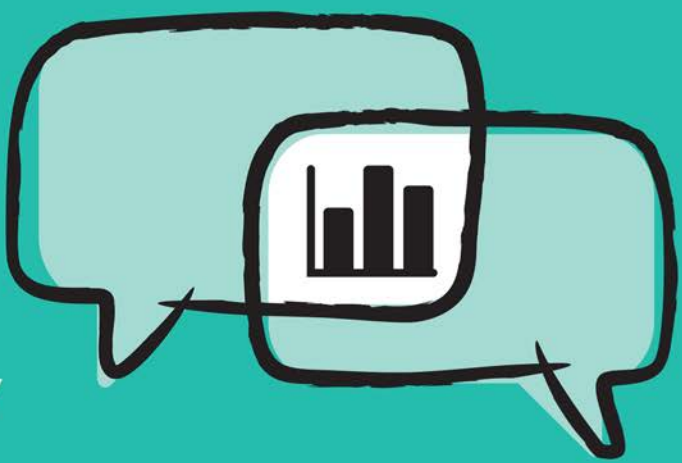
FT PUBLISHING

How to Talk about Data

Build your data fluency



Martin J. Eppler & Fabienne Bünzli

How to Talk About Data

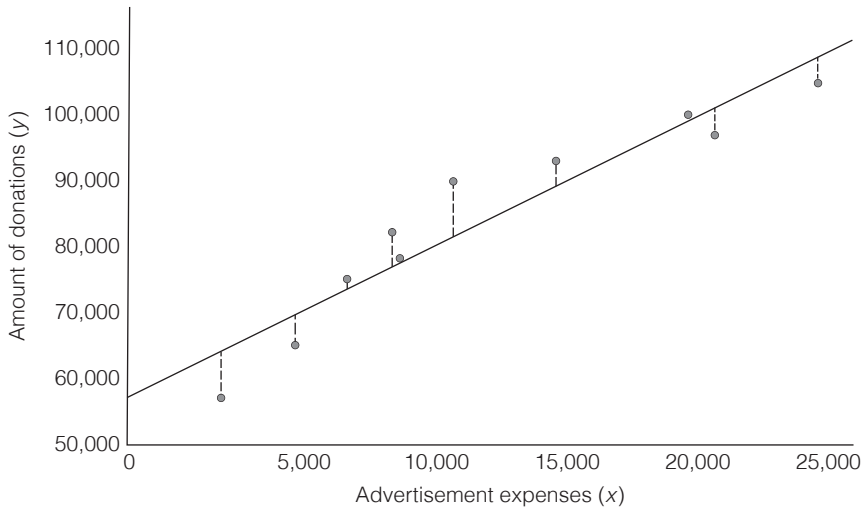


Figure 3.6 The distances between the data (the actual amount of donations) and the regression line (the estimated amount of donations) are marked with dotted lines. This is known as the error of the model or the residual. Each data point has one residual.

as possible to all data points? The way this happens is through minimising the error of the model, which is also referred to as the residual. And what's the error? It is the difference between the data that you collected (i.e., what you measured) and the prediction by your model (i.e., the regression line). Figure 3.6 illustrates what we mean by the distance between your data and the line. The bigger this difference, the worse the model is. The smaller the difference, the better the model is.

To conduct a regression analysis, we have to define a predictor variable, which in our case is the advertising expenses, and an outcome variable, which is the amount of monetary donations. The **linear regression model,** in its simplest form, is based on the equation $y = a + bx$, with a being the point where the line crosses the vertical axis (**the intercept**) and b being the **slope (or gradient)** of the line (Griffiths, 2008). Figure 3.7 visualises the components of the equation. The slope b indicates how steep the regression line is. The slope tells you how much you can expect the outcome to *change* as the predictor increases by one unit. This is visually represented in Figure 3.7 with the triangle that is attached to the regression line.

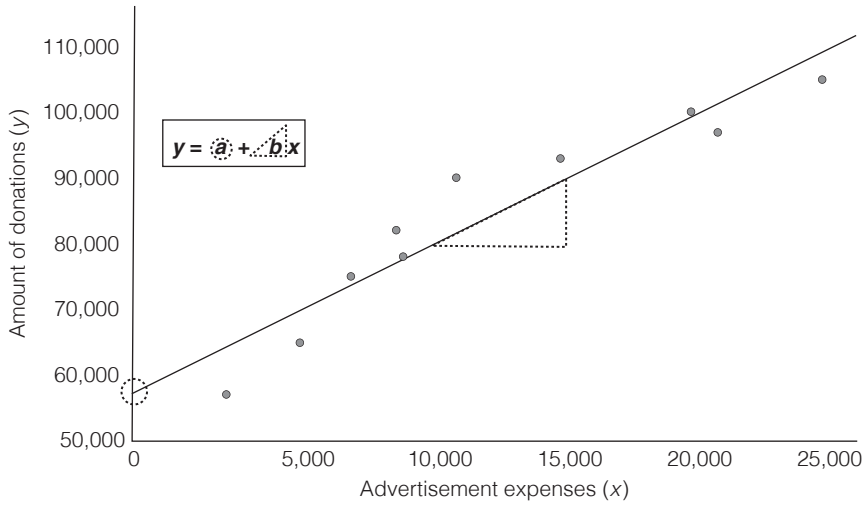


Figure 3.7 The line of best fit or regression line.

