Week 3: Regression – Descriptive and Predictive Analytics

Video 1: Descriptive and Predictive Analytics: Overview

Welcome friends! In the last module, Professor Sudhir talked about various data pre-processing steps, the types of data used in the business analytics, how to clean and wrangle data? How to impute missing values? How to do dummy variable coding for non-metric data? and basic statistics.

So, once we are done with the data pre-processing step, our data is ready for the next step, which is for the machine learning tools. So, today I'm very excited to talk about one of the most popular supervised learning tools called Regression. Regression.

Regression is probably one of the oldest, yet very powerful and very popular supervised learning method. Regression can be used for descriptive analytics as well as predictive analytics. So, join me folks today, to learn regression.

Video 2: ML Approaches in Business Analytics

So, there are primary 2 types of questions, businesses are interested in, from insights from the data. Either they are interested in understanding how variables affect each other, if you change one variable, how the other variable will behave, or they are interested in making predictions, how prices affect the sales, for example, which kind of customers will churn? What kind of drugs will be effective for what kind of disease? To understand all this, businesses use supervised learning tool.

The other kind of questions they are interested in from the data is, to identify the new and novel patterns which may not be visible to the naked eyes, which they are unaware of. For example, how many different kinds of customers exist in the market? What are different customers saying about my product on the social media? We can do the text mining, we can identify some hidden patterns and that might help them come up with the new products and offerings.

At the end, the entire machine learning domain is all about pattern recognition. Supervised learning models use to discover dependencies between 2 variables and the unsupervised learning models usually try to discover the interdependencies between the variable. And the popular supervised learning techniques are, decision trees, regression, linear discriminant analysis, logic models, etc., and etc.

The goal in the supervised learning model is to uncover dependencies. The relationship between 2 variables in presence of or maybe in absence of the other variables. More precisely, the supervised learning models try to identify the mathematical relationship between the variables of interest, often denoted by the Y, and the variables which might be affecting the Y variable, which are denoted by X.

The most amazing part of this approach, the machine learning is, that the machines identify the relationships. Unlike in the the usual approaches what we see, that the managers actually tell machines that use this, this, this and then kind of you know help them make the predictions. Here, the thing is completely upside down.

Here, the machines are told, "These are the decisions I made. Can you identify? can you tell me how I made the decisions?", and that's the best part, that the experts you can actually have the data from the past and then use that data to discover how the earlier managers or you, yourself or the other experts actually were making decisions, right? So, now you can use those relationships or the patterns you have recognized to make predictions, and one of the most popular technique for this kind of work is still a regression.

So, what is regression? regression? Regression is a supervised machine learning tool. So, we have always a variable of interest, we call it a dependent variable and is denoted by Y. And the other variables which we think are related to that dependent variables are called the independent variables, are denoted by X.

So, look at this picture. So, suppose I have 1 dependent variable, which we codded as Y and the independent variable, and we plot these variables, like we try to visualize it. As you can see, there is some pattern here, right? The goal of the regression is to kind of identify what is that pattern? To find the equation which best fits the pattern. That's what it is, a pattern recognition. The only thing which you have to remember in the regression, these patterns are identified as a linear pattern, and we'll talk about that in a minute.

Video 3: Regression: Motivating Example

So, let's look at the real-world example from the retailing industry, to understand regression. In the retail industry, managers are interested in knowing whether certain kinds of promotions are working or not, what is the impact of their promotions on sales of the products they are promoting? For example, what is the impact of featuring a brand in a flyer on the brand sale? So, what is a flyer? So, remembering your newspaper, often you see there are inserts.

In these inserts, your local grocery stores might be telling you that there are certain kind of products which are on deal, come to our store and get it, right? So, those are the things which call the feature. Managers would be interested in knowing, are these things working? Do we have any impact of putting these flyers in the newspaper on the sales of the things which we have described in the in the in the flyers? They are interested in knowing, when they put things on display, maybe end of the aisle, or close to the checkout counter, does it impact the sales of those brands?

They might be interested in knowing, if I change the price of the brand, how does it impact the sales of that brand? For that matter, if the competitors changes the prizes, does it impact my brand? If it does, how much? What if the, I have a store brand, if I change the price of the store brand, how does this impact the sales of the brand A? Are their synergy effects between the two different marketing activities we are doing? For example, if we do feature and display together, does it have a proportionately higher impact on the sales? Do they work in synergetic way? or they're cancelling out each other? They will be interested in knowing or predicting the sales for the next period. So, how can they do that? So, let me kind of you know go back a little bit and talk about the data driven approach and the four major steps one has to think through.

So first, you need to identify, what is the business challenge you want to address? what do you want to achieve? and why? Then, u kind of, you know, ok what do I know about this context? What kind of information we already have? Can we address a problem with the information currently we have? Or do we need additional information? And what kind of information?

