Learning Analytics Tools
Professor Ramkumar Rajendran
Department of Educational Technology
Indian Institute of Technology, Bombay
Lecture 44
Linear Regression - Example

In this video, we will see the example of how this linear regression and logistic regression is used.

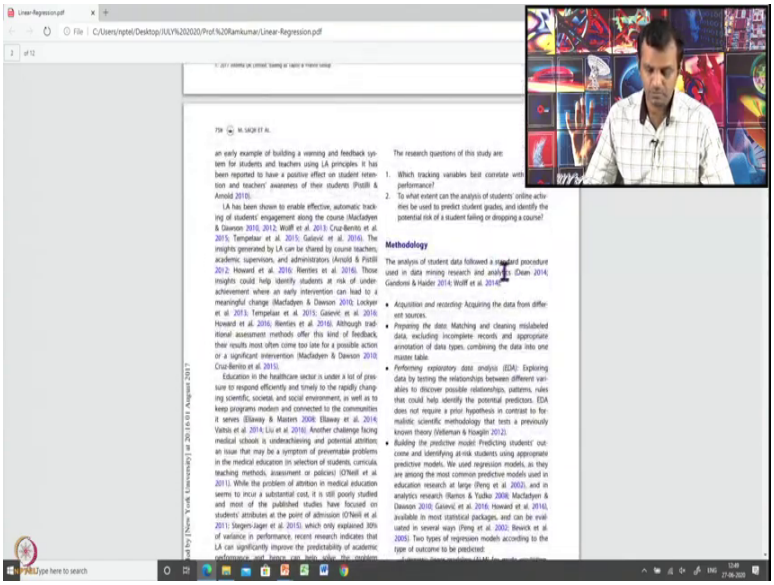
(Refer Slide Time: 0:26)

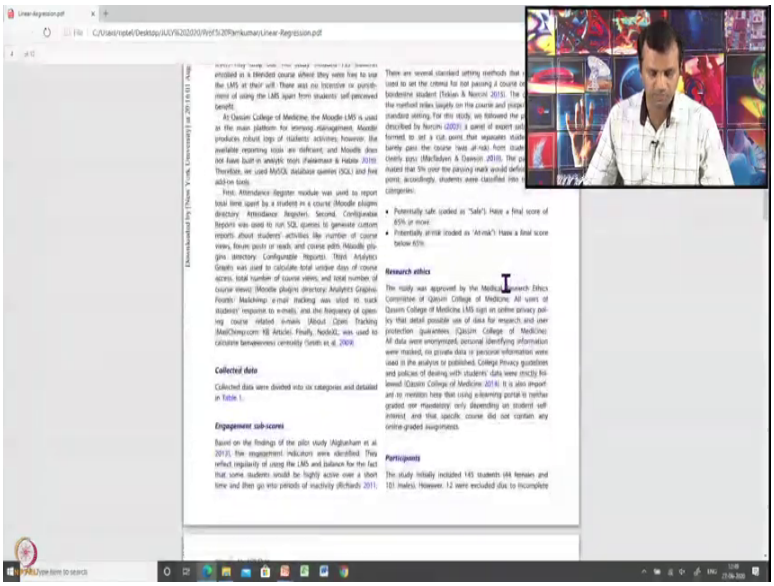
So let us look at this paper how learning analytics can early predict underachieving students in blended medical education course. It is hard to find a paper that explains linear regression in much detail in 2017 or recently because linear regression has been used in the education field for a long time.

But since the authors might be using it for the first time and it is in the medical education course, they explained it in detail. So this is a good paper to look at it, to understand our linear regression. But I do not recommend you writing a paper in such detail because

everyone now knows what is linear regression and they are looking for the response or the metrics, not the intercept value, each weight and other such things.

(Refer Slide Time: 1:09)





enrolled in a blended course where they were free to use the LMS at their will. There was no incentive or punishment of using the LMS apart from students' self-perceived benefit.

At Qassim College of Medicine, the Moodle LMS is used as the main platform for learning management. Moodle produces robust logs of students' activities; however, the available reporting tools are deficient, and Moodle does not have built-in analytic tools (Fakeerawe & Hatoto 2016). Therefore, we used MySQL database queries (SQL) and free add-on tools.

