Course Name - Operations and Revenue Analytics

Professor Name - Prof. Rajat Agrawal

Department Name - Department of Management Studies

Institute Name - IIT, Roorkee

Week - 01

Lecture - 05

Welcome friends. So, in our last class, we discussed the applications of predictive analytics for forecasting purposes. We discussed how you can use very simple methods of predictive analytics. You just need to understand the pattern of your data, and we started with a use case of forecasting that if you can see the pattern of the data and try to extrapolate the same pattern, you can make decisions about the forecast. We discussed how moving average methodologies are so simple to apply that you are taking the average of past periods, and that average of past periods is the forecast for the next period.

In the last session of the class, we discussed a more improved method in the form of exponential smoothing methods. We discussed how, on a regular basis, we can improve our forecast and try to reduce the uncertainty in our forecast because whatever uncertainties are happening in the current period, we are trying to incorporate some component of that uncertainty in our forecast. That is what we are going to discuss in this session also, and another very popular method of predictive analytics, which is regression analysis, will also be discussed in this particular class. Now, what we were discussing in the previous session about the exponential smoothing method is that exponential smoothing is simply the smoothing of the fluctuations by applying a particular way of incorporating some of these fluctuations in my base data, and I will continuously improve the base data. I was talking in my last class about how my past data is moving.

Now, here I am today, and I want to forecast for the next period. This is my period t, and I want to forecast for period t plus 1. Now, what we do, as we have already discussed, is that any demand data has two components. Some components we call the level

component because this data also consists of, let us say, this dotted line you see, which is the level data. So, I will determine this level value for the current period, which I call St, and this level value of the current period, I assume, will remain the same for the next period. So, this St for the next period will become my forecast for this next period, and for getting the value of St for the current period.

I have two very important pieces of data. This is t minus 1. So, at t minus 1, what was the base? St minus 1, and what is the demand in the current period Dt? So, the current base St is basically a function of the current demand and the base of the previous period, and this can be represented in this simple equation where this alpha is the smoothing constant.

$$S_t = \alpha(D_t, S_{t-1})$$

The value of alpha varies from 0 to 1, and you may take more common values, which are generally, though alpha is 0 to 1, but common is 0.1 to 0.3. These are generally the common values of alpha which we generally use in our practical purposes.

Now, if you just simplify that, this becomes alpha Dt plus 1 minus alpha into St minus 1.

$$S_t = \alpha D_t + (1-\alpha)\, S_{t-1}$$

So, whatever is your current demand, for example, if my current demand is, let us say, Dt is 100, St minus 1 is 90. So, St minus 1, 90 means for my current period, I forecasted 90 units, but the actual demand came to be 100 units. I want to forecast for the next period. So, what I am going to do, let me take alpha equals to 0.2.

So, with this data, 0.2 into Dt 100 plus 0.8 into 90, that becomes my St, and that is 20 plus 72, 92 is my current base value, and this 92, which is my current base value, will become the forecast for the next period. Now, in this, I have updated my current base from 90 to 92. And how could I do this? Because I realized that there is a gap of 10 units in my forecast and actual demand for the current period. So, I considered some fluctuation that I should take some of these fluctuations. I am not taking the 100 percent fluctuation.

I am taking only 20 percent of the current period's fluctuation for improving my forecast, and 2 units, 20 percent of 10 is 2, so 90 plus 2, 92 is my new base, and that new base is my forecast. Similarly, when I reach this particular period, I will get a new Dt1, and in that iterative manner, I will continuously improve my forecast. So, if I see one practical example, the data which we used in our previous class also, that this is the 4-period data W1, W2, W3, W4 for the month of April, the data is like 38, 35, 77, and 90. Now, exponential smoothing I have to use, alpha is given to me as 0.1. I have to determine the forecast for week 1 of May.

Now, the forecast for week 1 of May is basically as of week 4 of April. So, with this data, I will try to calculate what is the base for the fourth week of April. Whatever is the base for the fourth week of April, that will become my forecast for the first week of May. Now, to start this process in an iterative manner, let me start with the first period of the April month, and for that first period of the April month, let me consider the data is W1, W2, W3, W4. Now, for the first period, the forecast given to me is 60, and that 60 we have calculated by the process of initialization, and that process of initialization is simply the average value of four demands.