In other words, the slope is a ratio of change in the outcome (y) per one-unit-change in the predictor (x). In our example, the slope indicates how much we can expect the amount of donations to go up as the advertisement expenses increase by one dollar.

We will now elaborate a bit on the intercept and the slope to give you the necessary background to interpret these two key components of regression analysis. This will help you gain a more in-depth understanding of how predictions work. This is really key to understanding how predictive analytics work today. We acknowledge that the following pages are probably not the most enjoyable and pleasant to read. However, be ensured that we did our best to keep it as short and concise as possible. (Note that the second author's idea to include cat pictures to make these pages a bit more entertaining did not pass the editorial process.)

To find the line that best fits the data, we need to identify the values for the intercept (a) and the slope (b) that minimise the distances between the data that you have collected and the line. We will first look at the slope. The value of b that we are looking for can be calculated based on the following equation. No worries, we won't torture you with lengthy explanations about this equation. Just know the following: x is the predictor variable and y is the outcome variable. There are basically two things you have to

do in order to fill out the equation. First, you have to calculate the differences between the mean of the advertising expenses and the actual advertisement expenses you had over recent years. Second, you have to calculate the differences between the mean of all donations and the actual amount of donations made.

Equation: Regression slope b

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

where $\sum$ represents sum up, n denotes the sample size, x_i represents the ith value of x, y_i represents the ith value of y, $\bar{x}$ is the mean of x and $\bar{y}$ is the mean of y.

Anyway, let's leave this equation as it is. In our example, the slope would be 1.96. But what does this slope tell us?

To interpret the slope, we have to consider the units of the outcome variable and the predictor variable. The amount of donations and the advertisement expenses are both measured in dollars. A slope of 1.96 means that the amount of donations increases by 1.96 dollars for every 1 dollar increase in advertisement expenses.

Things become a bit more complicated when the outcome and the predictor are measured in different units. Suppose the outcome is measured as a 10-point work satisfaction scale (y) and the predictor is weeks of vacation per year (x) and you find that the slope is 1.4. What does this mean? Without considering the *units* of the outcome and the predictor variable, this number does not make much sense. Considering the units, you understand that a slope of 1.4 means an increase of 1.4 satisfaction points (change in the outcome) for every 1-week increase in vacation (change in the predictor).

And what about the intercept a? How do we know where the line crosses the y-axis? The regression equation is $y = a + bx$. The regression line represents the line of best fit and as such, the line goes through the means of the outcome and the predictor. Again here, we do not bore you with lengthy explanations about why it is like that. Just trust us. In our example, we take the amount of donations and the advertisement expenses and calculate the means for each of the variables. Moreover, we already know the value of b (the slope).

This allows us to figure out the value of the intercept (see equation below). The regression line crosses the y-axis at 59,890. We can interpret this value as follows: if no money is spent on advertising, our organisation is expected to receive donations of 59,890 dollars.

However, caution is warranted when interpreting the intercept. The point where our advertising expenses are zero (i.e., $x = 0$) lies outside of the range of data that we collected. Look again at Figure 3.7 and you'll see that we do not have data for advertising expenses below 2,200 dollars. As a general rule, you should never make predictions for a point that lies outside of the range of the data that you have actually collected. The relationship between the variables might change, but you don't know if it changes because you did not gather *that data*. For instance, it might be that the relationship between advertising expenses and donations is exponential and not linear for advertising expenses between 0 and 2,000 dollars. Hence, the intercept that we calculated based on *the linear model* would be a very inaccurate prediction. There are further instances where the intercept is meaningless: for example, when data near $x = 0$ do not exist (e.g., height: people cannot be 0 or 2 cm tall).

Equation: Intercept a.

$$\boldsymbol{a} = \bar{y} - b\bar{x}$$
$$a = 84{,}200 - 1.962*12{,}390$$
$$a = 59{,}890$$

As we have seen, the linear regression model allows you to calculate the line that *best* fits your data. Cutting-edge data visualisation tools such as Tableau fortunately relieve you from calculating the line of best fit yourself. These tools compute the line of best fit and give you various statistics including the correlation coefficient. Nevertheless, in the overwhelming majority of cases, *the best fit is not the perfect fit*: it almost never occurs that all data lie exactly on the line. There is almost always a discrepancy between the data that you collected and the values that you predict. Look at Figure 3.8. The dotted lines from the actual values to the predicted values visualise this discrepancy. If we want to consider that there is an amount of error in the predictions that we make, we need to extend the regression model with an error term (Figure 3.8).

So far, we have seen how to predict an outcome from one variable. But ask yourself: are your own decisions driven by just one factor?

