We have seen how to implement ordinary least square regression models, but it is also very important to understand the assumptions under which the ordinary least square beta coefficient estimates are valid. There are a series of assumptions; some of them are very necessary, some of them can be relaxed, but let's go through the assumptions first and then we will look at some of these assumptions in greater detail.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

First, the model is linear in parameters. What this means is: We have a model of the form $Y=\beta0+ \beta1X1+ \beta2X2$, and so on. We cannot have a model where $Y= \beta0+ \beta1^2X1+ \beta2^2 X2$, that is not a linear model. Notice that we are talking about the parameters, the beta coefficient, not the excess. If you have a model where $Y= \beta0+ \beta1X^2$, that can still be captured with the linear aggression model because we can convert that into a linear form by taking for example the log. If I had $\beta1X^2$ and I took log on both sides, my model will become $\log Y= \beta0+2\log \beta1X1$. So, essentially we are saying that in a linear aggression model the parameters which are the beta coefficients, that they are linear in nature.

Second, the data that we are using for a linear regression model is essentially a random sample from an underlying population and the errors are not correlated; they are statistically independent from one another.

The third assumption, the expected value of the errors is always zero. Remember, expected value is essentially the average; the average of the errors is zero.

Four, the independent variables are not too strongly collinear. Here, collinearity is essentially co-relation. We are saying that independent variables are not highly co-related among themselves. Remember, we are not talking about the co-relation between the independent variable and the dependent variable; we are talking about the independent variable themselves not being highly co-related with one another. The independent variables are measured precisely. The X variables are measured precisely with no error. The residuals have constant variance. The residual remember is the error, the difference between the predicted value and the actual value. We are saying that the variance of the error should be constant.

>>

The errors should be normally distributed with a mean equal to zero and finally the model is correctly specified. Meaning, that we are using the right variables in the right type and we are not missing good information. If all these assumptions are met, essentially we have what is called a BLUE model. OLS beta coefficients are Best Linear Unbiased Estimators of the true relationship of the X values and the Y value relationship. When we say best we mean minimum variance estimator. But let’s look at what these assumptions mean, especially some of them that are very easy and important to check.

>>

In real life remember though that all conditions may not be met.

- First, we need to check if the assumptions are holding up.
- And if they are not, then we may need to figure out how to correct for violations or decide if we need to use some other technique other than a linear regression model.

Therefore, when you build a linear regression model, you should also be checking are the assumptions under which the OLS beta coefficients are Best Linear Unbiased Estimators, are those assumptions valid, are they holding up or not.

>>

Let's start with the first assumption, linear relationship. In order to check that, what we can do is we can run a model and plot the residuals against each independent variable. The reason we are plotting the residuals and not the actual Y variable is because remember in a linear regression model we are looking at many X variables that influence the Y. If we look only at the plot of XY, then you are not including the impact of the other X's on the Y. The reason we are including residuals, remember residuals are nothing but predicted minus actual, we are essentially taking into account all the impact of multiple X's on your Y and now we are looking at the relationship between the residuals and each of the X's. Now, in this plot you should not see a relationship. You should see random variation. If you see random variation for each of these plots, residual against each X, then your model is correctly specified linearly.

>>

However, if the data is non-linear, if the relationship is non-linear then you will see a residual plot that looks like this, and so you know that you need to do something because you're not capturing a linear relationship. And what can be done, we will discuss.

>>

Another assumption that we should check for is that the residual should have constant variance, this is called homoscedasticity. If you look at the left plot you can see we have plotted the residuals and the variance of the residuals is constant, there is no pattern.

However, if you look at the plot on the right side, you can see there is a pattern to the residual. As the values of the X variables go up, the variance is reducing in the right plot, and this is called heteroscedasticity. Of course, this is one pattern and there could be other patterns as well but ideally what we want to see is no pattern. Why is that? Remember, we are essentially seeing that once I build my model, whatever I can't explain is random chance variation. If it is really random chance variation there should not be a pattern in the error top. If there a pattern in the error top then it is not a random variation. There is some other factor that is influencing this model that we are not capturing in our model and therefore, we should not see heteroscedasticity. If you have a good model, you will see constant variance which is homoscedasticity.

>>

What does it mean if you have heteroscedasticity? You run a model, you'll look at the coefficients, they look fine to you but you look at the residual plot and you can see heteroscedasticity. Now ideally, it's not a good idea for us to have heteroscedasticity but it does not imply that the coefficients themselves are wrong. However, when you have heteroscedasticity that is bias in the standard errors, remember the standard error is the variation or the variance of the distribution of the coefficient estimate. The standard errors are biased, and therefore the confidence intervals and the P value that we are getting will be biased. So, hypothesis test results will be biased leading to wrong inferences if you have heteroscedasticity.