Then comes, what kind of tools I can apply to actually get to that information? What kind of data would be needed to get to that insight? and of course then comes, you use that tool, you interpret the results, maybe make predictions and largely gain insights. So, let's put this in the context of retailing, right?

So, suppose a manager is interested in knowing how does the prize, my prizes and the promotional activities are actually impacting the sales of a particular brand. So conceptually, what he's thinking, the sales is my dependent variable, which I am interested in knowing how it is related to my own prizes. If I change the price, what would be the impact on the sales of this brand? What is the impact of the change in competitors prices on the sales of this brand? What is the impact of promotional activities like feature in display? How does that impact the sales of a brand? So, all these things which we talked about, the prizes, competitive prices, feature in display, these are your independent variable and the sales of a brand which you are interested in, becomes a dependent variable.

So, remember, in the regression, regression, regression assumes this relationship to be linear, right? So, if you think of this conceptual model, it will translate into a simple mathematical equation which is, on the right-hand side, the Y variable will be sales, on the left-hand side, there will be things like pricing, competitors price, display feature, whatever other things you think are related to, or affect the sales. And of course, there will be things which you don't have data about, and those are the things which go into the error term.

Now, what do I mean by the mathematical relationship? So, think of this as a equation of a line. I want to discover what is the intercept of this line? what is the slope of this line? and in this case, it's a multi-dimensional plane, right? So, same idea goes through.

So, let's come back to the data. What kind of data do I need to run or understand the things which I'm interested in. In this case, if you have a weekly sales data of a brand which you are interested in, if you have a prices for that brand, what prices you have set, and what prices competitors have set, you have the data about whether this particular brand was featured in a flyer, you have a data about whether this brand was on a display, whether it was displayed at the end of the aisle, or it was displayed next to the counter.

So, if you have this data, you actually can run this simple model. So, in this case, you will find this data on your LMS. It's a small dataset just to kind of you know, give the flavor of you know, how to run the analysis and how to get the insights. The data is from one of the big retail chains from one of these stores.

So, usually in the retail settings, manager set the price once a week, that's why we're using the weekly data. How much was the sales of brand A? And what was the price of brand A in that week? Price of competitors brands? whether this particular brand was in feature or display? All this information is coded. So, that's the row of the data, right? And I'm sure Professor Sudhir talked about the structures of the data. So, we're going to use this data and run the regression and see, can we answer some of these questions which managers are interested in?

Video 4: Regression Demo: Summary Stats

So, when you look at the Regression App, first thing you will notice is that it looks different from the Data Preprocessing App. So, let me walk you through the options here. On the left side, you have an option to upload the data. So, what we're going to do is we are going to

browse where our data is, in this case, we're going to go to the orange juice sales dataset which you have downloaded from the LMS and upload it on the App.

And once you have uploaded the data, the first thing you want to check is the Data Summary tab, to see whether the data is uploaded correctly. So, here you can see that the whatever data you have uploaded, whether it's loaded correctly or not, you can look at the number of observations, more observations.

So, what we see is that the data is correctly uploaded. Now, there are 116 rows in the data, there are 10 columns, so everything looks fine in terms of the upload. Now, the next thing you want to look at is the summary statistics. Here the summary statistics is divided into two parts. Summary statistics for the numerical variable.

So, remember sales is the continuous variable, numerical variable it can take different values. So, you are going to see, what is the minimum values its taking? What is the median of all the observation? What is the mean? and what is the standard deviation? And similarly for the price variable, what is the mean and what is the standard deviation and so on and so forth. And below that you will see the summary statistics for the categorical variable.

Now, the App actually is smart enough to figure out sometimes that what are the categorical variables. So, what happens is that if you can see, if you see here you see the Display is selected as a categorical variable and the way the App actually figured this out is by looking at the Display column and when it finds that there are some values which are character, basically, then the App knows that it must be a categorical variable and that's how you basically, automatically checkmarked it as a categorical variable.

But there is one more categorical variable in our dataset, which is the Feature, which takes the values zero and one depending on whether the Brand was featured that we got now. Now, why this is a categorical variable because we know that any intermediate value doesn't make any sense, for example value 0.5 has no meaning. So, because it's basically one is kind of a proxy for Yes, it was featured and zero is the proxy for No.

So, because App when you look at the column, it found that all these values are numerical, so, it doesn't know whether it's a numerical variable or the categorical variable and that's where we need to help the App to figure this out that tell the App that it actually is a categorical variable.

The moment you will checkmark it as a categorical variable, earlier it was thinking that the feature is the numerical variable, the moment you tell the App that it is a categorical variable, you will see it appears in the summary statistics of the categorical variable, Feature out of 160 observations, Feature basically and the 55 observations they took the value zero and for 61 observations it had a value of one.

And similarly for Display, Display can take three values, whether it was not displayed at all, whether it was displayed next to the checkout counter, which is basically marked as chk and whether it was displayed next to the end of the aisle. So, 20 for...out of 116, 20 observations are for 20 weeks. It was displayed next to the aisle and for 12 weeks it was, this particular Brand was displayed next to the check out counter.

