Regression Modeling on EDU-DATA in Technical Education System(TES)

Prof. Dr. P. K. Srimani[#1] **, Mrs. Malini M. Patil**[#2]

#1Former Director, R&D Division, Bangalore University, DSI, Bangalore, Karnataka, India ,
#2Assistant Professor, Dept. of ISE, JSSATE, Bangalore, Karnataka, India.
Research Scholar, Bharthiyaar University, Coimbatore, Tamilnadu, India.

ABSTRACT

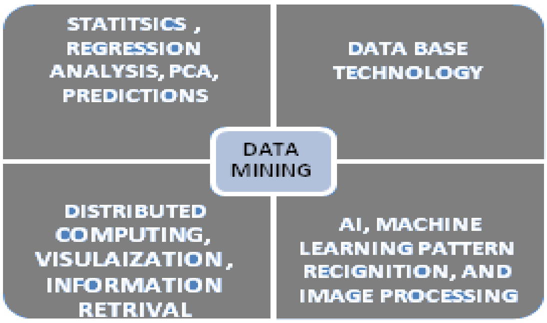
Mining Educational data(Edu-Data)is an emerging inter-disciplinary research area that mainly deals with the development of methods to explore the data stored in educational institutions. The technique of mining Edu-data is referred as Edu-mining. Queries related to Edu-Data are of practical interest. Data mining is concerned with the analysis of data and the use of the software techniques which are responsible for finding the patterns It is a confluence of many disciplines, in which visualization and statistics are major areas and are addressed in this paper. The paper aims at developing a regression model for Edu-Mining using the statistical approach. The statistical results obtained helps the management to predict the results in order to meet the vision of the educational institution. It is found that the linear regression model results are 80% accurate while the multiple and polynomial regression results are 100% accurate.

Keywords: Edu-DATA, Edu-MINING, Data Mining, Regression, Prediction, Multiple regression, Polynomial Regression.

Corresponding Author : Malini M. Patil

INTRODUCTION

Database can be defined, as the collection of data usually associated with any real world concept based on its various functions. Many organizations have accumulated vast amounts of data with the rapid advance of technology in data collection. Databases today can range in gigabytes, terabytes and peta bytes of size. Within these large databases, there lies a hidden information of strategic importance which is obtained through Data Mining(DM). Data mining is concerned with the analysis of data and the use of the software techniques for finding patterns in the data sets. The computational techniques are responsible for finding the patterns, which are previously unknown, presently useful for future analysis. DM is an integral part of Knowledge Discovery in Databases (KDD), which is the overall process of converting raw data into useful and structured information. KDD is more than DM. The knowledge discovery process comprises of six phases, Viz., Data selection, Data cleaning, Data enrichment, Data transformation or encoding, Data mining, Reporting and display of the discovered information. Data mining focuses on different ideas such as sampling, estimation hypothesis testing from statistics, search algorithms, modeling techniques machine learning theories from artificial intelligence, pattern recognition and machine learning, hi-performance computing,. Thus, data mining is represented as a confluence of many disciplines as shown in the fig. 1. A number of other areas also play key supporting roles. In particular data systems are needed to provide support for efficient storage, indexing and query processing. Techniques from high performance computing are often important in addressing the massive size of some data sets. Since data mining is a “Confluence of Many Disciplines“, among which the statistical approach plays a very important role in predictions.


STATITSICS , REGRESSION ANALYSIS, PCA, PREDICTIONS
DATA BASE TECHNOLOGY
DATA MINING
DISTRIBUTED COMPUTING, VISULAIZATION , INFORMATION RETRIVAL
AI, MACHINE LEARNING PATTERN RECIGNITION, AND IMAGE PROCESSING

