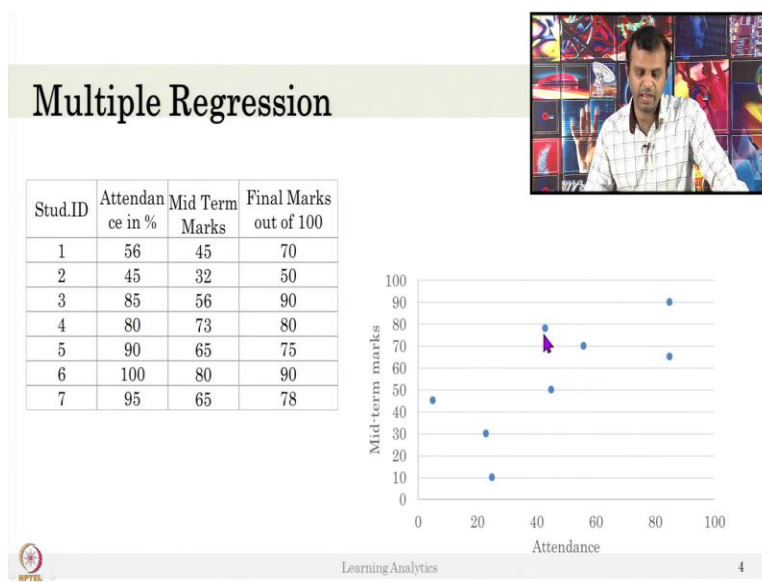


Learning Analytics Tools
Professor Ramkumar Rajendran
Department of Educational Technology
Indian Institute of Technology, Bombay

Lecture 8.3
Multiple Regression

In this video we will discuss what is multiple regression. So, again the same regression can be classified into simple or multiple. We talked about multiple regression, it has one dependent variable and multiple independent variable. So, which means there are two independent variable and dependent variable that is the final marks. This data we have seen previously multiple times, you know examples, so I am using same example so that you know how the same example can be used for different models, how different models predicts the score differently and you can compare that also.

(Refer Slide Time: 01:11)




We have already plotted attendance and midterm and want to predict by what will be the final marks out of 100? We cannot show it in this in a two-dimensional plot. The data and the plot is not matching but consider this is the plot, we cannot show the final marks in it, maybe we can do for each data say there is a data of say 45 percentage or 40 percentage and 80 percentage midterm marks, you might be able to plot what will be the mark something like that maybe a 70 or 80 you can write the marks on top of it, that might be good for classification algorithm not for educational algorithm, let us see how it be used for regression algorithm.

(Refer Slide Time: 01:46)

Multiple Regression Ypred

- $Y_{pred} = 31.45 + 0.404X_1 + 0.22X_2$
- X_1 – Attendance in %
- X_2 – Mid term marks
- $R = 0.85, R^2 = 0.72$
- $R^2 = 0$ to 1 , 0 means no relationship, 1 is perfect linear fit

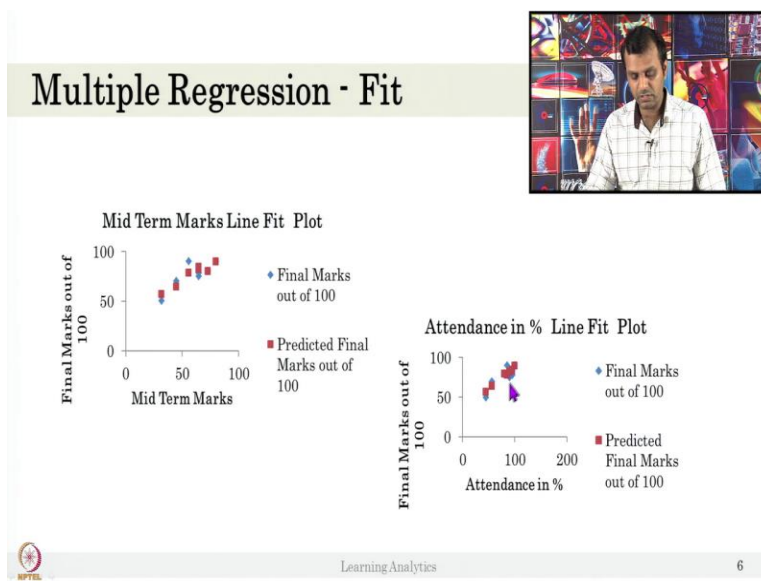


Learning Analytics 5

Y predict is equal to the 31.45 plus 0.4X1 and 0.22X2 that is the linear regression model for the data we have, X1 is attendance is in percentage and midterm marks is X2. And R equal to 0.85, R square is equal to 0.72. If you remember the correlation coefficient R that is what computed here the R square is this.

