We did a Multiple Linear Regression model using the data that was available to us. It had a target variable which was the birthweight of a baby, and it had data on some X variables on influences of birthweight. In our sample, the data we had was the gestation period in weeks, the race of the mother, the education level of the mother and whether or not she smoked during pregnancy. Now we got some output and we looked at coefficients, we looked at P values, but how do we finally decide that we have a good model or how do we validate a model?

There are many ways of validating a Linear Regression model and all of them should be evaluated. The first is to look at the R^2. The R^2 explains the amount of variation in Y because of the X variables. So the higher the R^2, the better the model, the more variation you explain, but R^2 is not the only measure of Model Fit.

We also look at what is called a Fit Chart and a MAPE measure. Let start with what a FIT chart is: Remember, according to our model, birthweight equals to: -2834 + 156.51*Gestation + 9.57* Years of Education - 168.9* Race - 174.8* Smoking. I got these numbers from the coefficient table.

Now, if this was a good model, if this was the straight line that best explains the relationship between the X variables and the Y variables, we can actually come up with fitted values of the birthweight. What is fitted value? Remember in our sample data, we had data on gestation, years of education, race and smoking. In fact we have a

thousand one hundred and fifteen observations for each of these variables. What happens if you substitute the values of these variables in this equation?

You will get a birthweight outcome and that birthweight is a fitted outcome. Fitted meaning, we fit the X values to this equation and we come up with a Y value.

>>

Now, we can do that manually by simply taking this equation and multiplying the X values with the coefficients and coming up with a birthweight. Or, we can automatically do that using a tool, for example excel.

>>

Remember, this is our data. When I run data analysis, and now I'm going to regression and now my dependent variable is till the Y variable range A, but my X variable range is now multiple variables. So, it's going to be gestate, years of education, race and smoking. So, I'm going to choose four independent variables. I'm requesting for confidence intervals of 95% output in a new workbook, but now I'm also going to ask for residuals. We'll talk about what residuals are in a minute.

>>

Now this is the regression output and these are my coefficients. If I want fitted values, essentially what do I need to do? These are my coefficients, I can put a fitted value like this: We can say Y, birthweight equals to intercept, minus 2834, plus 156.51*gestate (bw=-2834+156.51*gestate) and so on. In fact, let me do this calculation manually with the first set of X values.

These are my X values. So if I say Fitted Y here, I can actually calculate this as intercept, plus first beta coefficient times[*] the first X value, which is gestate, plus the second coefficient years of education times[*] years of education, plus the third beta coefficient which is race times[*] the value of race, plus the fourth beta coefficient, smoking times[*] the value of the smoking in the first observation. So, this is the fitted value of Y. Remember, if we use this equation and the X values available to you, we will come up with a Y value. So, this is the Y Fitted value. However, instead of doing this manually, we can automatically generate that in excel or any tool by asking for predicted values or residuals.

In fact, if you look at the output, you can see there are predicted values which are the same thing as fitted values and residuals. This predicted value 3251, is exactly what we calculated for the first row, 3251. So we don't have to manually calculate this, we can automatically come up with this. So remember the fitted value is the value of Y if this line is the best possible line that captures the relationship between X and Y.

>>>

How many fitted values will we get? We will get as many as the number of observations in our data set. The residuals are the difference between the actual Y and the predicted or the fitted Y, so this is really like an error.

>>>

Now, why are we doing this? Because, we can use the fitted values to validate the model. Remember, we have an actual value of Y and now we have a fitted or a predicted value of Y. If you have a good model, what would you expect to see between the actual value and the predicted value? They should be very close. The closer the actual and the predicted values are, the better your model is. So in fact, one way to validate the model is to generate the predicted or the fitted values for all values of X and compare the actual Y to the predicted Y.

>>>

Remember, we have the predicted Y here. I'm going to copy paste this and add another sheet, and next to that I'm going to use the actual values. Remember, this is my actual values. So now if I compare these two, I'm going to just use a chart, I'm just going to use a line chart.

This is a comparison of the actual values and the fitted values according to the model – "What is Y and what was actually Y." Remember, if we have a

good model we would expect to see the actual values and the predicted values to be very close to each other, very similar. If we see that they are not then we don't have such a great model. What about this particular output? You can see that in some instances they are close, but this is not really a great model because many times the model is off. If you look at the top half you can see many times the model is unable to predict when weights are greater than average, and when weights are lower than average, the model does a better job but still not so good. Ideally if this was a good model, you would expect to see a lot of overlap between the actual values and the predicted values. This is called a FIT chart.

>>

And, this is one way of validating a model. Remember R^2 gives us one measure; the FIT chart will give us another sense of whether or not we have a good model.

>>

Also, from the FIT chart we can calculate what is called "Mean Absolute Percentage Error." The difference between the actual value and the fitted value is the error. And if I look at the average of the absolute error, so I calculate the difference between actual and predicted, and I make an absolute value so I ignore the sign and then I take an average of that, I will get what is called as The Mean Absolute Percentage Error.

>>>

So, the calculation is pretty straight forward. I want to take the percentage of error between actual and predicted, so actual minus predicted, divided by actual, this is the percentage error. Now this percentage error could be negative or positive depending on whether the fitted value is higher than the actual value or lower than the actual value. If I take the absolute value of this then I'm ignoring the sign on the error, so this is my absolute error. The average of this is my Mean Absolute Percentage Error. So in fact, this is my MAPE, this is simply my error. So, my Mean Average Percentage Error is 11%, in other words the model is off on average by 11%. Ideally, you would want to see MAPE values that are low, you may want to see absolute percentage errors 5% or lower.

>>>

So, that is the third measure of model fit or model validation.

>>>

There are many ways of validating your regression model. We want to see high R^2, but R^2 is not the only measure. We want to make sure that we have a good FIT in the model and we want to make sure that our Mean Absolute Percentage Error is low. Many times people ask what is a high R^2? What is a good enough FIT? And we will talk about all of that when we look at how to improve our models. But

one final thing about the model in general, we talked about the regression model but how is this a predictive model in technique?

- Remember once we have a final validated model, given the regression equation, for values of X we can predict a Y variable.
- We calculated fitted values exactly like that. We said, if given these X variables what is my prediction of Y? What is my fitted value of Y? But if we are convinced that this is a good model then we can come up with predictions of Y based on the X variables.

>>

If we were told that we have another mother with ten years of education, race is African American, expected gestation period is forty weeks, she's not given birth yet, and she doesn't smoke, can we come up with a reasonable estimate of what the baby's birthweight should be? We can, we simply plug in those X values and we come up with a value of 3352 gms. This is a prediction, remember the baby's not born yet, but we can predict that the baby's birthweight will be approximately 3352 gms. How reliable is this prediction? Clearly, the better our R^2, the better our model FIT, the lower the percentage error, the more confident we are about our prediction. And therefore, we want to make sure that we have good validated models when we try to predict given some data. So, this is a predictable model because once we validate the model, once we convince that the straight line equation we have is a good line that captures the

real relationship between the X variables and the Y variables, we can use this equation to plug in expected values of X, to come up with a prediction of Y.

>>

And remember, we will believe those predicted values if we have a good model. In our particular example, we don't have a great model. We have R^2 of only 52%, MAPE is 12% which is high, the FIT also wasn't very good when we looked at the FIT visually.

So, what next? We need to improve our model. How do we do that? There are many ways of improving models and that is why lots of times analytics is called more an art than a science. But before we look at how to improve models, we will first review assumptions that we need to be aware of when we build a Linear Regression model using ordinary least square approach.

>>