s

Fig 1: Data mining as a Confluence of many Disciplines

The advancement of technology has resulted in the evolution of different techniques in the area of DM. New research findings resulted in new issues in each techniques. To quote some: association rule mining, classification, clustering, Support Vector Machines, data stream mining, image mining, text mining etc. Many organizations worldwide are already using DM techniques to explore the hidden useful information from the respective databases. Educational Mining (Edu-mining) is a method of exploring Educational data (Edu-data) which helps the technical education system(TES) to take useful decisions for maintaining the quality of the education system. Edu-data which is a large data repository consisting of data related to educational systems. It has earned lot of scope in educational research. Edu-data is evolved because of huge collection of data mainly from WWW, study material available in the internet, e-learning schemes, computerization of education system, online registration schemes for admission process in the universities, student information system, examination evaluation systems etc. Recent development of such data repository not only belongs to higher education system but also to the secondary education system. Educational Mining(Edu -mining) is a method of mining Educational data. Predictive analysis is one novel approach for proper predictions in student stakeholder of the typical Edu-mining system. Aim of the paper is to study the TES using regression analysis to provide a better predictions to teaching, learning and management process of the education system. This paper emphasizes a regression model using multiple and polynomial approaches by considering all the above aspects.

The paper is organized as follows: section II focuses mainly on the thorough study of the typical education system, which is taken as a benchmark system for Edu-mining. Section III discuses the related work about statistical approach in mining education data. Section IV focuses on Edu-MINING using linear regression analysis. Section V is about the implementation steps and Section VI is about the results and analysis respectively. Future enhancement of the work and conclusions are briefed at the end of the paper.

Technical Education system(TES) : A Bench mark system for Edu-Mining.

This section mainly focuses on the typical education system, which is considered as a bench mark system for the study of Edu-mining. The system is organized by three main components, which are called as stakeholders shown in the fig. 2. The three important stakeholders of the system are discussed as follows: Stakeholder one is **Management**, which is the supreme authority to manage the system. Stakeholder two is **Students** who are considered as the main revenue generators in the system, who work on a give and take policy. They have to join the institution to acquire a degree of their choice, pay the fees as per the norms and expect a home away from home atmosphere in the system. Stakeholder three is **Teachers** who are instrumental in strengthening of the system and are in teaching and learning process. With this brief overview of the system, we will try to analyze the system under three different headings, namely.,

- Goal seeking analysis,
- Optimization analysis,
- Sensitivity analysis.

The detailed discussion on these different approaches of analysis is based on the typical technical education system that we have considered.

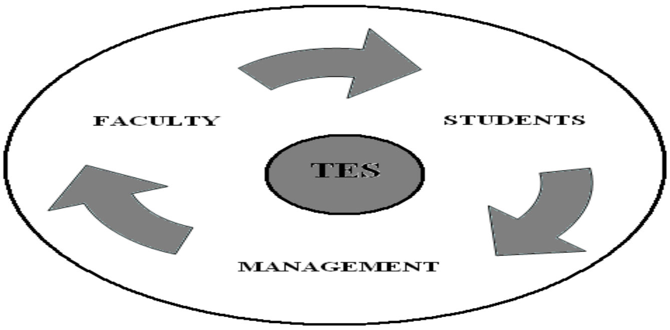


Fig. 2. A Typical Technical Education System(TES)

Goal seeking analysis

This analysis mainly focuses on the aims and objectives of the institution set by the management or in other words, it can be stated as the goal of the management. They are summarized as follows: The mission and vision of the management must be very clear. Management aims at keeping the brand image of the system as the best institution providing highly sophisticated infrastructure. Maintaining and following the proper evaluation and grading policies set by the statutory bodies and providing facilities for the faculty and students in the teaching and learning process are the two main necessary activities of the management. Conducting regular induction programs for students, encouraging them by providing extra tutorials for below average students to improve the results, arranging industry visits, promoting good hobby projects, arranging fests, sports, and other all extra co-curricular activities are essential. Selection of students is based on proper aptitude tests. Revision of institutional policies for in-house activities of both faculty and students must be carried out. Mentor system activities must be initiated to make the student-teacher relationship stronger.