So, the average value of four demands is calculated on the screen, and that is coming to be 60.

$$S_0 = (D_1 + D_2 + D_3 + D_4)/4 = (38+35+77+90)/4 = 60$$

So, this is the calculation of the initial base, and this initial base is the forecast for the first period, and now the actual demand for the first period is 38. So, you get F1 minus D1 equals to 22. Now, you see that the formula for our calculation is ST equals to ST minus 1. Now, S0 is given to you, we have to calculate S1, which will be alpha D1 1 minus alpha S naught.

$$S_1 = \alpha D_1 + (1- \alpha)S_0$$

Now, D1 is 3, alpha we have taken 0.9 into 60 this will come as S1. when I calculate S2 it will be alpha d2 plus 1 minus alpha S1.

$$S_2 = \alpha D_2 + (1- \alpha)S_1$$

Now, this S1 will come here d2 is given to us that is 35. Using S1 and D2 you will calculate S3 alpha D3 plus 1 minus alpha S2.

$$S_3 = \alpha D_3 + (1- \alpha)S_2$$

and then you will calculate S4.

$$S_4 = \alpha D_3 + (1- \alpha)S_3$$

Here you will use D3 equals to 77. Alpha D4 plus 1 minus alpha S3 where you will use D4 equals to 90 and this S4 is forecast for the first week of May.

So, in this systematic iterative manner you will calculate the forecast for the period of May and this is how the calculation will take place like 0.1 into 38 plus 0.9 into 60 and as I just explained you can continuously do these calculations.

And finally, you may check your calculation that F5 that is the S4, F5 is basically S4, this will come to be 60.90. After understanding this exponential smoothing method, it is also important to understand that exponential smoothing is possible in a variety of ways. Like for example, we in this particular case used only level, but you can use level and trend also. You can use level, trend, seasonal factor also and there may be fluctuations in trend also, there may be fluctuation in seasonal factor also. So, here you are using only alpha, here you will use alpha and beta and here you will use alpha, beta and gamma also. So, depending upon how many different types of characteristics you are taking in your past data more number of smoothing constants may be required.

All these smoothing constants have 0 to 1 variation, all these smoothing constants have 0 to 1 variation. Before I go further I would like to tell you one more thing that if you see this particular equation alpha Dt plus 1 minus alpha St minus 1.

$$S_t = \alpha D_t + (1- \alpha)S_{t-1}$$

So, generally all other equations will be also in the same way alpha Dt plus 1 minus alpha St minus 1 is the formula for calculating new base. Now, all these smoothing constants are between 0 to 1 whether it is alpha whether it is beta whether it is gamma. Popular values are between 0.1 to 0.3, these are popular values.

These are possible values. Now, between 0 and 1, sometimes you will see that you take extreme values of either 0 or 1. These are the extreme values, either 0 or 1. If you take an extreme value of alpha equal to 0, what happens? ST equals ST minus 1. If alpha equals 1, what happens? ST equals Dt.

Now, what do you infer from these two extreme cases? If alpha equals 0 and alpha equals 1, if alpha equals 0, our new base is equal to the old base. It means we have totally ignored the fluctuations. I do not want to include any of the fluctuations for the estimation of my new base. These fluctuations are for very temporary reasons.

They should not be used for updating the forecast values. So, I want to ignore these fluctuations, and when alpha equals 1, it means I totally ignore my past base. The demand has a new characteristic; there are some permanent changes in the base, and therefore, I am shifting to a new base level that is represented by new demand data. I do not want to include the fluctuations or the data of the past. So, these are the two extreme cases of alpha equals 0 and alpha equals 1.

In one case, I am totally ignoring the present happenings, and in another case, I am accepting only the present things and totally ignoring the past data. So, these are two special cases. Now, after understanding the basic method of exponential smoothing, we move to another very popular predictive analytics tool, which is regression analysis. And, regression analysis, I hope most of you are already aware of how useful it is, not only useful but also how efficient it is. Initial efforts are required in developing the regression analysis because it is something like having a dependent variable that depends on some number of independent variables.

