Interpreting Regression Results

We imagine there is a relationship between two variables, x and y, according to y = f(x), which states that y depends in some way on x. An example of such a relationship is a linear equation, y = a + bx. Another example is a non-linear relationship, $y = a + bx^3$. An even more complicated relationship is given by $y = ae^x$.

Our theories involve such relationships, usually among a group of variables, y = f(x, z, w). We want to know what happens to y when x increases or when w falls, and so on. We wish to test our theories and obtain information about the specific form of the relationship using approximations of these equations with real world data on variables x, y, z, and w. Every major public policy decision depends on this kind of analysis.

Consider the following linear regression equation,

Ln(y) = a + bLn(x) + e

where a and b are coefficients to be estimated, e is a random variable, and Ln(z) is the natural log of the variable z, which is a simple transformation of the original data (x, y). Suppose that we have used some data on x and y and estimated the coefficients. Let (a*, b*) be those estimates.

The first thing to notice is the sign of b*. Is it positive, or negative? This tells us a lot about the impact of x on y. If b* < 0, then an increase in x causes a decrease in y, and vice versa if x falls. If y is demand for GM products, and x is price, then we expect b* < 0 since this captures the law of demand. In a heuristic way, we can conclude that anything that raises the price x will lower demand for GM products. Something that lowers the price raises demand. For example, if we impose a tax on GM cars and trucks, fewer of those vehicles will be purchased. On the other hand, if b* > 0, an increase in x will raise y. So, for example, an increase in income (x) will increase the demand for GM products.

The second thing to notice is the magnitude of b*. How large is the absolute value of the estimate? The larger the estimate is in absolute value the greater the impact of x on y. If b* = - 0.57, this is larger than if b* = - 0.21.

Finally, we also have to consider whether the estimated coefficient is different from zero, or not. Suppose b* = 0.321. This is a small number. Since our technique is an approximation, we are never absolutely sure we are right. The data we have is imperfect. It is collected sometimes in a haphazard way. Our linear regression is an approximation. Sometimes it is a good one and sometimes not. Ultimately, we need to know whether the estimate reflects the "truth" in some sense. What we need to do is get at this notion of what the data is telling us and this involves hypothesis testing. There are other hypothesis tests than if the coefficient is zero or not but we will focus on the hypothesis of the coefficient being zero or not for simplicity.

When the coefficient is estimated we also get a standard error listed in a separate column in the Excel spreadsheet. The standard error of each coefficient is a measure of the variability of the estimated coefficient. It is also a measure of our confidence in the estimated coefficient and helps in hypothesis testing. An estimate b* that we have a lot of confidence in has a small standard error. We have less confidence in an estimate b* that has a large standard error.

Suppose $b^* = 0.025$ and the estimated standard error is 0.0125. The number 0.025 is very small and close to zero. The question is: Is it 'statistically' equal to zero? So our so-called null hypothesis is H_0: $b^* = 0$. Is $b^* > 0$ or not. If we find evidence that b* is larger than zero, then we reject the hypothesis H_0. If we do not find such evidence, then we cannot reject the hypothesis that $b^* = 0$.

To figure this out we can calculate the so-called t-statistic and compare it to a table of critical values that measures our level of confidence in the estimate b*. To do this, we have the following formula for the estimated t-stat:

t-stat* = b*/se,

where se = standard error. For this example, t-stat* = .025/.0125 = 2.0. We then compare the estimated t-stat, t-stat* = 2.0, with a table of critical values. The appropriate critical value depends on the sample size of our data, i.e., how many observations in the data, and on the level of confidence we wish to have in the estimated coefficient, b*. Levels of 1%, 5%, and 10% are popular levels of confidence. As it turns out, the critical value of the t-stat for a ten percent confidence level is about 1.68. This is commonly used as a measure of confidence in the estimated result b*. Since t-stat* = 2 > 1.68, we can be sure at a 10% level of confidence that our estimated coefficient b* = .025 is "statistically" different from zero.

Fortunately, Excel does this for you. The first column of output in an Excel regression exercise is the coefficients, the second column is the standard errors, the third column is the t-stat, and the fourth column is the p-value (for a Mac). In fact, the so-called 'p-value' listed in the output is the confidence level of the estimate. If the p-value is small, we have a lot of confidence in the estimate and we can reject the null hypothesis H_0. If it is 7%, for example, then the relevant estimated coefficient b* is statistically significantly different from zero with a 7% level of confidence. So, a small se coincides with a large t-stat and a small p-value; a large se coincides with a low t-stat and a large p-value.

As another example, suppose b* = - 10.2 with a standard error of 20.4. Ostensibly, |10.2| is large in magnitude and we might think it is much different from zero. However, this need not be the case. It depends on the standard error. Given the se = 20.4, the t-stat is given by t-stat* = - ½ = - 0.5. Since |- .5| = 0.50 < 1.68, we can not reject the null hypothesis that the estimated coefficient is zero.

Elasticities and Estimated Coefficients.

Suppose we estimate the following equation,

Car sales = a + $b_1$income + $b_2$car price,

and all of the data is in log form. Then the estimated coefficients b_1 and b_2 are elasticities.

To see this, consider the following explicit example. Suppose

Ln(y) = a + bLn(x).

The differential of Ln(z) in general is: dLn(z) = dz/z from the rules of calculus. Differentiate the regression equation,

dLn(y) = bdLn(x),

to get

$$dy/y = bdx/x.$$

Multiply through by x/dx to get

$$(x/dx)(dy/y) = b(x/dx)(dx/x)$$

or,

$$(dy/dx)(x/y) = b.$$

The left hand side is an elasticity (of y with respect to x). The estimate of b, b*, is therefore, an elasticity.

Note that the first difference Δ is an approximation of 'd' so dx = Δx. In that case, the formula becomes

$$(\Delta y/\Delta x)(x/y) = b^*.$$

We can write this in the following way to provide a convenient interpretation,

$$(\Delta y / y) \div (\Delta x / x) = b^*,$$

where $(\Delta y / y)$ is the percent change in y and $(\Delta x / x)$ is the percent change in x.

Suppose b* = 0.25, se = .05, and x increases by 10%, i.e., $(\Delta x / x)$ = 0.10. How much does y increase? First, t-stat* = .25/.05 = 5 > 1.68 so we reject H_0: b* = 0. Second, using the formula $(\Delta y / y) \div (\Delta x / x) = b^*$, substitute what we know,

$$(\Delta y / y) \div 0.10 = 0.25,$$

and solve

$$(\Delta y / y) = (0.10)(0.25) = 0.025.$$

When the elasticity is equal to 0.25 a 10% increase in x leads to a 2.5% increase in y.

In the car sales example, suppose b_1 = 1.2 and b_2 = - 0.36 with standard errors respectively of .4 and .5. The t-stats are 1.2/.4 = 3 and - .36/.5 = - 0.72. The first estimate is statistically significantly different from zero, while the second is not. Therefore, an appropriate equation for the data on car sales is

Car sales = a + 1.2income,
(0.4)

where we have listed the se below the estimated coefficient, following standard practice. We have left out price since its coefficient is not different from zero in a statistical sense. One conclusion from this analysis is that reducing price will not sell more cars, according to this data.