Optimization analysis

This kind of analysis in the system is mainly concerned with the qualitative measures of the system. They include standardization of policies related to administrative procedures. Proper faculty recruitment procedures, strengthening of training and placement activities by providing soft skills training program, training of technical staff with latest industry requirements, signing of MOUs with industries, developing Industry-institute-interactions, providing hostel, sports and transportation facilities, and constitution of anti-ragging committees are few important areas which needs to be focused. All these optimization issues are management dependent and the policy measures should be regularly improvised so that the institute can maintain a good ranking in the present educational scenario.

Sensitivity analysis

This kind of analysis deals with selection procedures for students in order to improve the quality of the intake and to maintain overall ranking of the institution. Perfect result analysis procedures are essential to find out short falls in teaching process, faculty feed back by students to improve the faculty responsibilities in teaching, and promotions to both teaching and non-teaching staff based on their attitude and aptitude. Motivational programs like arranging faculty development programs, encouraging the faculty to attend conferences and workshops, encouraging them to present papers, attending summer schools and winter schools, floating the idea of best teacher award etc. are the important sensitivity parameters.

The presentation of a typical education system in this compact manner is to find the key areas where proper improvements could be implemented in the system. The main objective of the present investigation is to provide recommendations directly to the students, faculty and management with respect to their personalized activities. Currently the paper concentrates on only one stakeholder i.e., student. The analysis is carried out based on the data available in the student database and faculty data base of an education system. Comparisons with respect to hypothetical data and real data are done and the results are presented and discussed.

RELATED WORK

A thorough survey of the literature reveals that very sparse literature is available pertaining to the present work. Some amount of work in this regard has been done and is outlined briefly in this section. The authors emphasize that with regard to edu-mining the only works are[1,2,3,4] where the authors have not used the statistical approach. Therefore the present investigation is carried out to provide an excellent platform for future research.

Most of the related work was found on the analysis and visualization of data. The objective of the analysis and visualization of data is to highlight useful information and support decision making. In the educational environment, for example, it can help educators ,course administrators, management to analyze the students course activities and usage information to get a general view of a student's learning. Statistics and visualization information are the two main techniques that have been most widely used for this task.

Statistics is a mathematical science concerning the collection, analysis, interpretation or explanation, and presentation of data [5]. It is relatively easy to get basic descriptive statistics from statistical software, such as SPSS[6]. Statistical analysis of educational data can tell us things such as: where students enter and exit, the most popular pages, the browsers students tend to use, and patterns of use over time, [7]; the number of visits, origin of visitors, number of hits, and patterns of use throughout various time periods [8]; number of visits and duration per quarter, top search terms, and number of downloads of e-learning resources [9]; number of different pages browsed and total time for browsing different pages [10]; usage summaries and reports on weekly and monthly user trends and activities [11]; session statistics and session patterns [12]; statistical indicators on the learner's interactions in forums [13]; the amount of material students might go through and the order in which students study topics [14]; resources used by students and resources valued by students [15]; Statistical analysis is also very useful to obtain reports assessing [16] how many minutes the student has worked, how many minutes he has worked today, how many problems he has resolved, and his correct percentage, our prediction of his score, and his performance level. Information visualization uses graphic techniques to help people to understand and analyze data [17].Visual representations and interaction techniques take advantage of the human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once. There are several studies oriented toward visualizing different educational data such as: patterns of annual, seasonal, daily, and hourly user behavior on online forums [18]; the complete educational (assessment) process [19]; mean values of attributes analyzed in data to measure mathematical skills [20]; tutor–student interaction data from an automated reading tutor [21]; statistical graphs about assignments complement, questions admitted, exam score, etc. [22]; student tracking data regarding social, cognitive, and behavioural aspects of students [23]; student's attendance, access to resources, overview of discussions, and results on assignments and quizzes[24];student's progression per question as a transition between the types of questions [25]; fingertip actions in collaborative learning activities [26]; deficiencies in a student's basic understanding of individual concepts [27] and higher education student-evaluation data [28]; student's interactions with online learning environments [29]; the students' online exercise work, including students' interactions and answers, mistakes, teachers' comments, etc. [30]; questions and suggestions in an adaptive tutorial [31]; navigational behavior and the performance of the learner [32]; educational trails of Web pages visited and activities done[33].

