We have looked at how to run a multiple regression model. In most business situations, there are many factors that simultaneously impact the dependant variable and in terms of ordinary Li square estimation, we will follow the same process and we have multiple independent variables. We want to identify the line that minimizes the sum of squared residuals across multiple dimensions.

>>>

Now if we run, in our example, a multiple linear regression model with birth weight as our target variable or dependant variable and the independent variables being the level of the Mother's education, the Race and the smoking status and the gestation period. Essentially, we are saying that this is the straight line equation that captures the relationship between the target variable and the independent variables.

$$Birthweight = \beta_0 + \beta_1{*}Gestation + \beta_2{*}\ Years\ Of\ Education + \beta_3{*}\ Race + \beta_4{*}Smoking$$

>>>

Therefore, the number of ß coefficients that we are now estimating using the ordinary Li Square method is 4 ß coefficients.

>>

Now when we look at the output of a multiple linear regression model, similar to the output of the simple linear regression model, you will see the following output. You will see some regression statistics and Anova table and coefficient table. Of course, depending on the tool we use, some of this information may be presented differently, but broadly this is the sort of standard output that you will get when you run a regression model.

Let us start with the bottom most table, which is the coefficient table. Remember because we have 4 independent variables, we are estimating 4 β coefficients. There is also the intercept estimate, which is the 5th β coefficient.

What are the key things to look for in the coefficient table? One is look at the coefficients and their signs; the second is to look at the P values.

Now what does it mean to look for coefficient sign? For example, if you look at the Years of education variable (YearsEduc), the sign on the coefficient is positive. Now what does that imply? Remember the β coefficient is interpreted as a change is Y for a unique change in X. So for this particular variable, essentially, we are saying that if the Years of education go up by 1 year, then the birth weight of the baby, which is the Y variable is going to go up by 9.57 grams on average. Is that intuitively correct? Intuitively that is something that we could explain. The higher the years of education of the mother, possibly the better financial condition the mom is and is able to afford good food, good nutrition. Therefore, we would expect that the child's weight is also quite healthy. So intuitively, this seems correct.

Similarly, if you look at the race variable, there is a negative coefficient on the race variable. Race essentially is created as 0/1 variable which is a dummy variable. 0 is if the race is non-African-American and 1 is if the race is African-American. Now this coefficient of -168.96 essentially says if you move from 0 to 1, the average birth weight of the baby is going to go down by 168.96 grams. Again intuitively, this may make sense. Because socio-economically, African-Americans on average have lesser income than Non-African-Americans. So intuitively, we can say that we understand that relationship.

But let us also think about that relationship and intuition for a minute. Should we always have an intuition about an expected relationship? Many times, we may. For example, most often we know when price goes up, the sales will go down. When marketing go up, we expect the sales to go up. So we have some idea of a relationship. But sometimes, the reason we are running the analysis is to test whether or not the relationship actually exists. Or sometimes we run the model or analysis to figure out what is the relationship; I may not have an idea about what that relationship is.

So we should think about expectation carefully. If you are essentially going in with a position that says I really don't know the relationship and I am going to let the data tell me what the relationship is, you may not have an expectation on the coefficient. But in most business situations, we have some sort of an expectations based on our business knowledge about the relationship between a X variable and a Y variable.

Now if that relationship is borne out by the data, it is great. What if the relationship is not borne out? For example what if instead of saying a negative coefficient on a price variable when you are modelling sales as function of price, you see a positive coefficient. When price goes up, the sales are actually going up.

Does it mean that the data is wrong? Maybe. But it could also mean that your product behaves differently. Sometimes we know some luxury goods, for example, that when price goes up, the sales may go up. So you want to carefully think about the expectation and evaluate the model-making sure what you are seeing is a real impact.

So remember, you could sometimes see non-intuitive results because either your data is bad or your hypothesis or expectation was wrong. You should be able to think through that carefully. In this particular example, do we see any non-intuitive coefficients?

- Years of education is positive – make sense.
- Race is negative - OK
- Smoking is also OK; if the mother smokes during pregnancy, then the birth weight on average will be reduced. That is intuitively understandable.
- And the gestation period is positive and that is intuitively understandable. The

longer the gestation period, the higher the average weight.

Now also remember in the coefficient table, we should look at the P value. Because P value tells us which of these coefficients are actually statically significant influencers of the Y variable. In this example, everything is statistically significant except for the years of education at a 5% level of α.

While we have a positive coefficient on the years of education, the model is suggesting that there is a 13% chance that this is because of random variation. What should we do if the independent variable is insignificant? It depends on what you want to do with the model. If our aim with the model is to understand the impact, then we just note that the years of education is insignificant, not a significant influencer of birth weight of a baby.

If our intention instead is to do some prediction, we may not want to include relationships that are not statistically significant that are potentially driven by randomness. In that sort of situation, you may finalize a model where we will drop the years of education variable because it is not significant.

>>>

We should also look at the R square table. Remember the R Square tells us what is the percentage of variance in Y that is being explained by these X variables. If you remember simple linear regression model, the R Square was 49%. The R Square has now gone up, but only a little bit. It is now 52%. So there is still a lot of unexplained variation in Y over and above the 4 independent variables we have used.

>>>

There is also the Anova table. Anova table tests the hypothesis that none of these β coefficients are different from 0. All the β coefficients are equal to 0 and we will reject the null hypothesis, because my P value is very close to 0. Therefore, we will conclude at least one of these β coefficients is significantly different from 0. This is how we interpret the output of a multiple regression model.

>>>

But we must also look at model fit. How good is the model? Why should we believe the model? One measure of how good the model is, is R^2,which we have looked at. But that is not the only measure and we should not rely on R^2 alone as a assessment of the quality of the model. There are other model fit/validation techniques that we will take a look at in the next section.

>>>