So, my first effort is the dependent variable, which I always know. Known means I know what I want to determine. Known means the value is not known, but I know what is to be determined; that is the meaning of known here. But, this factor Y, which I want to determine, on what factors is it depending? What can be X1, X2, X3, and Xn? This requires my understanding as well. So, identifying the right factors, independent factors, which may affect the value of Y, is the important challenge. That is number one. Second,

developing a proper mathematical relation between this dependent variable and independent variables is another challenge.

So, that these independent variables, with that proper mathematical expression, can explain the variation in the dependent variable. So, these are two very important things which are required for regression analysis. Now, in this particular session, I am talking about a special type of very simple case of regression method, which is the simple linear regression model, which has only one independent variable. This is the independent variable, this is the dependent variable, and since it is a linear function, the relation between x and y is a linear function, y equals mx plus c, something of this type. Where you have M and C. M is a kind of parameter for x, and C is also a kind of constant.

$$\mathbf{Y = MX + C}$$

With the help of M and C, you are able to develop a relationship between x and Y. So, whenever we are trying to develop a relation between Y and x, our job is to identify and calculate the value of M and C, so that a specific relation between Y and X can be established. However, there may be more complex relations between Y and X, but we are beginning our discussions, so I am considering only a simple linear regression model. In some cases, let us say Y equals X squared, this is another possibility, Y equals the square root of X, Y equals sin X, all these are other possible relations. And there may be many more such relations. It is not simply that these three relations are complete.

You can create infinite types of relations between y and x. But my interest is in this linear relation. However, it is quite possible that the value of this constant C may be 0 also in some cases. And if it is 0, then in that case, it may be simply y equals mx. So, it is also a linear relation. So, linear relation may also take some forms; it may be y equals minus mx, which is also acceptable. But there cannot be a power to the x variable; that is the meaning of linearity.

You can see on your screen some of the relations between y and x. Here, these pictures are good enough to tell you that because y equals mx plus c, we are developing this relationship. And in this relationship, if y is like this, then it means as x is moving, y is also moving in a similar direction; that is a strong relationship. But, if you see the right-

hand side images here, Y is quite scattered, and when it is scattered, it is not very sure because when X is moving, let us say here one possibility of Y is this, another possibility is this, and another possibility is this also. So, here in this case, the relationship which we will be developing, y equals mx plus c, is not going to give you a very good, you can say, estimation or prediction of y because for one value of x, there are different possible values of y. But my relation will give you only one value, and y may take any other value also. So, therefore, our regression analysis will have low predictability here and very high predictions there.

If you have a strong relationship, and sometimes these strong and weak relationships are possible because of the wrong identification of X, whether the correct X is taken or not, on that also these strong and weak relationships may vary.

This is a very generic, you can say, expression for representing your regression equation.

$$Y_i = \beta_0 + \beta_1 X_{i+} \varepsilon_i$$

Y is the dependent variable, as we have already discussed. This is because the regression equation is like this. This is x, this is y, and this is that line which is the line of regression.

Now, this line is intercepting the y-axis here. This interception of the y-axis is beta naught, or which I was saying earlier as c also. Then there is a coefficient of the slope of x, and then there is a random error term also. If you see on this particular diagram, it is very well articulated: x-axis, y-axis, this is the regression line, and on this regression line, if you see, this is the intercept on the y-axis, which is B naught, this intercept. The slope of this line can actually be seen from here, that is beta 1, which is the coefficient for the independent variable Xi, and then the random error term, the meaning of this random error is that for any value of X, Xi.