But it is not the correlation coefficient like we did with one variable for X and one Y instead this is with the two variables, so that is called regression coefficient. Let us see that. If it is 0, it means no relationship between the variables X1 and X2 to Y, if this near took 1 or 1 is a perfect relationship perfect linear fit between X1 and X2 to Y, near to 1 is good, so it is good.

(Refer Slide Time: 02:52)



So, I plotted this linear regression in excel sheet using data analysis tool and here it considers only the midterm marks to predict the final marks score, final marks in the out of 100. You know this is a predictive value, this is the actual value, the fit is kind of okay. If we consider only the attendance the fit is again good, it's not very bad, so individually each variable is doing good and we have seen that this variable as high correlation, we discussed that in one of our correlation matrix video, we know about that. So, this is good. So, these two variables are correlated and they also fit best with the final marks out of a 100.


(Refer Slide Time: 03:44)


Predicted Value

- $Y_{pred} = 31.45 + 0.404X_1 + 0.22X_2$

Stud.ID	Attendance in %	Mid Term Marks	Final Marks out of 100
1	56	45	70
2	45	32	50
3	85	56	90
4	80	73	80
5	90	65	75
6	100	80	90
7	95	65	78

$$Y_{pred_1} = 31.45 + 0.4 \cdot 56 + 0.22 \cdot 45$$
$$= 63.75$$



Learning Analytics7

So, what that means in multiple regression. Suppose you have this particular prediction value that is the linear model learn from this 7 data and you want to predict the future marks, you can use this model to predict it. Let us, see what is this predictive model do for the given data that is on the training set itself, for predicting a student one data I would add 31.45 plus 0.4 into X1 is 56 and the 0.22 into X2 0.22 into X2 is 45, 0.22 into X2 is 45, if you compute this the marks will be 63.75, the final mark is the actual mark is 70, the predicate mark is 63.75.

So the error is 63.75 minus 70 into square that is a error. And if you compute the other like that for all these values and a weight of that year is the least mean square error value. So, then cut the best fit model. So, this is how it is computed.

(Refer Slide Time: 04:53)


Predicted Value

- $Y_{pred} = 31.45 + 0.404X_1 + 0.22X_2$

Stud.ID	Attendance in %	Mid Term Marks	Final Marks out of 100
1	56	45	70
2	45	32	50
3	85	56	90
4	80	73	80
5	90	65	75
6	100	80	90
7	95	65	78

Observation	Predicted Marks
1	63.84
2	56.57
3	77.94
4	79.60
5	81.91
6	89.21
7	83.93

Att 75%, mid 50
 $y_{pred} = 31.45 + 0.40 \times 75 + 0.22 \times 50$

Learning Analytics8

If I compute that for all the models their answers like this, so this value is bit different, you know 63.7584, it is because I ignored the values beyond the two decimal point. So, this says that this is best fit for some of the data for examples student number 4 it is good and also student number 6 is good, but others it is not so so great or so perfect, but that is why and this is the model we get it, best fit model.

Now, using this model if it is very interesting that if you want to predict some student, you have no idea what the student is the student have say 75 percentage attendance also he scored midterm marks say 50, something like that, so what will be that students performance?

So, consider you obtained the data in the table and you got it from historical data by last year last two years data kind of like a 100 data you have or 200 students data. You created this model and this is the student one in a current semester, so if the student one, this is a student one in a current semester, yes attendance and midterm marks, now you want to predict what will be the students final score in the exam.

So, you apply this model. So, the students final score in exam, $Y_{predict}$ equal to 31.45 plus 0.40 into 75 percentage that is you can use 75 as it is and plus and 0.22 into 50 midterm marks, why I asked you to use as it is maybe that system would have used as it is, so I used actually the same mark, I did not put the convert that into a like 75 points something something like that, anyway.