First, Attendance Register module was used to report total time spent by a student in a course (Moodle plugins directory: Attendance Register). Second, Configurable Reports was used to run SQL queries to generate custom reports about students' activities like number of course views, forum posts or reads, and course edits (Moodle plugins directory: Configurable Reports). Third, Analytics Graphs was used to calculate total unique days of course access, total number of course views, and total number of course views (Moodle plugins directory: Analytics Graphs). Fourth, Mailchimp e-mail tracking was used to track students' response to e-mails, and the frequency of opening course related e-mails (About Open Tracking MailChimp.com KB Article). Finally, NodeXL was used to calculate betweenness centrality (Smith et al. 2009).

Collected data

Collected data were divided into six categories and detailed in Table 1.

Engagement sub-scores

Based on the findings of the pilot study (Alghanham et al. 2013), five engagement indicators were identified. They reflect regularity of using the LMS and balance for the fact that some students would be highly active over a short time and then go into periods of inactivity (Richards 2011).

There are several standard setting methods that [illegible] used to set the criteria for not passing a course or borderline student (Tolsgaard & Norcini 2015). The [illegible] the method relies largely on the course and program standard setting. For this study, we followed the [illegible] described by Norcini (2003): a panel of expert [illegible] formed to set a cut point that separates students barely pass the course (was at risk) from students clearly pass (Mladenovic & Ivanovic 2010). The [illegible] stated that 5% over the passing mark would define [illegible] point; accordingly, students were classified into [illegible] categories:

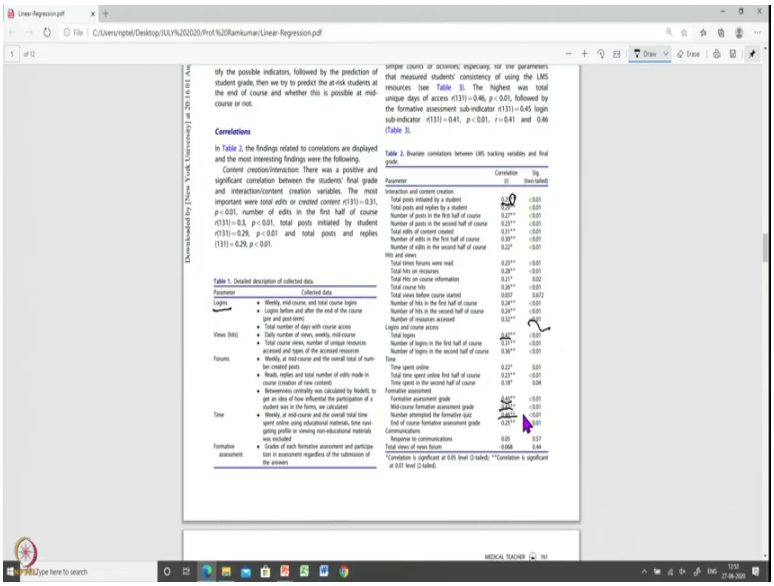
- Potentially safe (coded as "Safe"): Have a final score of 85% or more.
- Potentially at-risk (coded as "At-risk"): Have a final score below 85%.

Research ethics

This study was approved by the Medical Research Ethics Committee of Qassim College of Medicine. All users of Qassim College of Medicine LMS sign an online privacy policy that detail possible use of data for research and user protection guarantees (Qassim College of Medicine). All data were anonymized, personal identifying information were masked, no private data or personal information were used in the analysis or published. College Privacy guidelines and policies of dealing with students' data were strictly followed (Qassim College of Medicine 2014). It is also important to mention here that using e-learning portal is neither graded nor mandatory; only depending on student self-interest, and that specific course did not contain any online-graded assignments.

Participants

The study initially included 145 students (44 females and 101 males). However, 12 were excluded due to incomplete



tify the possible indicators, followed by the prediction of student grade, then we try to predict the at-risk students at the end of course and whether this is possible at mid-course or not.

Correlations

In Table 2, the findings related to correlations are displayed and the most interesting findings were the following.

Content creation/interaction: There was a positive and significant correlation between the students' final grade and interaction/content creation variables. The most important were *total edits* or *created content* $r(131) = 0.31$, $p < 0.01$, number of edits in the first half of course $r(131) = 0.3$, $p < 0.01$, total posts initiated by student $r(131) = 0.29$, $p < 0.01$ and total posts and replies $(131) = 0.29$, $p < 0.01$.