So, you can look at the summery statistics. And below that what you will see is that the missing data. So, if there are any missing values App will actually tell you that these are rows in which the missing values are there. In this particular dataset, there is no missing value, so, it's basically telling you zero rows with the missing value.

But in the other dataset, if you had missing values, you know in the Data Preprocessing you have to handle the missing values, but I have also created an option here in this App where if your dataset has missing values, you can either impute those missing values or you can drop those rows. In our case again, we didn't have any missing values.

So, even if I choose any of these options nothing will change. You will see 116 observations remained. Okay. The next tab basically tells you again the basic statistic of the visualisation of the dataset. So, here you see the histograms, how the price of the Brand A basically the range is from 1 to 3 and a majority of the time is basically aligned within 2 and 2.5.

And sometimes it was very low priced between 1 and 1.5 and display because it's a categorical variable, it tells you how many observations are zero. How many observations basically belonged to the end of the aisle, how many observations belong to the display at the checkout counter.

So, below that you will see another kind of information which tells you graphically how the two variables are related. For example, if you look at the sales of Brand A, how it is related to the price of Brand A. So, price of Brand A is plotted on the X axis and the sales is plotted on the Y axis.

And you can visually see some kind of pattern here. So, this make as the prices increase sales is dropping. And that is what is captured here in the correlation also. If you look at the correlation between the sales of Brand A and the price of Brand A is minus 0.714. This minus sign is basically telling you the relationship between the price and the sales is negative.

Okay, and for example, in the other case, you can see that the whenever you see the positive value basically like for example, in the week case, you will see that as the week increases sales is going up a little bit. The strength of the relationship basically comes from this number. So, this is how you interpret the correlation table and look at the kind of get the sense of how different variables are related to each other in the dataset. Go to the summary OLS tab and look at the output. In the output what you will notice is that it tells you what Y variable you have selected.

So, you have selected the sales as Brand A so, here it will show that Y is the sales of Brand A and then it also tell you the estimates of the regression module. These are the beta coefficients. We will talk about that in a minute and then it will tell you the significance of these betas through the stars. It also tells you how good is your model basically through the adjusted R square, you can get the sense of how well is your model performing. Then the next you want to look at the Residual Plots.

The residual plots basically tell you whatever model you run, how good that is, in another words, so the X axis basically is the actual Y variable. The Y was the sales of Brand A and this is the predicted sales. If the actual sales and the predicted sales are matching, there you will see all these observations rely on this a 45 degree line.

But as you can imagine, the model is just the approximation of the reality and we don't have all the pieces of information so there will always be some kind of error in terms of the model prediction. And when you look at the actual data and that's why you see these observations are scattered around this, this line.

The other thing you want to look at is maybe the residual plots its the same information, but presented in a different way. It tells you like remember there were 116 observations in the dataset. So, observation one how much is the error, observation 100 how much is the error.

Now, as I said in the ideal world if your model was perfect, all this will stack and you will have the zero errors.

But we don't have enough information or we don't have all the variables which affect the sales and that's why you see, these points are not all lined up at this line. So, the way the OLS does is basically measures the distance between this point, each and every error point. This is...these are the errors. This distance is basically the error in the prediction for observation one and for observation 100 this is how much is the error and it tries to minimise that. It tries to basically come up with this best fitting line in the sense. So, that these errors are minimised.

And that is what is basically you can see here numerically also that you take all those errors, you square them and normalise it and you would see the value here. The lower the number that means that there is a less error so all these points are all stacked up around this line then that number will be very low, if they're all very far away that number will be big. So, that's another way to kind of get the sense of how well is your model.

Now, the next tab basically tells you whatever data you uploaded, what is the predicted value and what is the actual value. So, remember these are the difference between these two is the error and the OLS tries to minimise this error by betting the right beta coefficients. Now, when you build the model whence you have satisfied that by selecting the correct variables then the next thing you want to do is you want to use that model for the prediction purpose in the new dataset.

So, where you don't have the sales available, for example, you want to predict the sales in the next period. What would be the sales look like if you put something on Feature and Display, what the sales would look like if you don't put that on the Feature and Display. So, that you can build that dataset with all the X variables or you already have that X variable, you can upload that dataset here and then once you upload that dataset, it will basically give you the predictions.

So, this is the data you have uploaded. Of course, I have uploaded the same dataset that's why you see the Y variable is also present here. But in your case, because you are interested in predicting the sales if you already knew the sales what's the point of building the model. You didn't know the sale and that's why you wanted to build the model to make the predictions.

This is the original dataset and it has... here are the predictions. So, here are the predictions for observation one, observation two and again this looks very similar because I have uploaded the same dataset which I use for the input for the model calibration. I'm actually using the same dataset for the prediction and just to show you how to use this tab.

