Let's review how to interpret results of a simple linear regression model. Whether we use excel or any tool, we will broadly get similar output. You can see that there are two distinct tables in the output. There is a regression statistics table, there is an ANOVA table, and then there is a coefficients table. We are going to start with the coefficients table.

>>>

Remember in a linear regression model, we are trying to estimate a straight line that best captures the relationship between birth weight and the gestation period. If we look at the bottom of table, that generates the coefficients. According to our data, according to the model that has been generated on the data that we are using that best possible straight line is actually this line.

Birthweight = -3245.44 + 166 * Gestate

How is that? Remember we are trying to estimate a straight line and a straight line has an equation Y = mx + c or

$Y = \beta_0 + \beta_1 x$

β_0 = Intercept

β_1 = coefficient on the X variable

That is exactly what this coefficient output is generating. The intercept value is -3245, the gestation coefficient value is 166. If we say $Y = \beta_0 + \beta_1 x$, Y in our example is the birth weight in grams equals to the intercept -3245.44 + β_1 which is the coefficient on gestate 166.44 * gestate variable which is the X variable. So coefficients are essentially helping us to estimate the best possible straight line that captures the relationship between your Y variable and X variable.

>>

What is it means to say that the line is -3245 + 166 * Gestate. What is the interpretation of this coefficient? Let's start with the Beta coefficient on the gestate variable. If you think about it what it essentially says is that if I increase gestate by one unit then I expect my birth weight on average to go up by 166 grams. Remember if Y = mx + c, m is the slope. It is the rate of change of Y relative to a unit change in X. so if I increase gestate by one week, then I expect the birth weight to go up by 166 grams.

>>

For every unit increased in a gestation period, we expect to see an average increase in birthweight by 166 grams. The positive sign on the coefficient on gestation implies that there is a positive relationship between gestation and birth weight. What does that mean? If gestation period goes up, we expect birth weight to go up as well.

What if there is a negative sign on a coefficient. Then we expected inverse relationship. If X goes up Y will actually come down. When we say unit increase in X what does that mean? It is a unit in whatever scale X is. In our example, X is a gestation period is measured in weeks. So when I say a unit increase in gestation period, we are essentially talking about one week increase in gestation. But you could have data that is captured in hours, could be captured in days, and could be captured in other units grams etc. So a unit increase essentially means one unit increase in whatever the scale of X is.

We have to be careful when we interpret the Beta coefficients. Do we expect that every time there is a one week increase in the gestation period for any mom, the baby automatically will add 166 grams in birthweight? Intuitively that doesn't make sense. It is not necessary that every time there is a one week increase in gestation period, the baby must add a 166 grams of birthweight.so what we are talking about is an average result. On average for every one week increase in gestation period we expect to see an increase in 166 grams of birthweight. But this is an average response.

Why is it an average response? Because remember we are dealing with one sample from an underlying population. What we expect to see is if we repeat this experiment, meaning we take many samples from an underlying population and we run a regression on each of those samples. On average we expect to see that an increase in gestation period of one week results in an average increase of 166 grams of birthweight.

>>

What about the intercept? We have talked about the Beta coefficients on gestate which is 166. The intercept remember is the value of Y when X is zero. In this example we are saying that when there is zero gestation weeks, we expect the birth weight to be negative. Of course that doesn't make any sense. But remember if the gestation period is zero, then there is no birth weight either.

The intercept is simply away of baselining the outcome and you may not be able to interpret the intercept the way we interpret the Beta coefficient. However in order to correctly capture the relationship between the birthweight and the gestation period, this intercept value is required.

>>>

Another thing to be careful about. Remember we are saying that when gestation period increases by one week, birthweight increases by 166 grams. Can we say it the other way around that if the birthweight is increased by 166 grams then gestation will increase by one week? Because remember there is cause and effect. An increase in gestation period will increase birth weight but it is not true the other way around. An increase in birthweight does not mean that it will lead to an increase in gestation. Y, remember is a function of X not X being a function of Y.

Finally what about the error term? How does that influence the interpretation of results? Remember that this is a stochastic or a statistical model. We are saying that my Y variable which is a random variable is influenced by factors that are outside of anyone's control. So when we try to capture the relationship of factors that influence Y, we know that we will not be able to capture it 100%. There will always be some variation in Y that cannot be explained by the X that we have and that is because of random variation.

What we do is we assume and this is one of the assumptions of the regression model that the average error, the average variation because of randomness is zero. Remember if it is random variation it does not have a pattern. Overall if we look at average error or average variation, assume that is equal to zero and therefore we don't have to estimate the error term.

That was the first piece of information in the regression table. There is another piece of critical information in the coefficient table, which is the p value. What is the p value? A p value is a probability value negotiated with a null hypothesis. It is the probability of rejecting the null hypothesis when in fact it is true. When we have p values clearly we have a hypothesis test.

What is the hypothesis test in this regression model? The hypothesis test is in fact testing whether or not the Beta coefficient is actually zero.