Table 1. Detailed description of collected data.

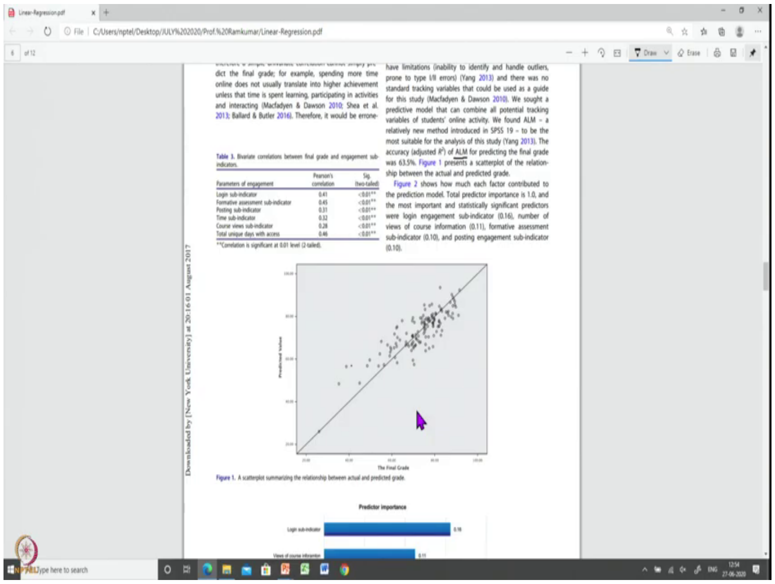
Parameter	Collected data
Logins	• Weekly, mid-course, and total course logins • Logins before and after the end of the course (pre and post-term) • Total number of days with course access
Views (hits)	• Daily number of views, weekly, mid-course • Total course views, number of unique resources accessed and types of the accessed resources
Forums	• Weekly, at mid-course and the overall total of number created posts • Reads, replies and total number of edits made in course (creation of new content) • Betweenness centrality was calculated by NodeXL to get an idea of how influential the participation of a student was in the forms, we calculated
Time	• Weekly, at mid-course and the overall total time spent online using educational materials, time navigating profile or viewing non-educational materials was excluded
Formative assessment	• Grades of each formative assessment and participation in assessment regardless of the submission of the answers

simple courses or activities; especially, for the parameters that measured students' consistency of using the LMS resources (see Table 3). The highest was total unique days of access $r(131) = 0.46$, $p < 0.01$, followed by the formative assessment sub-indicator $r(131) = 0.45$ login sub-indicator $r(131) = 0.41$, $p < 0.01$, $r = 0.41$ and 0.46 (Table 3).

Table 2. Bivariate correlations between LMS tracking variables and final grade.

Parameter	Correlation (r)	Sig. (two-tailed)
Interaction and content creation		
Total posts initiated by a student	0.29**	<0.01
Total posts and replies by a student	0.29**	<0.01
Number of posts in the first half of course	0.27**	<0.01
Number of posts in the second half of course	0.25**	<0.01
Total edits of content created	0.31**	<0.01
Number of edits in the first half of course	0.30**	<0.01
Number of edits in the second half of course	0.22*	<0.01
Hits and views		
Total times forums were read	0.25**	<0.01
Total hits on resources	0.28**	<0.01
Total Hits on course information	0.21*	0.02
Total course hits	0.26**	<0.01
Total views before course started	0.057	0.672
Number of hits in the first half of course	0.24**	<0.01
Number of hits in the second half of course	0.24**	<0.01
Number of resources accessed	0.32**	<0.01
Logins and course access		
Total logins	0.45**	<0.01
Number of logins in the first half of course	0.35**	<0.01
Number of logins in the second half of course	0.36**	<0.01
Time		
Time spent online	0.22*	0.01
Total time spent online first half of course	0.23**	<0.01
Time spent in the second half of course	0.18*	0.04
Formative assessment		
Formative assessment grade	0.45**	<0.01
Mid-course formative assessment grade	0.46**	<0.01
Number attempted the formative quiz	0.41**	<0.01
End of course formative assessment grade	0.25**	<0.01
Communications		
Response to communications	0.05	0.57
Total views of news forum	−0.068	0.44

*Correlation is significant at 0.05 level (2-tailed); **Correlation is significant at 0.01 level (2-tailed).