Video 5: Regression Demo: Input Data

So, before we move on to interpreting the beta coefficients in the OLS model, let me talk about the categorical variable for a minute. How the OLS regression handles the categorical variable?

So, remember in our orange sales data set, display was the categorical variable, and it can take three values, right? It can take the value zero, chk, or end, depending on whether the particular brand was on a display next to the checkout counter or on the display next to the aisle. So, what regression does is, basically, take this original data and creates a multiple columns out of it. How many columns? These number of columns is exactly equal to the number of different values this particular variable can take.

So, for example, in this case, it can take three different values, and that's how the three different columns will be created. And now, how we populate this entire matrices? If you look at the observation 1, the display was coded as zero.

So, in the new, in the dummy variable regression, the display zero column will be on and will coded as one, the other two will be called as zero. And if you look at observation number 3, the display takes the value of chk, and that's why you see the display chk column is on and the other two columns are off. And that's how you convert the categorical variable into the dummy variables, which is also called the one hot coding.

Now, one thing to remember is that in the regression, you will see, for the categorical variables, you will see multiple betas associated with every categorical variable. So, for example, for the display, you will see two different beta coefficients, and these two beta coefficients will belong to the last two kind of, you know, columns.

For technical reason and for the identification purpose, the first column once you convert the categorical variable into the dummy variable, the first column is always dropped, right? And that's why you get only two beta coefficients and how to interpret those beta coefficients? We will talk in a minute when we talk about the how to interpret the beta coefficients from the OLS model.

Video 6: Regression Demo: Interpretation

Let's look at how to interpret these results, right? So, for that, click on the "Summary OLS Tab" and there you will find the model and the output of the model. So, pay attention, there are a lot of information here. A lot of managerial relevant questions, which you were asking, the answer is here.

Now we have to figure out how to get the answer from this output, how to extract those insights, how to change this mathematical form, back into the the language that managers can understand, or you can understand. So, for example, remember the question we were interested in? How does the prize of brand A changes its sales?

So, what you're going to do is, you're going to look at that price of brand A row in your summary of OLS tab in the output, and what you will see is the estimate which is -24,561. And then, there is a bunch of other information.

So, let's start with interpreting what does this -24561.45 mean. It means that, if you change the price of brand A by one unit, in this case the units were a dollar, if you change the price of brand A by one dollar, the impact on the outcome variable, which is the sales of brand A is

24,561 units, ok? What does this minus sign means? minus sign means? Minus sign means, if the price of the brand A is increasing the sales is going in the opposite direction.

It makes sense, right? If you increase the price of something, what do you expect the sales? The sales will go down and that's what this minus sign is telling you, the relationship is inverted, right? So, if you increase the price, the sales go down, downwards sloping, right? This kind of.

Now, are you confident about this number? Can you be certain about that this is the right, this is what you would expect? For that, you have to look at the standard errors, t value and the probabilities, right? But here, in this app, what you will see is, is a very easy way to, for you to identify that. At the end, you will see the stars. The three star means that you are very very very, very, very confident that -24,561 number is a very statistically significant.

So, that means, if you look at three stars means, the probability of this number being zero is .001. That means you are 99.9% sure that this relationship is holding, again, roughly. If you see two stars, then you're 99% sure. If you see one star, you're 95% sure. And if you see a dot, that means you're 90% sure. And if you're not 90% sure, then pretty much you can safely assume that relationship is not robust enough to survive, or what you have identified, the relationship doesn't exist between these two variables.

Now, there was another question as a manager you were interested in. For example, how does the display affects the sale of a brand? If I put a brand on a display, this particular brand on a display, either on the checkout counter, next to the checkout counter or end of the aisle, does it have any impact on its sales? And remember this is beyond the impact of the price changes or the changes of the competitors prices or weather this brand was in feature and all.

So, if you look at these numbers, so you see only two outcomes, right? So, although display can take three values, it's a categorical variable, remember? It was a categorical variable. But remember from the technical note, that one of the columns is dropped before running the regression and the column which was coded, that display was not on or this particular brand was not on display, that particular column is dropped, and that dropping happens for the identification purpose, and I'll talk about how to interpret that.

So, first thing you notice is, if this particular brand is displayed next to the checkout counter, you see the number -234. Well, the minus is telling you that if you actually display a brand next to the checkout counter, the sales are actually going down. Wow! That doesn't make a lot of sense. Oh! Okay okay. I haven't seen the stars. Oh! There are no stars in front of it. So, pretty much what -234 is, is no different from zero. That's what the stars are important, right?

Now you see, why we have to look at the stars at the end, which basically tell you how confident you are. Now, if you look at the the, the second variable which is the display at the end of the aisle, what you see, the impact of that is positive. If you put a brand next, at the end of the aisle, if you display this brand, the sales go up by 20,450 units on an average. Wow! That's a pretty strong effect. Can I be confident about it? Let's go and look at the stars. Oh! There it says three stars. Wonderful. That means this particular result is 99.9 percent I'm confident that this holds.