>>>

Remember there is a p value here that is associated with each coefficient. Let's start with the gestate coefficient. The p value on the gestate coefficient is 2.54E-166 which is essentially a value that is very, very close to zero. 0.1650 followed by two. This is very, very close to zero and we are testing via this p value is the outcome of the hypothesis test that is testing whether or not the coefficient on the gestate variable is equal to zero.

What does that mean? The coefficient on the gestate variable is zero, essentially we are saying X has no influence on Y. So we are testing whether or not the gestate coefficient is influencing Y at the statistically significant level. What do you think the answer of the outcome of the hypothesis test is? Because the p value is very, very low, we reject the null hypothesis that the coefficient on gestate is actually equal to zero.

In other words we conclude that gestate is a statistically significant influencer of birthweight. Remember we want low p values to reject the null hypothesis. When you look at a regression output, and look at the coefficient tables, the first thing to look at is the coefficient themselves, but the next thing to look at is also the p values. How many of these coefficients are statistically significant and we check that by looking at the p value and making sure that those p values are less than 0.05.

Supposing the p value is greater than 0.05, let's say 0.3 or 0.4, what would we conclude? We would conclude that, that X variable may not be statistically significant influencer of the Y variable. Even though we see a relationship via coefficient, the relationship may not be statistically significant. In other words, it could be driven much more by random variation rather than a real relationship between the X and the Y variable.

>>

How do we actually test the hypothesis? We can show that the distribution

$$(\hat{\beta}_j - \beta_j) / se(\hat{\beta}_j) \sim t_{n-k-1}$$

Let's step back for a minute and think about where does the distribution come in? Remember this is a statistical model. We are taking one sample from an underlying population and we are looking at the data in the sample to come up with the relationship between X and Y.

However there are many possible samples that we can take from that underlying population and essentially we are saying that for each of the samples, if we run a regression model, we will get different estimates of the Beta coefficients. It is not necessary that in every sample, that we will get a same value of the Beta coefficient because Y is a random variable.

If we look at the distribution of Beta coefficients then we are saying that the distribution of the Beta coefficient across the samples which is

$$(\hat{\beta}_j - \beta_j) / se(\hat{\beta}_j) \sim t_{n-k-1}$$

The difference between the estimated coefficient and actual coefficient in the population divided by the standard error of actual coefficient in the population (remember standard error is standard deviation by square root of sample size). That is distributed with a t distribution with n-k-1 degrees of freedom. n is the total sample size, k is the number of variables or parameters. So you understand that the p value is being generated of a t-distribution.

>>

Finally what we are interested in. Remember in any hypothesis test, is whether or not we are rejecting the null hypothesis. In this particular example, we are certainly rejecting the null hypothesis that gestate has no influence on the birthweight of the baby.

>>

There are some other output in the coefficient table as well. There is standard error, there is t statistic, and there are confidence intervals. Remember we are looking at a distribution of coefficients. This is one possible coefficient from a distribution of coefficients that can come from multiple samples. The coefficient is a point estimate, the average of the coefficients.

The standard error is the standard deviation of the distribution of coefficients. The test statistic is the distance in the t distribution and the p value is the associated probability of outcomes greater than the critical value which is the t-stat. because this is a distribution, sometimes we are more interested in 95% confidence level. What is that mean? When x increases by one unit, it does not mean that Y will always increase by exactly β_1. Sometimes Y is going to increase little bit more than β_1, sometimes it may increase less than β_1. Because there is a distribution of Beta coefficients.

We know than every time X increases, Y does not increase exactly by β_1. But what we can say is we can talk about the confidence level. We can say 95% of the time when X increases, Y will increase by these values and we get that by the standard error. Essentially if you look at the lower 95% and upper 95% confidence intervals in this example, we are saying that every time X increases by one unit 95% of times Y will increase 156.5 grams and 176.37 grams. So we know 95% of the time this is the values that show how much Y will increase by.

Let's think about these confidence levels. They are coming directly from standard error. Remember you have point estimate and then you have a standard deviation. The larger the standard error, the greater the range of the confidence interval. The lower the standard error, the narrower the ranger of the confidence interval. The lower 95 to upper 95 is called a confidence interval.

Intuitively would we want a large range of the confidence interval or the low range in the confidence interval? We would want to low range.

Because when we come up with an estimate, we want to be pretty sure about that estimate. We don't want it to be the very, very wide interval. Because if we say that for example, the lower 95% had been 100 and the upper 95% had been 200 that's a very wide range. But the narrower it is, the more precise we are about the estimate of the Beta coefficient.

Remember the confidence levels are sometimes more appropriate to report than a coefficient, a point estimate and the narrower the confidence interval the better your estimate is. A confidence interval width depends on the standard error of the X variable that is reported in the output. So that was about one table which simply talks about the coefficients, p values, and the confidence level. But there were other tables as well. There was the ANOVA table and the regression output r^2 table. Let's understand what they tell us next.

>>