dict the final grade; for example, spending more time online does not usually translate into higher achievement unless that time is spent learning, participating in activities and interacting (Macfadyen & Dawson 2010; Shea et al. 2013; Ballard & Butler 2016). Therefore, it would be errone-

have limitations (inability to identify and handle outliers, prone to type I/II errors) (Yang 2013) and there was no standard tracking variables that could be used as a guide for this study (Macfadyen & Dawson 2010). We sought a predictive model that can combine all potential tracking variables of students' online activity. We found ALM – a relatively new method introduced in SPSS 19 – to be the most suitable for the analysis of this study (Yang 2013). The accuracy (adjusted R^2) of ALM for predicting the final grade was 63.5%. Figure 1 presents a scatterplot of the relationship between the actual and predicted grade.

Figure 2 shows how much each factor contributed to the prediction model. Total predictor importance is 1.0, and the most important and statistically significant predictors were login engagement sub-indicator (0.16), number of views of course information (0.11), formative assessment sub-indicator (0.10), and posting engagement sub-indicator (0.10).

Table 3. Bivariate correlations between final grade and engagement sub-indicators.

Parameters of engagement	Pearson's correlation	Sig. (two-tailed)
Login sub-indicator	0.41	<0.01**
Formative assessment sub-indicator	0.45	<0.01**
Posting sub-indicator	0.31	<0.01**
Time sub-indicator	0.32	<0.01**
Course views sub-indicator	0.28	<0.01**
Total unique days with access	0.46	<0.01**

**Correlation is significant at 0.01 level (2-tailed).

Figure 1. A scatterplot summarizing the relationship between actual and predicted grade.

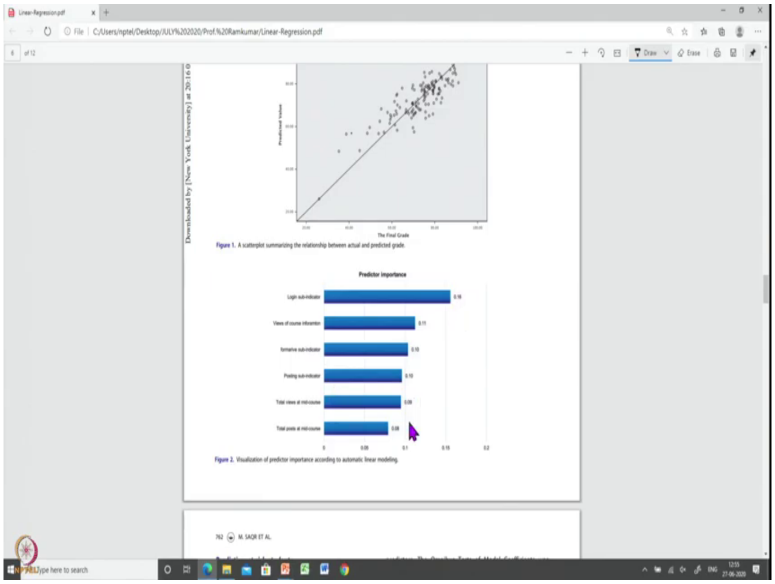
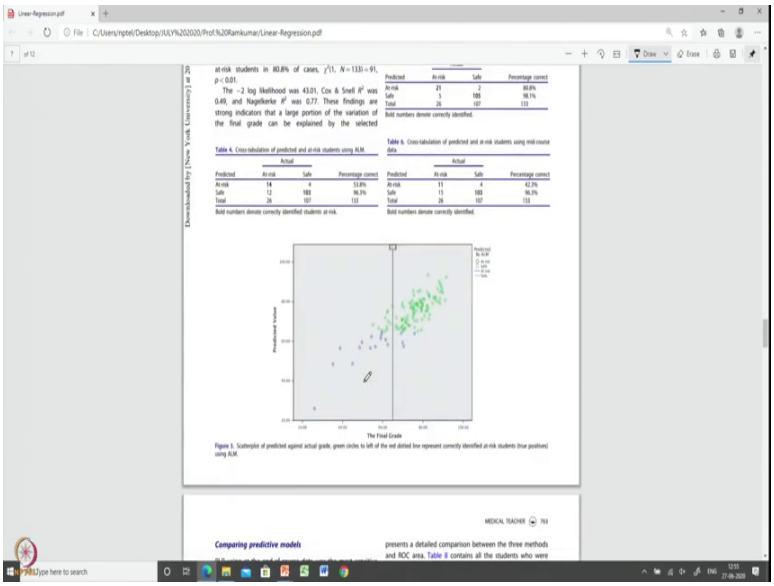
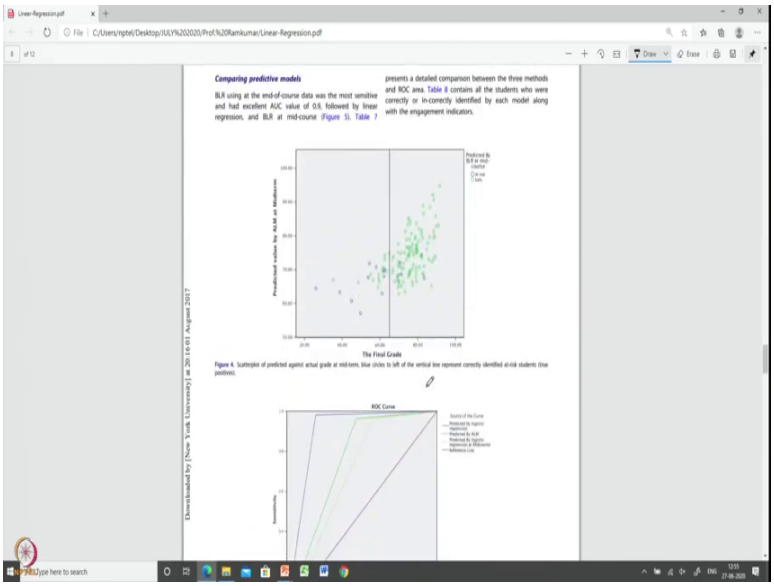