That's how you interpret. Now, one small thing which I wanted to cover here, when you are interpreting the categorical or the factor variables, technically what you're saying is, relative

to the base value and the base here is the column which is dropped. In this case, the brand was not on a display, it becomes the base.

So, these effects are relative to that, right? So, if the brand is not on a display compared to that, if you put it on the display next to the checkout counter, you have seen no effect because it's not significant. But if you put that brand next to the aisle on a display, then you see a significant effect and that effect is the sales increase in that week by 20,450 units.

So, I hope you understood how to interpret the, this table which has lot of numbers, but really the numbers which are really relevant for you is the first column. What are these Beta coefficients? The estimates of the Beta coefficient and the star values. Is it a three star, how confident you are? Or is it a two star or one star. By the way, Just just for your additional knowledge, let me tell you one thing.

These stars are generated by the T values, right? So, another way, even if you don't have a star, you can look at the T values. If the absolute number of the T value is greater than two, you will see it has at least one star, and you can check on yourself in the output.

Video 7: Regression Demo: Evaluating the Model

So, now you understood how to interpret the, interpret the results of the regression model. Interpreting the results basically is all about understanding the relationships, that falls under the, we're going to call it as a descriptive analytics.

Now, as I said, this start, regression is also used for the prediction purposes. So, let's see, how that is done. Before you go to the prediction, you want to see how good is your model. You have seen in terms of the interpreting the beta coefficients that whether this relationship was significant or not by looking at the stars and all.

So, can we say something about the overall model, the full model, with all the variables together, how is it performing? To do that, let's start with, look at the "Input Data with Prediction" tab. What basically this tab is showing you, is that what was the actual Y, which is the sales of brand A? And what is model predicting it to be?

So, for example if you see the first row, is predicting, given all the X variables, the model is saying the Y hat will be negative number 1651. Negative sales, what does that mean? and why this model is predicting? What can we do? But before we come to that, let's try to understand what really the model did, and why is it predicting 1651 negative number in one particular instance? To understand that, let's look at this picture.

Remember, what is a regression analysis regression analysis? Regression analysis is technically what is doing is is looking at the different data points and trying to fit a best line. What is the best line? Best line could be, if you see there is a yellow line, maybe that's the best line. Or is it a green line which is a better line? Or is it a blue line which is better? So, how does a model figure out whether the line should be blue, green or yellow?

The way it figures out, it basically measures the distance between that point and the distance from the line. It basically measures all these distances, take the squares of these distances and sum it up and runs the, kind of you know, algorithm in the background and try to minimize this error. So, this is the error, right? Actual data is that point, and the line is what model is saying the relationship is.

So, there is a deviation between the actual data point and what the model is actually picking up. And what you want to do is, you want to minimize these errors and that's what basically

the background regression is doing, it's measuring, it's creating different lines and measuring these errors, taking the squares of it, adding it up, and then doing it for the another line, for another line, another line, and kind of picking the best line.

Now, in this case it just turns out the best line is whatever it is, right? And you have seen the prediction, how it was predicting. But this is the best you could do. Now what is the other way to look at how good my model is doing? So, actually if you go back to the model summary tab, what you will see, below all this beta coefficients, there are a few more things which were there. For example, one was the Multiple R-squared and that number was 0.6888.

And the other number next to it was the Adjusted R-squared. And these are the things which basically is a measure of how good is my model. So, the higher the number, the better my model is. So, what's the difference between multiple R-squared and the adjusted R-squared?

For all managerial purpose the one thing you remember is Multiple R-squared is a raw number and the Adjusted R-squared basically adjusts for the number of parameters in the model and we'll come to that in a second why that is an important thing.

Other thing often people look at is the F-statistics and you see the p-value there and you see the p-value is less than 2.2e to the power -16. This is a very small number. That number is basically is telling you the probability that this model is not very good and the probability of this model being not very good is very small, so we are pretty confident, right? The probability that this model is not good is 2.2e to the power -16, 16 zeroes and then point two, this is very small for all practical purpose.

We can say the probability of this model being not good is zero. Now the next thing you will do is, okay, we have done all this, and suppose you have figured out and you are satisfied that this is a good model because when you look at the R-squared, the R-squared, the R-squared is 0.6888 which means the value of the R-squared lies between zero and one. If it's zero, that means model is not good, if it's closer to one, it's very good.

Now, depending on the context, .3 value could be very, very good number and sometimes people look for value of very high numbers and that all depends on the context and we will come to that in a minute too. The other thing, once you are satisfied with the model, it's okay, my model is good .7 is is good enough, is very close to one, actually it is a pretty good number if you ask me in this context, because what it's telling you, the sales of a particular brand is quite nicely explained by these five variables which we pick, its own prices, it's competitors prices and the promotional activities, right? And then basically it's telling you right again intuitively think about 70 percent, right? .688 is roughly 68 percent or 70 percent, ok.