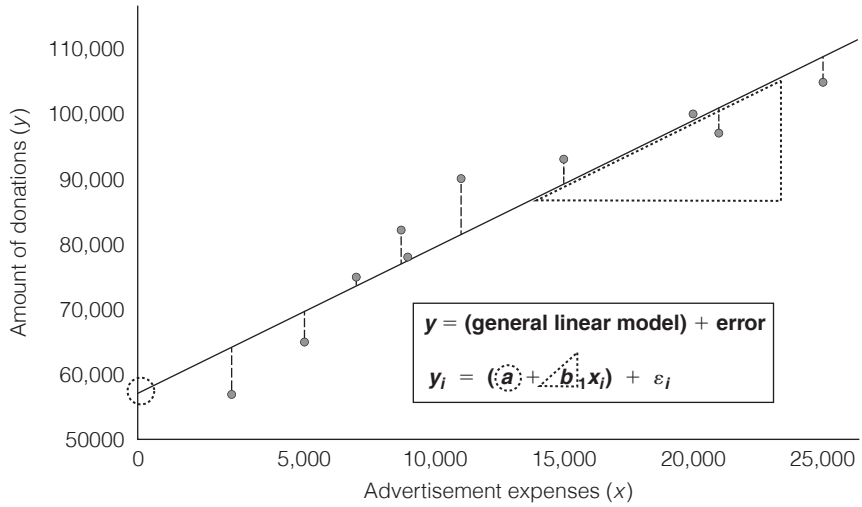


Figure 3.8 Linear regression model with error term.

Do you buy your clothes just because of the price tag? Or, do you donate money to a charity just because you are familiar with that non-profit organisation? In both cases, your answer is probably 'no'. Usually, our decisions are affected by a number of factors, with some of them exerting a stronger influence than others. Often, it therefore makes sense to consider the influence of several predictors on a given outcome. One of the compelling advantages of the linear regression model is that we can expand it and include as many predictors as we want. An additional predictor can be included as shown in the following equation:

Equation: Regression with multiple predictors

$$\mathbf{y = (linear\ regression\ model) + error}$$

$$y_i = (a + b_1 x_{1i} + b_2 x_{2i}) + \varepsilon_i$$

where y is the outcome value, a is the intercept, b_1 denotes the slope of the first predictor, b_2 is the slope of the second predictor, x_{1i} is the ith value of the first predictor variable, x_{2i} is the ith value of the second predictor variable and ε_i is the error.

As you can see, there is still an intercept a. The only difference is that we have two variables now and therefore two regression slopes (two b-values). Moreover, the visualisation of the data looks slightly different. Instead of a regression line, we now have a **regression plane**. Just as with the regression line, the regression plane seeks

to minimise the distances between the data that you have collected and the values that you predict. The aim is to minimise the vertical distances between the regression plane and each data point. The length and the width of the regression plane show the *b*-value for the predictors (Field, 2018). It is relatively easy to visualise regressions with one or two predictors (Figure 3.9). However, with three, four, five, or even more predictors visualisations are not readily made because we cannot produce visualisations beyond three dimensions.

Let's take our example with the advertising expenses and the amount of monetary donations per year again. Imagine we also want to know whether the number of newsletters sent to our members influences the amount of donations. Hence, advertising expenses and number of newsletters would be predictors. The slope for advertising expenses is 1.51, whereas the slope for the number of newsletters is 363.17. But what does the slope for the number of newsletters mean? As we have said earlier, we have to consider the units of the predictor and the outcome to interpret the *b*-value. A slope of 363.17 means that the amount of donations increases by 363.17 dollars for every newsletter we send to our members.

The *b*-values come with a huge disadvantage: if the predictors have different units, the *b*-values are not directly comparable. However,

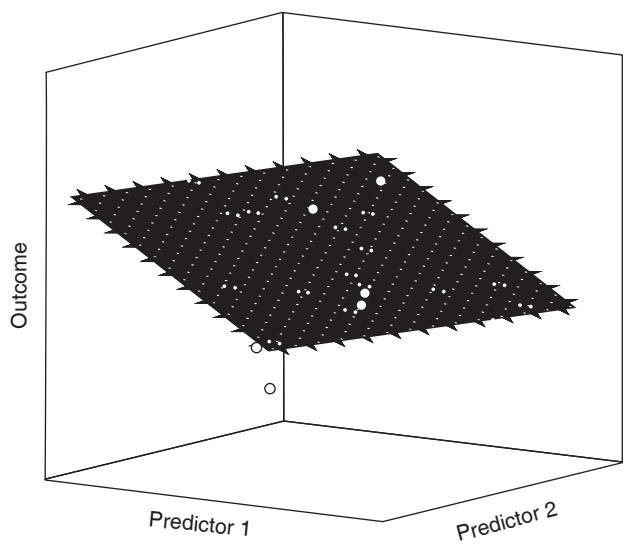


Figure 3.9 Regression with two predictors (regression plane).