Figure 1. A scatterplot summarizing the relationship between actual and predicted grade.

Figure 2. Visualization of predictor importance according to automatic linear modeling.





But this paper is explaining in detail. Let us look at this paper. The paper shows how learning analytics can early predict underachieving students in a blended medical education course. So in this paper, they want to predict that students who are at risk i.e. were going to score below 65 per cent in the final eaam. So potentially safe students who have finally scored more than 65 per cent will be considered as pass students.

How to predict the students were going to get less than 65 marks in the final score, that are at risk. So there are 145 students, however, they have excluded many, so they have

only one category of students over the period of six weeks. So what is the data they are collecting, the data they collected is in the blended learning approach, that is, they use MOOC and classroom. The data they collected is log in like weekly, mid-course, total course logins.

How many times he logs in a week, in mid-course how many times he logs in before the mid sem, how many time total login times and login before and after the end course of the exam. Also views, the number of views daily, weekly and mid sem till mid-course or also total course views, number of unique resource accessed.

A number of unique resource means how many resources he looked at it? How many papers, how many videos he watched everything and the data type of this resource he accessed. All the forums, like number of posts created, read or replace, number of edits made in the post, the number of likes and also how influential a particular post he created is, based on the number of people who reply to that post.

And the time he spent on each of these sessions, like weekly meets and over all the time on this particular learning environment also the grades at each formative assessment and participation assessment like assignment, regardless of submission of the answers whether the students submit the answer or not. If he participated in the quiz or answers the pop-up questions in the video, if he answers those questions, is also considered as a formative assessment, not the answer score.

So they took all this and multiple features from this like if you see the login can be classified into total login, number of login in the first half and second half. They computed the correlation with the final score. So this you can compute using the correlation matrix.

So you know how to compute it now or they compute each one individually and also they identify whether this has significant value or not because there are 132 students we know whether this correlation is significant or not, significant is not telling the strength of the correlation, significant tells that whether this correlation is reliable or not.

Significant does not mean this correlation is high or not that is a very important thing we have to understand. So what they do, they see this 0.29, the two stars(asterisk) significant by 0.01, 0.01 level, yes here. So all of them are like average correlation. Not really good maybe this is a good correlation number of times you log in and the assessment is a very good correlation with the final scores.

Yes, as expected assessment is really highly correlated with the final score based on the performance in the quizzes or mid-term assessment grade and formative grades. So what they did, they combined these feature values again this split given to us, six variables. Let us look at it. They made it into six parameters of engagement. See the login indicator so formative assessment, posting time, course views and total uniqueness.