>>

Another assumption that we should check for, are the residuals normally distributed? This is easy to do, I did it for our gestation model. Take the residuals and simply generate a histogram. You should ideally see a probability plot or a histogram which is normally distributed for the errors.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

What if the residuals are not normally distributed? Hypothesis test outcomes may be invalid though this is really not so much of an issue when you have large datasets. When you have large datasets then you don't really need to worry about residuals being normally distributed, approximately normal is still okay.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Another assumption that we should check for, the independent variables are not too correlated. When the independent variables are highly correlated we have an issue called multicollinearity. What does it mean to talk about independent variable correlation?

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

In excel, suppose I was looking at the correlation between the independent variables. Remember I have four independent variables in this model, years of education, race, smoking and gestation period. I'm looking at the correlation for each of these combinations of X variables and I should not see high correlations. Higher than 40%, 50% or 60% correlations essentially means that the independent variables are highly correlated, in which case we have a problem called multicollinearity.

>>

What is multicollinearity and why is it a problem? If two independent variables are highly correlated, essentially what we are saying is that there is very little additional information coming from the second variable, because the relationship between X - Y is already being captured by let's say the first variable. So, if X1 and X2 are highly correlated then the relationship of X1 with Y, the beta coefficient is already capturing a lot of the information about the relationship of X2 with Y, because X1 and X2 are highly correlated.

Again, the estimate themselves maybe fine but the standard errors are inflated, and one way to check that is to generate what is called a Variance Inflation Factor. Now Variance Inflation Factor is generated easily in specialised tools like SAS or R, not in excel. If you are using excel then you can look at the correlations themselves and see if there are variables that are highly correlated. But remember, the problem of multicollinearity is not that the estimates are wrong but that the standard errors are inflated. Again it means your confidence level intervals are inflated and your hypothesis tests outcomes are not necessarily valid.

Typically when you have multicollinearity, what is recommended is that you combine the variables, because even though the beta coefficients are fine we cannot be sure about how precise those estimates are because X1 and X2 themselves are highly correlated. So, typically what is done in a modelling process is to reduce the correlation by either transforming the variables or combining the variables.

>>

Remember that running a model given data is an easy task given that the computation is done with tools like SAS or Excel. The skill of an analyst lies in generating the right model to understand and solve for the business issue at hand. What we've looked at is a very simple, naïve model. You put all the independent variables, you use the target variable and simply generate a regression, but that is only the starting point. Depending on the domain understanding and modelling techniques knowledge, many models are run before arriving at a final model. And we will take a look at what it means to run many models when we look at our case study for the linear regression model. How can multiple models be run using the same data?

>>

There are techniques called Step-wise Regression where essentially you add one variable at a time or you remove one variable at a time. And the idea is you look at the model, what happens to it when you start adding one variable at a time, because remember, the X variables are also interacting with one another. So, it gives us an idea of what is actually happening, which variable influences other variables, what is a stable model. So you can do Forward Regression, Forward Step-wise Regression where we add one variable at a time or

Backward Step-wise Regression where we remove one variable at a time.

>>

But these are still essentially tricks to sort of understand correlations, understand how the X variables behave in relative to the Y. In real predicted modelling, especially at the business level, the Step-wise Regression models gives us an idea of what a stable model is. But to build a good business friendly model, there is still a requirement of domain understanding and analyst skill. And the analyst will use the Step-wise Regression model to gain an understanding of behaviour, but eventually the final model typically will be built manually.

>>

There are other things that we can do to manually build models. Remember, sometimes once you run an initial model, you may realise that the variable is not being captured correctly or not being created correctly.

- Sometimes we will do variable transformations, log transformation. S
- Sometimes we may do interaction variables, trying to capture impact of two variables together.
- Sometimes we will aggregate variables, sometimes we will disaggregate variables.
- Sometimes we will do other data preparation in order to capture the exact relationship.

So what we've looked at so far is a very, very simple approach of linear regression. We've looked at what is a linear regression model, when is it used, how are ordinary least square estimates generated, what is the interpretation of the least square estimate and what are the assumptions of a regression model.

Next we will look at case study where we will understand how regression techniques are actually implemented on business data, and how multiple models are tried before finalising on a good model. A good model is a model that is technically good, the FIT is good, the R^2 is good, but also makes sense from a business perspective.

>>>