares due to the regression, SST is the total sum of squares, and SSE is the sum of squared residuals (often also called sum of squared errors, hence the acronym). The total sum of squares is the sum of the squared distances between the points and the mean of y. The sum of squared residuals is the sum of the squared distances between the points and the prediction. You can see these formulas below:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SSR + SSE$$

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

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

The coefficient of determination is a number between 0 and 1. The closer it is to 1, the better the fit of the model. The closer it is to 0, the worse the fit of the model. We interpret it as the proportion of variance in y that is explained by x. In other words, it is the proportion of variance in y that is explained by the regression line.

The correlation coefficient r is the square root of the coefficient of determination. It carries the same sign of the coefficient b_1. It is a number between -1 and 1. The closer it is to 1, the better the fit of the model. The closer it is to -1, the better the fit of the model. We interpret it as the strength of the relationship between x and y. In other words, it is the strength of the relationship between x and y.

Multiple linear regression

Simple linear regression is good for learning the concepts, but it is not very useful in real life. In real life, an outcome variable is likely affected by not one but many variables at a time.

For instance, let's say that we want to predict the grade of a student based on their stress level and their sleep hours. We can use multiple regression to estimate the relationship between more than one predictor variable and the outcome variable. Not only multiple regression is

more useful in real life, it is necessary to avoid bias in the estimates. If we are interested in the relationship between stress and grades, we need to control for other variables that may affect grades, such as sleep hours. If we don't control for sleep hours, we may be overestimating the relationship between stress and grades. This is called omitted variable bias, and it is a very important concept in regression. However, we do not have time to cover it here.

In multiple regression, the formula is the same, but we have more than one predictor variable. The formula is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \epsilon$$

Where y is the outcome variable, x_1 and x_2 are the predictor variables, β_0 is the intercept, β_1 and β_2 are the slopes, and ϵ is the error term. Notice how we'd be able to estimate the relationship between stress and grades while controlling for sleep hours and many other things, which is why there is an ellipsis in the formula (…). For the estimated model, we include the estimated version of each parameter. For the intercept, we use b_0. For the slopes, we use b_1 and b_2, and so on. For the error, we use e. The formula is as follows.

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + e$$

We would still need to use the criterion of ordinary least squares to estimate the parameters. The formula is the same, but we have more than one predictor variable. This makes the problem a bit more complicated, but the idea is the same. We want to minimize the sum of squared residuals. We will need vector calculus to solve this problem, but we won't cover it here. It is sufficient to know that to minimize the sum of squared residuals, there is a specific way to calculate the intercept and the slopes. I also do not give you a formula for the estimates because I'd need to use matrix notation, which is beyond the scope of what you're learning right now.

What will be the regression line? We will no longer be able to graph a simple line like before. This is because we have more than one predictor variable, so our plane would be multidimensional. We can still graph the relationship between x_1 and y and the relationship between x_2 and y separately. We can also graph the relationship between x_1 and x_2. However, we cannot graph the relationship between x_1 and x_2 and y at the same time, unless we use something called the Frisch-Waugh-Lovell theorem, which is beyond the scope of what you're learning right now. The fitted values will still be the result of plugging the values of all the covariates into the regression equation. The residuals will still be the difference between the actual values of y and the fitted values.

The interpretation of the coefficients is the same. The intercept is the expected value of y when all x are equal to 0. The slope is the expected change in y for a one-unit increase in x, *holding all other variables constant.*

Consider the following example. Let's say that we want to predict the grade of a student based on their stress level and their sleep hours. We can use multiple regression to estimate the relationship between more than one predictor variable and the outcome variable. Imagine we get three estimates: $b_0 = 5$, $b_1 = -0.5$, and $b_2 = 0.2$. This means that the expected grade of a student who has 0 stress and 0 sleep hours is 5. This is the intercept. The expected change in grade for a one-unit increase in stress, *holding sleep hours constant*, is -0.5. The expected change in grade for a one-unit increase in sleep hours, *holding stress constant*, is 0.2.

The more variables we add, the more effects we *control* for. We are essentially partialling out, or separating, the effects of several things. You will notice that if you include more variables, all the coefficients will change. This is because the coefficients are conditional on the other variables. The more variables you include, the more you are controlling for, and the more the coefficients will change.