So, this is the mark, if you compute this, this is a mark the student want to get it. You know in a current situation we may see that lot of institutions, campuses or universities are cancelling the end sem exams the better way to do it let us take the students the current students who are in the fourth year, third year or the previous historical data on particular course.

Take all this data, and compute is there any fit by computing the midterm marks or term one marks and attendance or any other variables you can collect. Create a model and make that model very accurate I precisely doing it, then apply that model to predict the score that will be a better way to do it.

However, given the current scenario things are not that way, it was just based on 80 percent marks in last year's score or some university goes by midterm marks, some value or the student's performance still mid sem, student's performance till mid sem is still good you know, why? Because that actually proved if you compute the last five years data or something the mid sem mark is actually highly correlated with end sem marks.



So, but this should be the right method to do, if you are interested, if your data access to the data you can go and take the date of last 5 years students data and compute the current students data and see how it works. Then you may not have the actual values, that is fine, but you can predict it as much as good.


(Refer Slide Time: 08:48)

Activity

Multiple Regression

- $Y_{pred} = 31.45 + 0.404X_1 + 0.22X_2$
- What is the significance of 31.45? What is intercept!



 Learning Analytics 9

It is interesting that we saw in simple linear regression and also in multiple regression, this intercept like 31.45 in this equation or the some other values in the previous equation, what is this intercept means? What is significance of this 31.45? What is this intercept? Take a moment, you know this intercept is the line extended towards the 0 index value, that is the where it is crossing in y scale, that is a value good, but what is that mean? Take a moment, think about it write down the answers after write it down resume the video to continue.

(Refer Slide Time: 09:28)


Activity


Multiple Regression

- With no values is X_1 and X_2 , Y will predict intercept score
- Mean value of Y with all independent is set to zero
It can be positive or negative

Don't try to interpret the meaning, if the X values can't be zero.

$$y = c + x_1 w_1 + x_2 w_2$$



 Learning Analytics 10

So, intercept means with no values of X_1 and X_2 , Y will predict the intercepts score. The Y actually gives the if X_1 is 0, X_2 is equal to 0. It is value of Y with all independent value set to 0.

And it can be positive or negative, that is very very important, so it is very very important. Do not try to interpret the meaning of intercept that is not correct, this intercept is very very important for linear regression to create a best fit model, but there is no meaning to it, do not try to identify the meaning of the intercept, some model with a better intercept, no not really true.

In educational settings if the student is not attending the class, also he got 0 mark in the mid sem without attending the class, whatever the final score? So, you should may not even allow the student to write exam, the final score be 0, but the model might say 34 marks, it is not correct, if we allow the student to do that you might get the 31 marks we do not know or some cases, some particular examples, there is a possibility that a person may not even come to work at all and person may not able to perform the duty still he might get a some basic minimum wage, as far as the norms or something like that.

That kind of that is a minimum wage anyone can get it, even if you work do not work that can be a intercept, but do not try to interpret the meaning of this intercept that is the idea. So, maybe if the X values cannot be 0, like mid sems cannot be 0 then do not even try to interpret. So, in educating setting, do not do try to do that.

Most important thing in linear regression or simple regression or multiples regression is, the most important thing is Y say C plus $X_1 * W_1$, $X_2 * W_2$, so now you thought I said do not interpret the meaning of C , what is W_1 and what is W_2 ? What is the significance of W_1 and W_2 ? Keeping the other value constant consider the X_2 is same, keeping the other value constant, how much the X_1 as relationship with Y is defined in the $Y W_1$?

I said in the last video what is the advantage of using the linear regression or regression model is to create the indicator of each variable with the Y that is called the indicator. If you keep all other variable values constant, except that variable X_1 all other equals X_2 X_3 X_4 all other variables keep constant this particular relationship or relationship between X_1 and Y is given in this weight.

Similarly, if we keep all other things constant except W_2 , relationship between X_2 and Y is given in the weight W_2 . That is how the relation varies, if it varies by 1, this will vary by say if it

is 0.8, if it is 100 this will be 80, something like that. The you can interpret the meaning of the weight is not C, this indicates how strongly this particular variable is correlated with Y, keeping all other variables constant, that is the linear regression. Thank you.