The next thing when you are satisfied with this, you want to use this model for the prediction purposes. So, how can you do that? I mean the whole idea of, that you have learned some patterns, the relationship between Y and X variable, you want to exploit those relationships, for making predictions, and for that again, pretty easy, you go back and look at the left side of the panel, you will see there is a tab which says upload new data for prediction.

The only thing to remember there is, the structure of the data should be very similar to the structure of the data which we have used for the analysis. I mean, of course, you won't have the Y variable there, but you must have all the X variables. So, what the model will do is, basically will compute all the X variables, we have already know the Beta Coefficients. beta coefficients, beta coefficients will be correctly multiplied with the right X variable and then mathematically you will get the answer, what is the Y variable.

Video 8: Orange Sales Data: Improvement

So, now you have seen, how to make predictions using this particular app. But, before we go there, let's kind of a roll back, are there ways I can improve the model's predictive power? Remember in the prediction, the first row was predicting the negative number. Is there a way actually to improve your model? Maybe capture a better relationship. Now, when we say a linear regression model, the assumption is linear, doesn't mean that the X variable have to be linear. You can put the X as square.

So, look at this picture, you see the bunch of data scattered around which are plotted. The topmost plot shows multiple R-squared, if you run that regression between X and Y is 0.07044. But if you look at carefully, maybe there is a quadratic relationship, there's a non-linearity. So, do I have to run a non-linear regression model? No, no no, no, No, no, no, no, no, no. You still are in the domain of a linear regression model. The only thing you have to do is to capture that quadrature or the curvature. You're going to put the X squared term.

So,as you can see in the picture below, you can capture the non-linearity by putting the X squared term. Or you can actually think of you know, maybe this is, maybe more non-linear, more than non-linear than the second order. So you can put the cubic term once. You can keep on increasing. In fact, you can complete the entire Taylor series. You can put all X to the power n terms, and what will happen? It will actually fit the data better and better and better and better. So why we don't do that? Maybe we should put all the higher order terms, X, X X squared, X cube, X four, X five.

You know what, what what will happen? Yes, the model actually will fit the data back better. But that's not the goal. the goal. The goal is to kind of you know, infer the relationship, not closely fit the data. So, it is yes, it's a pattern fitting exercise, a pattern recognition. But there's a there's a trade off, and that trade off is, you know, if you are including the X squared terms or higher-order terms, always you will see the multiple R-square is increasing. I mean, at least will never decrease. And that's why, there is something called the adjusted R-squared.

What adjusted R-squared does is, basically captures this effect that how many variables you are putting in the model. The more variables you put, although the multiple R-squared may increase, but the adjusted R-squared will start decreasing. And that tells you actually, when you're putting more variables it comes at the very expense of over fitting the data.

And in the later model we'll talk about this idea a little bit more. But for now, still to capture the non-linear effects in the data, what people do? We put the higher-order terms and usually at the best in these kind of regression models, we never go beyond the second order effects, right? So, X and X squared, we rarely see the X cubes and X fours, because I have not seen. Or, the other very popular approach is, to transform the X variable, and instead of putting X, you use the log of X and log of Y. That's called the log-log model.

Now you must be wondering, okay, this was simple, I have uploaded the data, I figured out what are the X variables, I need to pick that, and then, the next step was within that, figuring out which were the categorical variables, it was all straightforward, then I look at the output and interpreted all fine, this life was very simple and easy.

So now, this is another challenge that we have to identify, should I put X squared term, X cube term? And now we have to remember five different variables, for which variables X squared terms need to be added, and for which not to add. How will we come to all this? And that's where a lot of managerial knowledge comes in.

Managers should have a fair idea, are these effects linear or non-linear. For example, if you spend more and more on advertising, do you think the sales will increase, keep on increasing? No, there is a saturation effect. The more you advertised after a while, the effect of advertising will start plateauing out, right? So that's a non-linear effect. That means, you need to kind of you know, capture that through not only putting the advertising but the advertising squared terms also.

Past research will guide you and that's where the theory becomes very important. Theory will guide you. What kind of you know, non-linearity you should expect in the data and must capture. In fact, there is a whole area of looking at the residuals and I'm not going to go into those things in this particular module, which also helps to identify what kind of variable transformations are needed.

Should you put the X squared term? Should you take the log of a variable instead of it's just the X term? But for all practical purpose for your, I would suggest the easy way out. Easy way out is, the old goal standard, trial and error. It cost you not much, right? You just have to upload the data and interpret the results and keep looking at the adjusted R-squared term. If the adjusted R-squared term is improving, you say, "my model is better than the last model", right?

So,okay, let's get back to the example we were looking at. In the retailing context, the idea was to predict the sales based on the input variables like your own prices, the competitive prices, the feature and display, and the way we modeled it with a simple linear equation, what you're seeing on your monitors, right? It's a very simple linear model.

