Simple Linear Regression Model

The most commonly carried out statistical tool in economics is the use of the regression. A simple regression tries to estimate a linear relationship between variables x and y. There is an assumed causal relationship between which implies that the independent variable x causes the dependent variable y and not the other way around. The dependent and independent variables are also called left and right hand side variables respectively.

The researchers often estimate a *linear* relationship between variable x and y which looks like the one below.

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where α and β are constants which the researcher wishes to estimate and ϵ_i, known as the error term, gives the departure of dependent variable y_i from the line defined by $\alpha + \beta x_i$, and is treated as a random variable.

Estimation

The most common way of getting estimates of α and β, which are written as $\hat{\alpha}$ and $\hat{\beta}$ is to minimize the sum of squared values of the error terms or residuals given by $\epsilon_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. These residuals are just the deviation of the dependent variable from the estimated line $\alpha + \beta x_i$) implied by the estimates. Practically speaking, one is just drawing a line through the scatter of point in the $x-y$ space such that the square of the vertical distance between each point and the line is minimized.

It is important to remember that the "*true value*" of α and β would never ever be known unless the researcher has an infinitely large data set. With a reasonably sized data set, the estimates $\hat{\alpha}$ and $\hat{\beta}$ would be in the vicinity of actual true value but we can never be very sure how close or far they are.

What we can be sure of is that if we increase the size of the sample, it would start becoming likely that the estimates would start getting closer to the true value. The variance or the standard error of the estimates tell us how likely it is by desribing the probability with which we would find them in the vicinity of the true value. Without getting into horrifyingly graphic detail of what these probability distributions functions looks like, it is safe to say that the lower the variance or the lower the standard errors of the estimates, the more likely its get that the estimates are closer to the true value.

I am sure by now you are wondering how do we get to the standard error of an estimate? First we have to find the variance of the residuals ϵ_i^2 which is $y_i - \hat{\alpha} - \hat{\beta} x_i$, the deviations of the dependent variable from the estimated line.

The variance of the residuals is just the average of the squared residuals. (The squaring makes sure that the positive deviations do not cancel out the negative deviations.)

Similarly, one can also calculate the variance of variable x_i. This is done by finding the average of the square of the distance between each observation of x_i and its mean $\bar{x}$. Once you have it, you can get the standard error of $\hat{\alpha}$ and $\hat{\beta}$ by using the following formula.

$$\text{variance of } \hat{\beta} = \frac{1}{N} \cdot \frac{\text{variance of residuals}}{\text{variance of } x_i}$$

Testing

Not only does linear regression provide a neat way of finding statistical relationships among the data, but it also provides a framework for testing the hypothesis that each estimate is actually 0.

The thumb rule is that if the t statistic is greater than 2, you can be fairly confident that the coefficient is significant and you should keep it in the regression. If it is not, just leave it out, unless leaving it in is making a point of some sort.

To calculate the t statistic, the following long and complex formula is used.

$$t = \frac{\hat{\beta}}{\sqrt{\text{variance of } \hat{\beta}}}$$

To check is a particular variable is significant or not, make sure you read what is written at the bottom of the table. Some papers report the t statistic making it easy. Some report the estimated coefficient's standard error. Most paper helpfully put a star next to the estimated coefficient, which makes is easy for the reader.