You are estimating Y here; this is the estimated value, the predicted value of Y. But the observed value of Y may be different, like here we have shown, but it can be here also; the predicted value will always be on the line, the predicted value of the variable will always be on this line of regression. It is possible that the observed value of Y may also be on the same line, or the observed value may be at the height or at the lower place of

the line also. So, the difference between the predicted value and the observed value is basically your error term, the random error at a particular value of X. So, that is how in this diagram you are able to understand the meaning of this linear regression equation. Generally, for our calculation purpose, we are using this equation: this is Y plus B0 plus B1 X1, which is the case of the simple linear regression model.

$$\mathbf{Y_i = b_0 + b_1 X_i}$$

And as we discussed already, whenever we are going to use any kind of model, we will be using that model for estimating the values of B0 and B1.

And here, if you see, we can use Excel very conveniently because it is easy to use. But you can always find so many other software options also which are dedicated to this kind of predictive analytics. You can use MATLAB, you can do programming in Python also, or you can use SPSS. All these are the different types of software and solutions which are available for getting you a regression equation. There may be some minor differences for the same data if you use different computing solutions. There may be some minor differences because of the rounding issues in the calculations for B0 and B1.

For example, if I take this particular case where a particular real estate agent wishes to examine the relationship between the selling price of a home and its size, generally, we expect that if the size of a particular house is bigger, the selling price will also be bigger. Now, I want to develop the relation that the price of a house is a function of the area of the house. Let me say house price is a function of the area of the house. This is what relationship I think is there. Now, whether the price of a house is double the area, triple the area, four times the area, or whatever is the relation that we want to see in this particular case.

Now, we have taken a sample of 10 houses, and we know from that sample of 10 houses the price and area. House numbers 1 to 10, their prices are given to us, and areas are also available to us. Then, with the help of area on the x-axis and price on the y-axis, if the area is this much. This much area, what is the price? Area is this much, what is the price? Area is this much, what is the price?

For the same area, it is possible that two prices may be available because you can easily understand that house prices may also be very much dependent on the location. Not only the area but also the location where it is. House prices may depend on various other factors as well. How close is your house location to the market? So, these are the different factors which may affect the house price.

So, based on those things, we may go for a more complex regression equation, but since in this example we are considering only one variable, which is the area of the house. And using this data of 10 houses: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Now, this is the independent variable, this is the dependent variable, and we are looking to develop this relationship. Now, there are ready-made software available, and this is the scatter plot you are going to get when you put this data in any Excel table, and you can generate this scatter plot from the data. Now, if you see, the regression equation will be something like this, which is also known as the line of least squares. The line of least squares.

Because this line, the line of least squares, means this line is the best representative of this data. This line is expressed as b naught plus b1 x1. So, this line best represents this data set, and that is our regression equation as well. And for that purpose, if you see, we can do this in Excel as well, where we have put this data in the table, and you can see the screenshot of the Excel on the screen, which is all the table data I was presenting here. We have punched that data into Excel, and then in the data itself, you can invoke this regression option, and when you invoke the regression option.

You can see that one field requires the Y range and another field requires the X range. The Y range means the dependent variable, the X range means the independent variable, and therefore, we have given the appropriate address for both these ranges: A1 to A11 and B1 to B11 for these two things. We have also clicked the label because A1 and B1, here you see A1 and B1, these two fields contain the labels of the data: house price and square feet, respectively. And then we have clicked 'OK,' and then we can see the regression outcome. The regression outcome that comes is like this type of outcome in the Excel table, which is easily readable.

Now, there are many things that are available in this table. I will not go into all of them at this moment. I am only interested in this outcome where you get the value of B0 and this is the value of B1. B0 is, let us say, 98. Let us not make it very complicated.

So, you can say y equals 98 plus 0.10 multiplied by x. So, that is the simple equation you are getting where the value of B0 is 98 and the value of B1 is 0.10. Now, you can put into this equation the area of any new house, and you can estimate the possible price for this particular house. That is the predictive analytics that you can predict: if the area of the house is this much in the city, what can be the possible price for that house? We may discuss other things that are available in the outcome of this regression analysis. But since we have not gone into that aspect, it is not advisable to have that discussion at the moment. So, we are leaving that remaining discussion, and I hope with this, we have understood the basics of predictive analytics, particularly in the use of time series analysis.

In the use of moving averages and exponential smoothing, as well as in the use of regression analysis. With this, we come to the end of this particular session. Thank you very much.