REGRESSION ANALYSIS

Regression analysis finds the relationship between a dependent variable and one or more independent variables. Regression is also used in data mining technique used to fit an equation to a dataset. The objective of the analysis and visualization of data is to highlight useful information and support decision making. In the educational environment, for example, it can help educators and course administrators to analyze the students' course activities and usage information to get a general view of a student's learning. Statistics and visualization information are the two main techniques that have been most widely used for this task. Statistics is a mathematical science concerning the collection, analysis, interpretation or explanation, and presentation of data. Several DM techniques have been used for this task and the most common are association-rule mining, clustering, and sequential pattern mining. But no work is available where in the regression models are studied.

The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, polynomial regression allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation, cubic equation respectively. A regression task begins with a data set in which the target values are known. For example, a regression model that predicts the sales of present year for each salesman of a company keeping last year sales as dependent variable and years of experience as independent variable. Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values. Regression modeling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling, and environmental modeling.

How Does Regression Work?

Regression analysis determine the values of parameters for a function that cause the function to best fit a set of data observations that are available. Equation 1 shows that regression is the process of estimating the value of a continuous target variable (y) as a function (F) of one or more predictor variables ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), a set of parameters (θ_1 , θ_2 , ..., θ_n), and a measure of error (e).

$$y = F(\mathbf{x}, \theta) + e \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

The predictors can be understood as independent variables and the target as a dependent variable. The error, also called the **residual**, is the difference between the expected and predicted value of the dependent variable. The regression parameters are also known as **regression coefficients**. The process of training a regression model involves finding the parameter values that minimize a measure of the error, for example, the sum of squared errors.

LINEAR REGRESSION

Linear regression is a statistical technique that is used to learn more about the relationship between an independent (predictor) variable and a dependent (criterion) variable. A linear regression technique can be used if the relationship between the predictors and the target can be approximated with a straight line. Simple linear regression with a single predictor is shown in fig 3.

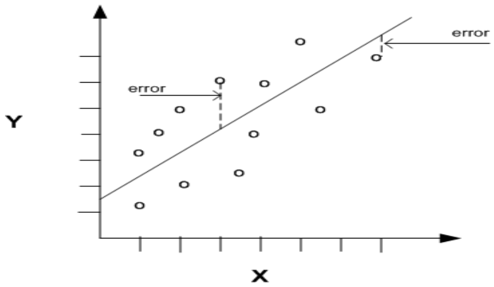


Fig 3. Linear Regression with a Single Predictor

Linear regression with a single predictor can be expressed with the following equation.

$y = \theta^2\mathbf{x} + \theta^1 + e$...2

The regression parameters in simple linear regression are:

- The **slope** of the line (θ^2) — the angle between a data point and the regression line
- The ***y* intercept** (θ^1) — the point where **x** crosses the *y* axis (**x** = 0)

The Coefficient of Determination

The **Coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination ranges from 0 to 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable.
- An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

MULTIPLE REGRESSION

Multiple regression is a statistical technique which consists of more than one independent variable in the analysis. Suppose as a test case it is necessary to check for an organization to predict how much an individual is satisfied with his/her job. Variables such as salary, academic qualifications, age, gender, number of years in full-time employment and socioeconomic status might all contribute towards job satisfaction. By applying multiple regression technique for the data collected pertaining to employees of some organization, it is found that job satisfaction is most accurately predicted by the type of occupation, salary and years in full-time employment, with the other variables not helping us to predict job satisfaction. Type of occupation, salary and years in full-time employment would emerge as significant predictor variables, which allow us to estimate the criterion variable. But as human behaviour is inherently noisy, not possible to produce totally accurate predictions, but multiple regression allows us to identify a set of predictor variables which together provide a useful estimate of a participant's likely score on a criterion variable.

The criterion for using Multiple Regression are:
Multiple regression technique can be used when exploring linear relationships between the predictor and criterion variables – that is, when the relationship follows a straight line.