Consider a situation where you predict grades from stress levels. Imagine you get a coefficient of -3. This means that the expected change in grade for a one-unit increase in stress is -3. The more stress, the less grade. What if we were to add income? The coefficient of stress will change. It may become -2. This means that the expected change in grade for a one-unit increase in stress is -2, *holding income constant*. Why does this happen and what does it mean? Well, when this happens it is because the two predictors that you include are related to each other and to the predicted variable as well. People with more income often have better grades, and people with more income often have less stress. Before including income, the regression did not know how to difference low income from high income, so all the variation that was caused by income was attributed to the stress coefficient. When we include income, we partial out the effect of income, and the stress coefficient changes. This is a live image of omitted variable bias.

Goodness of fit

The goodness of fit of a multiple regression model is measured by the coefficient of determination R^2. The formula is the same, but we have more than one predictor variable. However, when we have more than one predictor variable, we need to be careful about how we interpret R^2. Mathematically, adding more variables never reduces the coefficient of determination, even if the variables have absolutely nothing to do with the outcome variable. This is because the coefficient of determination is the proportion of variance in y that is explained by x. The more x we add, the more variance in y we explain. However, this does not mean that the model is better. It just means that we are explaining more variance in y.

To solve this problem, we use the adjusted coefficient of determination R^2_{adj}. The formula is the same, but we have more than one predictor variable. The adjusted coefficient of determination is always lower than the coefficient of determination. The more variables we add, the lower the adjusted coefficient of determination. This is because the adjusted coefficient of determination is the coefficient of determination adjusted for the number of variables in the model. The

more variables we add, the more we adjust the coefficient of determination, and the lower it becomes.

Categorical variables as predictors

So far we've assumed we are working with only continuous or numeric data as predicted and predictors. There are ways to also predict categorical data (i.e. analyzing the determinants of a student either failing or passing a course), but their analysis is more involved.

It is more basic to understand how to include and interpret categorical predictors in a multiple regression context. For instance, imagine you want to include gender in the grades and stress example. How could you do it? You probably have data, which is a word, in your table.

We'd include it by creating dummy or binary variables. If we have a categorical variable with two categories, we include one dummy variable. It could be that the variable is 0 for men and 1 for women or the other way around. It doesn't matter for calculation, but it matters for interpretation. If we use 1 for women, the coefficient that we acquire for this dummy variable will be the extra grade that women get, compared to men. If we use 1 for men, the interpretation will be the other way around.

How about for multiple categories? In ethnicity, we might have more than two categories. In that case, we add one dummy variable for each category, minus 1. For instance, if we have three categories, we add two dummy variables. If we have four categories, we add three dummy variables. If we have five categories, we add four dummy variables. And so on. Each dummy variable will be one for the category that it represents, and zero for all other categories.

How to interpret coefficients for multiple categories? You will get numbers for all dummies you include. If we have an ethnicity variable with black, white, latino and asian, we will add three variables. We add them based on the choice of a reference group, which is the group that we compare to when interpreting the estimates. Let's say we choose white as the reference group. The coefficient for black will be the extra grade that black students get, compared to white students. The coefficient for latino will be the extra grade that latino students get, compared to white students. The coefficient for asian will be the extra grade that asian students get, compared to white students.

In the gender example, the reference category is men. It is always important to know what is the reference category, because it changes the interpretation of the coefficients. You will never be able to add all categories, because you will always need to leave one out, so that we have a reference group. In fact, if you try to include dummy variables for all categories, you will get an error. This is called the dummy variable trap.

Summary

- Regression is a method to predict a continuous outcome variable from one or more predictor variables.
- The regression line is the line that minimizes the sum of squared residuals.
- The regression coefficients are the intercept and the slope of the regression line. In multiple regression, we have more than one slope.
- The intercept is the expected value of y when all x are equal to 0.
- The slope is the expected change in y for a one-unit increase in x, *holding all other variables constant.*
- The goodness of fit of a multiple regression model is measured by the coefficient of determination R^2. The formula is the same, but we have more than one predictor variable.
- The adjusted coefficient of determination R^2_{adj} is always lower than the coefficient of determination. The more variables we add, the lower the adjusted coefficient of determination.
- We include categorical variables as predictors by creating dummy or binary variables. If we have a categorical variable with two categories, we include one dummy variable. If we have more than two categories, we add one dummy variable for each category, minus 1. Each dummy variable will be one for the category that it represents, and zero for all other categories.
- The reference category is the group that we compare to when interpreting the estimates.

Up next

- Non parametric statistics
- SPSS I and II