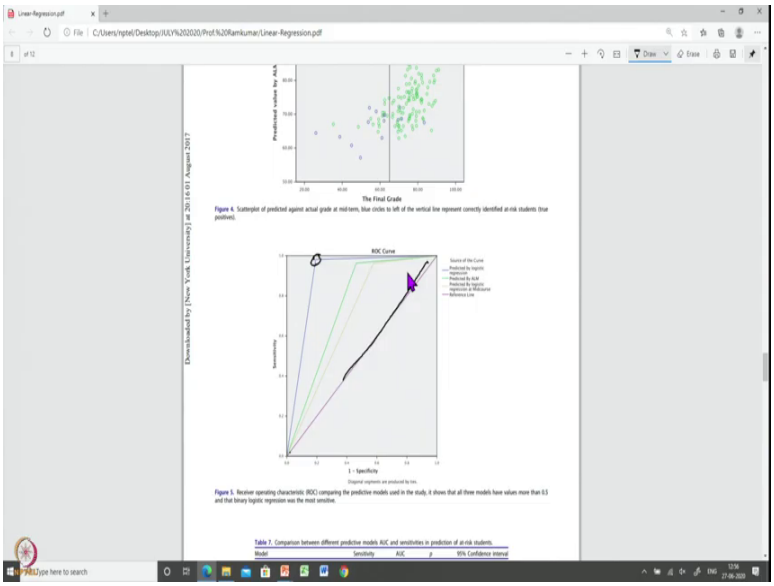
They made into six variables $x_1, x_2, x_3, x_4, x_5, x_6$. They used SPSS software for automated linear modelling that they call this ALM here. But SPSS is proprietary software I do not recommend using that. But if you have access to SPSS 19, please go ahead and use it. But they use SPSS and they are reporting the values here. Let us look at the values. So the predicted final grade is here and the actual final grade is given in this given in these dots but the regression scale is like this.

So this is the difference between the final grade and the predicted value. So also, they are giving the weight of each variable. But as I mentioned, this is the paper that explains every detail about linear regression, which is not needed for the current setting because

everyone knows what is LR. So but yeah see the weight 0.16, 0.1 and formative assessment is 0.10.

So there is a like a weight of each of these variables and the intercept is not given. And that is a problem you know, the intercept is not given. We do not know what is the value but we do not want to interpret the intercept. So we want to interpret only this value. So login is the most strong indicator of the performance and the answer.

(Refer Slide Time: 7:07)



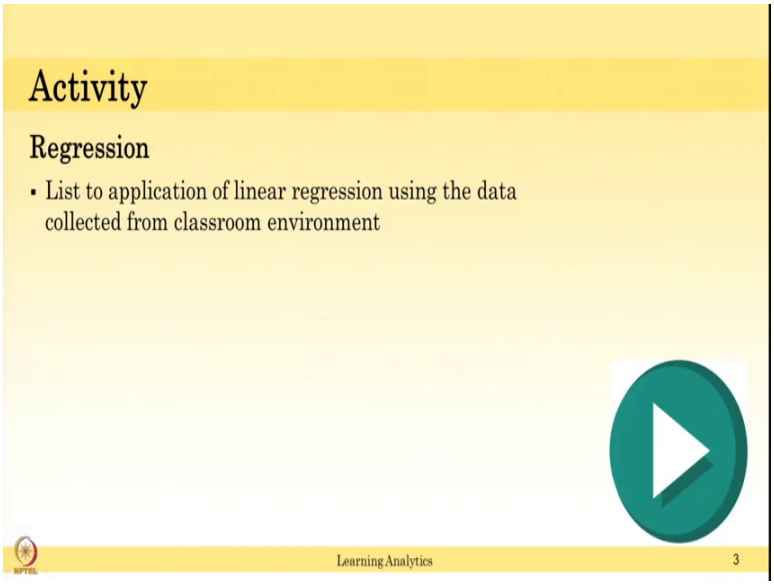
So they computed the students' performance at risk, and they also used the logistic regression to do that, and they also plotted ROC curve, predicted using-

1. logistic regression
2. a linear model
3. logistic regression at the mid-course

Mid-course means that we dont have to wait till the end rather we use data till mid sem to predict their final exam performance.

If you use the mid sem, so you know you will see this. I hope you know what is this line means. This line indicates that the area under the curve is 0.5 anything below this is not good. This particular curve indicates case 3 (i.e. using logistic regression at the mid-course). The graph shows that predicting using logistic regression is the better.