But the retail managers know from the training that the elasticity is the central idea when they're thinking about the prices. And the relationship between the sales and prices can be much nicely captured, if we actually run a different kind of model. The model where the, instead of running the regression of prices on sales, you take the log of sales and log of prices.

Interestingly, when you take the log of sales and log of prices and run the regression, then you can interpret those beta coefficients which are coming out of the regression directly as elasticity. So, what is elasticity? elasticity? Elasticity is a measure which tells you, instead of directly saying, if one unit of exchanges what is the impact in terms of sales of brand in units, it tells you the percentage changes in prices, how does it affect the percentage change in the sales. And that's basically what the elasticity means.

Elasticity is a measure of percentage change in something and how does it affect the percentage change in something else which you are interested? And how do we run this model? Simply, just take the log of Y and log of X, run the regression, interpret the beta coefficient and that beta coefficient directly basically tells you what the elasticity is.

If you are a retail manager or your friends are a retail manager, you will immediately recognize that how important elasticity is, and now you can see how easy it is to actually compute using the regression. So, basically take the log of sales, log of prices, run the regression, in that case the coefficient in front of a log of price, which is the beta one coefficient, tells you the elasticity of prices with respect to the sales. What does the beta two coefficient tells you? It tells you the cross elasticity. If the competition changes the price by one percent, what would be the impact of sales of a brand in the percentage terms? What about feature and display? In the feature and display, remember, these are categorical variables, so you can't take the logs of it.

Although feature is coded as zero, one, but still you can't take the log of it because as said, it's a non-metric variable. And by the way if you take a log of zero you know what will happen,

right? So, you can't take the log of the categorical variables, there are different ways to handle this problem, usually through the quadratic terms.

So, what I would like you to do is, open that CSV file which has the original data, transform or create two new columns, a column, log of price A, which is equal to the log of prices which were given to you, and similarly the log of price B, log of price C columns, and also a log of sales column, and upload the new data into your app.

And this time, remember to pick a Y variable as log of sales, then X variable as log of prices, and the feature and display. And do not pick the original variables, which were just the simple price and the sales. And try to interpret the beta coefficient and tell us what you found.

Video 9: Log-Log Regression on Orange Sales Data: Debrief

So, I hope you were successfully able to run the log-log model. And I hope you found the very similar effects. So, what I asked you to figure out what is the coefficient in front of a log of prices and what does it mean? So, you see, the coefficient is -2.51144. It's much smaller, than what we've seen earlier, 25,000 or something.

But remember, this is the percentage change in the prices and its effect on the percentage change in the sales, right? Because the sales was measured in thousands, that effect is kind of, you know, what has reflected, right? The change in the sales was big, but when you divide it by the the scale, then the percentage changes are usually small, right? and that's what it is.

But at the end, qualitatively still telling you, it's a negative number. If you increase the price, the sales decreases and more importantly, the number itself is a measure of elasticity, -2.5. Is it significant? Yes, if you look at the stars, in front of that row, there are three stars, so that means, you are 99.9 percent confident, and if you click the probability values, its to the power, e to the power -16, which is 0.0000001, is the probability that this number is, is a zero, right? So, you're very confident. So fine, good.

Video 10: Regression Variants: Main & Interaction Effects

So, let's look at another very interesting questions usually managers have. For example, in the retailing context, managers often think then if you do multiple promotions, they are more effective than these promotions individually put together alone. In other words, suppose your manager has a hypothesis.

When the brand has featured, as well as displayed, the total effect is far more greater than the combined effect of these things. So, can we test this hypothesis? And the answer is yes. In the regression framework, these hypothesis is tested through the interaction effects. So remember, our conceptual model was that the sales is a function of price, your competitors price, whether the brand is on display and feature, and so on and so forth.

Now you're adding the addition term there, that when the feature and display are together, they actually somehow create some kind of synergic effect, and together the sum of these two is greater than the individual components. To do that, what you have to do is, you have to create another variable.

The other variable, where basically you will multiply these two variables, where you think the interaction effects are operating and then try to interpret the coefficient in term, in front of that variable, what is the better coefficient? and by interpreting that, you will know whether these

interaction effects are positive, or negative, or there is synergistic effect, or they are cancelling out each other, is it significant or not, and so on and so forth. So, better coefficient basically will tell you all about, and this is above and beyond the individual effect of feature in display, right?

So, to do that, again what you have to do remember? It's, open your CSV file, create a new column where you're going to multiply these two things, display and feature, save that dataset, and now go back to your app, refresh it, and then upload the new data. And then choose this new column which has a display and feature. Let's do it, do it yourself and report, what did you find?

Video 11: Regression – Descriptive and Predictive Analytics: Summary

So, I hope you enjoyed the lecture and now you understood how to do the regression, how to interpret the results and to how to use it for the prediction purposes. Remember, regression is the workhorse model for the supervised learning. It's pretty straight forward and very powerful. What it does basically? It tries to identify the linear relationships, right?

