We were looking at how to interpret output of a liner regression model and we started with the first table, the lower most table in the output which has the coefficients, the standard error, the p values and the confidence intervals. The coefficient values tell us what is the impact on Y for a unit change in X. In this example, if you look at the Beta coefficient on the gestate variable which is an X variable. When X changes by one unit, In this case one week Y changes by 166.4 units.

There is a p value associated with this coefficient which is a number which is very close to zero, 2.54E – 166. Essentially 1650 followed by 2. How do we actually generate a p value? We are working with one random sample from an underlying population. If we took multiple random samples from that underlying population and we created regression models from each sample then we will get a distribution of Beta coefficients

This particular coefficients that we have generated is one coefficient from the possible distribution of Beta coefficient on gestate and we are checking is this coefficient different from 0 or not. How likely is it that we will get a coefficient of 166.44 if really the underlying impact was zero. We are rejecting the null hypothesis and concluding that this is statistically significant influencer of the Y variable.

The confidence intervals are based on the coefficient estimate and the standard error. Essentially the standard error is the standard deviation of the distribution of Beta coefficients. The smaller the standard error, the narrower the confidence interval and the more confidence we are about the preciseness of this estimate. This is only one table. But there are other outputs that is generated when we run a regression model.

>>

We are trying to find the best possible straight line that explains of captures the relationship between X and Y. But how do we know how good this straight line is? One way to do that is to look at a measure of explainability or R^2. That is what percentage of the variation in Y is being explained by your X variable. There is mathematical way to calculate the R^2

$$R^2 \equiv 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}}, \quad \text{Where,} \quad SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2; \quad SS_{\text{err}} = \sum_i (y_i - f_i)^2;$$

>>

But essentially what R^2 does and you can see R2 listed in the regression statistics output. It explains what is the proportion of the variance in Y that is being explained by X. Because we are running a simple linear regression model here, remember we have only one X variable. 49% of the variation in your Y variable which is birthweight is currently being explained by your X variable which is gestation.

>>

- The higher the R^2 obviously the more variation in Y you are able to express and therefore the better your model is
- However R^2 is one measure of how good your model
- It is a good measure but it is only one measure of model fit
- Sometimes you have multiple models with the same variable with the same R^2 but different model fit. We will understand that what model fit is shortly

>>

There is also a measure called adjusted R^2. Adjusted R^2 is actually better measure of R^2. Adjusted R^2 essentially captures variation and the adjusted R^2 typically will go up only if significance variables are added to the model. One of the things that happens with R^2 is as you add more and more X variables to the model your R^2 is going to go up. Irrespective of whether or not that new variable is the significant variable. But adjusted R^2 usually will go up only if significant variables are added to the model. So if we are interested in using the regression equation for prediction then we should look at adjusted R^2 rather than R^2.

>>

Finally there is an ANOVA table. ANOVA and regression are actually similar kind of models. They are called generalized linear models. But in an ANOVA table essentially what we are checking for is whether or not at least one of the Beta coefficient is different from zero. In this case we have only one X variable. So one Beta coefficient but if you have multiple X's ANOVA table will tell you are at least one of the Beta coefficients different from zero.

In this example p value < 0.05, so we reject the null hypothesis and we conclude that at least one of the Beta coefficients in the model is different from zero. We have looked at how to interpret basic output of a simple linear regression model. But in real life we rarely work with simple regression. We work with multiple linear regression. Because very often there are many factors that influence a Y variable. So next we will take a look at how to implement and interpret the output of a multiple linear regression model.

>>