(Refer Slide Time: 8:21)

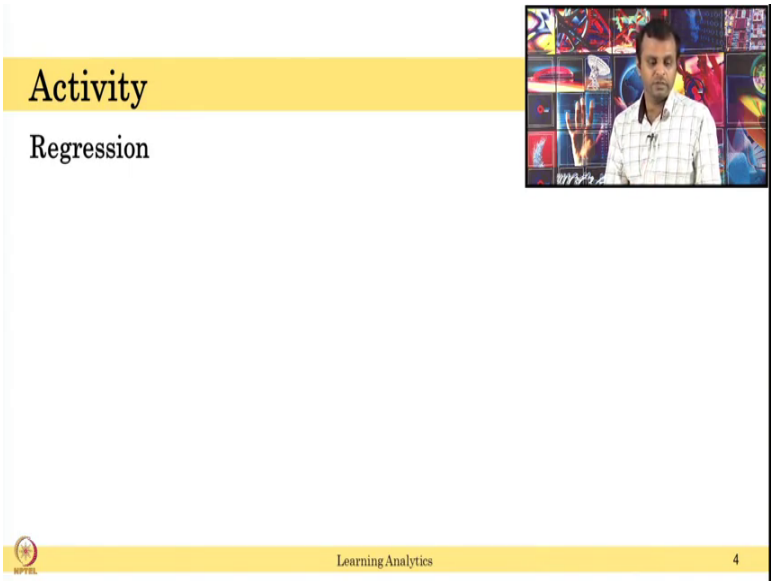


This paper discusses what we learned in general about our ROC curve, the area under the curve and also logistic regression introduction. Please go and read this paper understand how to collect data, how to collect features and how to use those features to predict the final score. So you saw, what is linear regression, also the logistic regression, can you list down one or two applications of linear regression using data collected from classroom environment?

The paper we just discussed now is applying linear regression on blended learning that is, the student is interacting with the MOOC kind of environment where they have to log in, watch videos, read things, post in discussion forums in a Moodle something like that.

Can you think of the application of using linear regression in a classroom environment and which data you can collect?

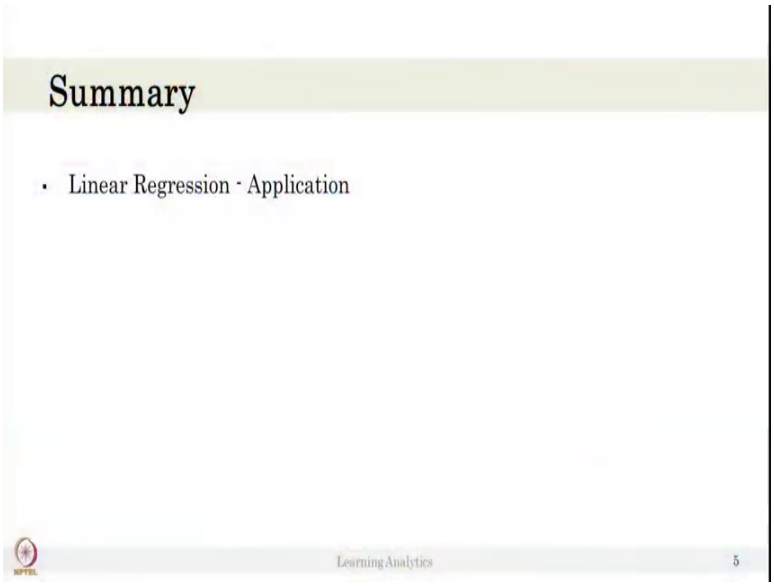
(Refer Slide Time: 9:17)



So list down your answers and resume to continue. So I am not writing any answers here because there are a lot of things that can be predicted I also discussed that debate at the starting of this week's course. You can predict students' performance, student's engagement. If you listed down something else, that is also good.

If you can access this data, you can go and collect the data, please go ahead and collect data and see which one works.

(Refer Slide Time: 10:02)



In this week we discussed only linear and logistic regression and not in detail. This week is kind of less on new learning but I request you to go and explore the tools demo to you and use linear regression, logistic regression, collect data and use the existing data. Go and check Internet for the data, use the data and try to apply and understand. Thank you.