The linear model and the background to regression analysis – Part 1 in "Econometrics 101"

This article series begins with the foundation, the backbone of what underlies many models. And we shall keep things simple. The level of complexity will get high enough as this educational series progresses, and there is no need to move too fast.

So to the foundation then, and the title of this article: the linear model. A topic which feels fairly intuitive and for those with a good memory, you might remember the concept of linear models from high school and algebra.

The linear models is based on the relation below

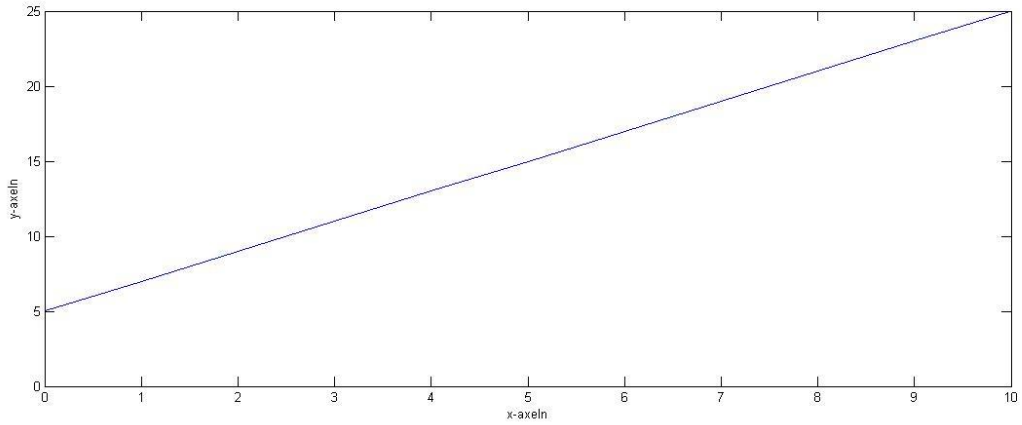
$$y = m + k * x$$

In the relation above, y is the variable we wish to explain and is called the dependant variable. In this example, we are trying to explain our phenomena with a linear model; we simply make the assumption that there is a linear relationship that explains y. In this context, x is the variable that explains y and is called the independent variable.

If we insert the line from the relationship above into a graph, the line will cut through the x and y axis at a certain point; this intersection point on the y axis is called the intercept which is represented by the m variable in the formula above. In short, this means that the line intersects the y axis at point m.

In relation to the x-axis, the line has a tilt we refer to as variable k in the formula above.

The values you give m and k will affect how x related to y. An example where we put m=5 and k=2 is found below in the figure below.



The linear model is the foundation underlying the linear regression, which most of you have encountered in the form

$$y = \alpha + \beta * x + e$$

From the simple model above, we get the quite famous concepts ”alpha” and ”beta”. One thing at a time though, how do we go from the linear model to regressions, and what is a regression?

Simply put, a regression can be said to be a practical implementation of the linear model. More correctly though, a regression minimizes the square difference between your observation and the value given by the model. Given a set of observations (data), we wish to figure out if there is a linear relationship between ex. the return y of company A and the return x of an index. The variable e represents the factors we do not know; these hidden factors can also be interpreted as an error term.

Such a relation, if kept simple and with few data points can be solved by hand. In practice, where thousands if not millions of observations and more than one independent variables are involved, we use computers to solve a regression. The reader should be aware that we can use more than one independent variables, and more than one parameter (more betas); a regression using multiple variables is called a multiple regression.

When we have calculated a linear relationship (with the help of regression analysis), how do we know that the relationship holds? We might have just accidentally found a model which seems to connect the amount of cows in Sweden to the level of rain in Zimbabwe. Many factors may lay in the way of generalizing based on results obtained using linear regression analysis. Traditionally, both academics and practitioners talk about whether a relationship is “statistically significant” or not. Such a statement is abusing the language somewhat,

as what is meant by ”statistically significant” depends on how you choose to validate your model.

With validating, we mean that we wish to make sure as best as possible that there is actually a relation and not a random coincidence based on errors that we have found. In other words, we wish to make sure that the model is more often right than wrong. With model we mean the variables you have chosen to explain y in the linear model and the assumptions you make about why there is or may be such a relationship.

In order to validate a regression model, a measure of what is called statistical significance is used. The more known measure of statistical significance is called R^2, and is generated automatically by most statistical software packages. The measure takes the variance of our explanatory variable x in the regression divided by the total variance, i.e. the part of y that is explained by x is measured. This relates back to what we discussed regarding minimizing the square difference between observed and calculated values.

Let us now problematize linear regression. When a linear relationship is assumed between y and x, the model requires that a unit change of x has an equal effect on y independent of how large or small x is. This is often a fairly improbable assumption, as you for example do not get the same satisfaction from 10 kg of bread as from a slice of bread in order to dampen a feeling of slight hunger. Other improbable assumptions are that the “error value” is supposed to be independent from x in general; we will never be able to check this assumption, but assume it none the less. This is an important problem to ponder further about; many of the assumptions that are being made in econometric contexts are assumptions that we will never be able to test (in a way that gives concrete answers); a Schrödinger’s cat of Finance. There are of course many questions to discuss regarding the assumptions that lay the foundations of how the results of regression analysis are to be interpreted in the financial industry.

We return instead to the concepts alpha and beta in the regression model. The perception regarding what these parameters mean is an interpretational matter entirely.

After having explained and gone through the linear model, the reader can see the simplicity that actually underlies these parameters, and that over-interpreting results may be a dangerous zone we often end up in. Regression analysis in a financial context which usually involves over-performance as the variable to be explained (or predicted)

$$R_r - R_f = \alpha + \beta * x + e$$

where y now describes the difference between a return and an assumed risk free rate (for those who still believes that there are states that can stand for an interest rate which is free of risk). With the classical assumption we learned at university, it is not possible to achieve over-performance, and y becomes zero. This would be the reason many of us use the term “to chase alpha”, which would entail deviating from the assumption of an efficient market and zero over-performance.

Do think about what we discussed earlier though, regarding interpretation. And what alpha and beta stands for; do not over-interpret.

By: Amira Roula

Founder, White Raven Group