Now, these relationships are linear in beta coefficients, not the Y and the X variables. You see, I can actually transform the Y variable and use log of Y, or I can use log of X, or I can actually throw in the X square term, but still, it is called the linear regression model. Why? Because the betas are all linear, and that's why it's called the linear regression.

So, just to recap, whenever we have the dependent variable and we are interested in identifying the relationship of that variable with other variables, which we code under X variable, often we use the regression model.

Regression model essentially identifies the linear relationship, which basically means, figuring out the right beta coefficient, and once you have identified the beta coefficients, you can use that to interpret the strength of the relationship, and how they are related to Y variable, and so on and so forth.

So, when you're doing the regression, what are the things you probably will focus on? Just to kind of you know, recap what we did in the lecture. We want to look at how good is my model? How do we do that? We look at the R squares and particularly the adjusted R squares because remember, if you keep throwing in more and more variables, the multiple R square probably will keep on increasing, it will definitely not decrease.

But adjusted R squares actually are a much better way to look at it, because it tells you that you are, kind of, you know, in some sense, not trying to over fit the data. Because remember, the idea is not to completely fit the data, rather try to infer the relationships the relationships.

The relationships which are more sturdy, which you can actually take outside this particular data and maybe use it for the prediction outside this data for another data. What else we do, we basically look at the beta coefficients that tells us, whether the particular variable matters or not matters by looking at this statistical significance of the beta coefficient and looking at the values of the beta coefficient itself, whether they are negative, positive, what is the magnitude, and so on and so forth.

We also looked at how to include, whether we should include the non-linear terms? Do we need to actually include the quadratic terms, and the way I said, either the theory will guide you or your managers have some intuition about it, that these effects are non-linear from the past knowledge, or you can just do the simple trial and error.

You can also test the interaction effect, are there synergetic effect between two variables by creating a new variable, which is the, the multiplier of the two variables which you are interested. And how will you know if these synergetic effects are positive, then that beta coefficient on that new variable will be positive. So, in together, these two variables actually do a much much, much better job than putting them together simply.

Or, if there are non-synergetic effect, that they're cancelling out each other, that beta coefficient will turn out to be negative. And the other minor thing, which in the app, which is done automatically for you, you'll be careful about picking the categorical variables, because there is a fundamental difference between how the categorical variables are handled and the regression versus the continuous variable.

So, this, the thing what we have learned today is also called the ordinary least squares. Well, there's nothing ordinary about it. It's still a workhorse analytical tool for the real world business problems. Very widely used in business and management, in social sciences, and if you look at the academic literature, you will see quite often, a regression is the tool which they have used to identify things.

As I said, it's a very, relatively very old, I think, probably one of the oldest method, but, it survived the entire era. Of course, there are better versions of it, but at the end, all kinds of regressions have the dependent variable and the independent variables, the big picture idea remains the same that we want to understand the relationship between the independent and the dependent variables. Now, there are few things which basically, we assume when we're doing the simple regression or the linear regression.

That one thing which we haven't talked about, that the independent variables usually are assumed to be un-correlated. Now, it's not a very technical thing, and usually, if you are thinking about using the regression model for the prediction purposes, you don't have to worry about any of these things. The only time you have to worry about these things, when you're trying to interpret the relationships, then you have to worry about the multi-colinearity, maybe you have to worry about the autocorrelation, homoscedasticity and things like that. Now, of course, these are advanced things. But let me tell you one thing, what we have covered in terms of the variable transformation, nonlinear effects, interaction effects, pretty much you have covered, if not 98% of the regression, definitely 95%.

And just to cover the last five percent, it takes quite a bit. But for all practical purpose, 95% is is a very very good for all managerial decision making. Remember, all these models are not the substitute of your decision making. They actually are helping you make the better decision. At the end, you should use them as a guiding principles, not kind of a, offload your work to them. As long as you remember that, you will see the regression is a very powerful tool.

And as I said, most of the business problems now you can think of can be transformed into the regression framework. You have some independent variable, which you are interested in, how it is related to the other variables, which we call the dependent variables.

For example, we looked at how the promotions and the prizes affect the sales, how the R&D spending have the effect on the patents, number of patents the firms are generating, how the gender composition of the company board affects it's financial performance, how the cultural orientation has an effect on the brand extensions.

Within the HR, you can think of how the different levels of satisfaction under different things, how it affects the overall satisfaction, and how that leads to, how likely somebody is to stay with the firm, or churn or go away or resign? But, at the end, there is a Y variable and there are X variables and the regression identifies that relationship, and you can understand those

relationships, get insights and more importantly, you can use them for the prediction purposes. So, I hope you enjoyed the class. We'll see you next time, and we'll discuss another kind of supervised learning tool which is the logistic regression.