Welcome to this section that talks about a predicative analytics technique called Regression. A regression analysis is a very popular technique used in analytical problem solving. A many business problems typically will lend themselves to a regression-based approach. Especially, when we are trying to predict or understand future outcomes using data that already exists on consumer behaviour.

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

In the course of this class, we will start with an intuitive understanding of regression analysis. Once we understand how a regression works and more importantly, the assumptions that are required to make sure that a regression approach really works, we will take a look at more complex regression methods. A couple of things to remember when we start understanding regression techniques. Regression theory is a statistical technique. Essentially, we will need to understand probability distributions, outcomes and inferential statistics to understand well how a regression analysis works and how it is applied. Therefore, it is recommended that you brush up on basic statistical theory before starting regression modelling.

It is also very important to understand the underlying assumptions that are required before creating and interpreting a regression output. Because otherwise you may have very misleading results. So what we will start with is understanding what the simple linear regression is and talk about

ordinary Li squares, how to run a regression and how to interpret regression results with a simple linear regression. Then we will add multiple linear regression, which is multiple explanatory variables, show you how to run this and interpret the results. We will spend some time understanding the assumptions that have to be satisfied to correctly apply regression models.

So far we are still talking about linear regression. But there are many other regression models including logistics regression which is very extensively used in Analytics. For now, we will focus on linear regression models which are the simplest regression models and intuitively the most easiest to understand. We will of course look at the regression techniques in the context of business analytics. So we will understand how regression models are used for business applications and business problem solving. Once we review the regression assumptions, we will then look at an application of a linear region to a business dataset. And we will see how linear regression models can be built in R.

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

A regression model is used to understand and quantify Cause-Effect relationships. What is the cause-effect relationship? Suppose that we have a brand of shampoo and we are looking at sales of that brand and suppose in this week, I offer a discount of 15% on the price of that shampoo. What do you think is going to happen to the sales of that brand in that week? We would expect the

sales to go up. More customers may buy the product because the price is now lower and therefore the sales will go up.

What is the cause here? The cause is reduction in price, which leads to an increase in sales. So the effect is an increase in sales. A regression technique is used to understand and quantify cause and effect relationships.

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

We already know that the effect of a decrease in price is an increase in sales. But what if we also want to know how much the sales will go up by? If I reduce the price by 15%, how much will my sales go up? That is the quantification of the impact of the reduction in price. The regression technique tells us what is the impact and what is the quantification of the impact when we have the cause and effect relationship.

In fact, if you look at the definition of a regression, Regression analysis is a statistical technique that is used to infer the magnitude and direction of a possible causal relationship between an observed pattern and variables assumed to have an impact on the observed pattern.

What is the observed pattern in our simple sales and price example? The observed pattern is the change in sales. The variables assumed to have an impact on the observed pattern is price. We are assuming that the price will have an impact on the sales and therefore when we change price, we

should see impact on sales. Why are we saying possible causal relationship? Because sometimes, we think that there is a cause and effect relationships and we use a regression technique to see if that relationship actually exists.

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

If you look at the definition, there are some important keywords.

Regression is a statistical technique. It is a mathematical approach that assumes that the pattern of interest, for example the sales variable, and the variables that impact the pattern are all random samples from an underlying population. So a regression technique is a statistical technique.

We are looking at the magnitude of the impact – how big is the impact – 2 times, 10 times and the direction of the impact – positive or negative. For example, what is the direction of impact when we change price? If price goes up, the sales will go down. If the price goes down, the sales will go up in most times, not necessarily always. But that is what we expect for most products.

So the direction of relationship between sales and price is negative in this case. Remember we are looking at Cause and Effect relationships and we need to be very clear what is the cause and what is the effect. The magnitude of rain fall has an impact on crop yield. But, does the crop yield have an impact on the magnitude of the rain fall? NO.

The observed pattern is the target variable, also called dependant variable. It is the variable that we are interested in understanding what causes changes in that variable. So that was a very simple overview of a Regression Analysis.

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

Let us start with understanding a simple linear regression. Now in order to understand linear regression models, we will start with a problem statement.

Suppose you work for a hospital and you are trying to understand the factors that may influence the birth weight of a baby. Now Let us think about a birth weight of a baby. For a baby to be born healthy at a healthy weight, what sort of factors do you think influence baby's birth weight?

- Certainly maternal diet – what is mom eating
- The period of gestation - Is it a full term pregnancy, 40 weeks or is it under term?
- Does the mother have other health issues?
- Maybe, the ethnicity of the mother
- Maybe, the age of the mother

So there are many factors that we can think of that influence the birth weight of a baby. Now you may want to figure out, how do I come up with model that will tell me, given these values of these variables, what should be the expected birth weight of a baby? In other words, can you come up with a relationship model that says birth weight of

a baby depends on these factors and this is how it depends on these factors. This is the magnitude and size of impact of each of these factors on the birth weight of a baby.

Now suppose I wanted to understand and work on this problem, how would I actually approach it? Let's think about the data itself.

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

Is it possible to analyse the population? No. Because many children are born every day across locations, across countries, across geographies. But we don't have to analyse a population, we can rely on inferential statistics and analyse a smaller sample. As long as the sample is representative, we know we can make inferences about the population with a certain degree of confidence based on what we find in the sample.

Let us look for a sample dataset. Here is an example of a sample dataset. This is 1115 observations from a hospital in the US. What does it capture? – it captures

- the weight of the baby when the baby was born,
- the weeks of gestation (40 weeks, 26 weeks, 36 weeks, etc.),
- the mother's education in years - Did the mother have any education or not?
- Race (Race is what is called an indicator variable. We will talk about indicator variable and how to interpret it as well,

But essentially, it has values of only 0 and 1. The value is 1, if the mother is African/American and 0 if the mother is not African/American. So this is essentially capturing Ethnicity) and

- We have a health related variable, which is whether or not the mom smoked during pregnancy (1 means that the mother was a smoker and continued smoking during pregnancy and 0 means that the mother did not smoke during pregnancy).

This is the data we have and we want to use this data to understand what factors influence the birth weight of a baby or Do these factors influence the birth weight of a baby? If yes, how do they do that? Before we start the analysis, Are these the only factors that impact birth weight of a baby? Not necessarily. There are other factors as well.

But we don't have data on those factors. So what do we do now? One of the things to remember is that in a business context, the data availability is variable. Sometimes the data is available and it is being captured actively. All the data that you need for an analysis are available to you. But very often, very little bit of the data that you need for the analysis may be available. May be 50%, 70% of the variables that you would like are available to you. Because some of the data are simply not being captured or it is being captured passively which means that it may not be completely reliable.

So in such a situation, what do you do? Ofcourse you try to do the best with the data that you have. At other times, you may be able to start a project where you design a survey to collect the data that you need during the analysis. In that case, you have lot more control over the data that is available to you. Obviously the reliability and the predictive power of analytical models will depend on the quality and comprehensiveness of the data available.

For understanding these models and the techniques, we will assume that we have fair amount of data available. But in real life, sometimes, you may not have all the data available. So part of the problem may be finding enough data to model or making the best of the limited data that you have. It does not mean that we will not create analysis or models. But we understand that the models can take us only part of the way. So coming back to this dataset, we have 1115 observations; we have 5 variables. The dependant variable or the variable that we are interested in is the birth weight of a baby.

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

We want to use this data to understand how do these factors or do these factors influence the birth weight of the baby? If yes, how do they influence it? Remember, we are doing cause and effect. The effect is baby birth weight and possible causes of variation in birth weight are in this dataset the number of week of gestation, the

mother's education, the race and smoking during pregnancy.

Now Let us just step back for a minute and think about how do we actually assess relationships between variables. There are many ways of doing that. One way is to generate some visualizations of the data and other is to generate correlations and possibly the 3rd option is to run a regression model. But why do we need regression models? Why can't we just visualize the data, the pattern, how the variables interact with one another and why don't we look at correlations

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

Regression is a better way to identify relationship between variables over and above a visualization or correlation. Why? One thing to remember is that when we do correlations or even when we do visualizations, typically, there are looking at two variables at a time. But in real life, multiple variables have an impact on that effect. For example, in the birth weight dataset, the gestation weeks has an impact on the birth weight and smoking during pregnancy also has an impact on the birth weight. If I simply looked at a correlation between the gestation weeks and birth weight and got a correlation coefficient of 80%, I am not including the impact of all the other variables that also impact birth weight.

The regression allows us to capture these multiple factor impacts on the effect. Also Regression will tell us the statistical significance of the impact. So

when we do a visualization, we can eyeball it and say there seems to be a relationship; there seems to be a pattern and we may be able to come up with an approximation of the size of the impact. But what it won't tell you is whether or not the impact is statistically significant. Because remember, we are dealing with random samples from an underlying population. So we should look at statistical significance of the impact. In order to understand how all of this works; we will start with the simplest type of regression models which is called Linear regression model.

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