- The criterion variable that is used to predict should be measured on a continuous scale (such as interval or ratio scale).

- The predictor variable should be measured on a ratio, interval, or ordinal scale. Dichotomous variable must be coded properly. For example, sex is acceptable (where male is coded as 1 and female as 0) but gender identity (masculine, feminine and androgynous) could not be coded as a single variable. Instead three different variables must be created each with two categories (masculine/not masculine; feminine/not feminine and androgynous/not androgynous). The term dummy variable is used to describe this type of dichotomous variable.

- Multiple regression requires a large number of observations. The number of cases (participants) must substantially exceed the number of predictor variables. A more acceptable ratio is 10:1. Other co-efficients like ANOVA, Correlation, Beta, R, R^2 Adjusted R^2 are discussed in the following section.

CORRELATION AND ANALYSIS OF VARIANCE(ANOVA) IN MULTIPLE REGRESSION

If two variables are correlated, then knowing the score on one variable will allow us to predict the score on the other variable. The stronger the correlation, the closer the scores will fall to the regression line and therefore the more accurate is prediction. Multiple regression is simply an extension of this principle, where we predict one variable on the basis of several other variables. Thus, in the above example above, people might vary greatly in their levels of job satisfaction. For example, we might be able to say that salary accounts for a fairly large percentage of the variance in job satisfaction, and hence it is very useful to know someone's salary when trying to predict their job satisfaction. A current trend in statistics is to emphasise the similarity between multiple regression and ANOVA, and between correlation and the *t*-test. All of these statistical techniques are basically interested in explaining the variance in the level of one variable on the basis of the level of one or more other variables.

Beta (standardised regression coefficients)

The beta regression coefficient is computed to assess the strength of the relationship between each predictor variable to the criterion variable. The beta value is a measure of how strongly each predictor variable influences the criterion variable. The beta is measured in units of standard deviation. For example, a beta value of 2.5 indicates that a change of one standard deviation in the predictor variable will result in a change of 2.5 standard deviations in the criterion variable. Thus, the higher the beta value the greater the impact of the predictor variable on the criterion variable. When only one predictor variable exists in the model, then beta is equivalent to the correlation coefficient between the predictor and the criterion variable. This situation is a correlation between two variables. When more than one predictor variable exists , it is impossible to compare the contribution of each predictor variable by simply comparing the correlation coefficients.

R, R^2, Adjusted R^2

R is a measure of the correlation between the observed value and the predicted value of the criterion variable. R-Square (R^2) is the square of this measure of correlation and indicates the proportion of the variance in the criterion variable which is accounted for the model by set of predictor variables.

POLYNOMIAL REGRESSION

Polynomial regression is a form of linear regression in which the relationship between the independent variable *x* and the dependent variable *y* is modeled as an n^{th}order polynomial. Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding conditional mean of *y*, denoted E(*y*|*x*). Few examples to quote are: description of the nonlinear phenomena such as the growth rate of tissues, the distribution of carbon isotopes in lake sediments and the progression of disease epidemics. Although *polynomial regression* fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(*y*|*x*) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression. Polynomial regression models are usually fit using the method of least squares. The goal of regression analysis is to model the expected value of a dependent variable *y* in terms of the value of an independent variable (or vector of independent variables) *x*. In simple linear regression, the model

$$y = a_0 + a_1 x + \varepsilon, \text{...} 3$$

is used, where ε is an unobserved random error with mean zero conditioned on a scalar variable *x*. In this model, for each unit increase in the value of *x*, the conditional expectation of *y* increases by a_1 units. In many settings, such a linear relationship may not hold. For example, if we are modeling the yield of a chemical synthesis in terms of the temperature at which the synthesis takes place, we may find that the yield improves by increasing amounts for each unit increase in temperature. In this case, we might propose a quadratic model of the form

$$y = a_0 + a_1 x + a_2 x^2 + \varepsilon. \text{..} (4)$$

In this model, when the temperature is increased from *x* to *x* + 1 units, the expected yield changes by $a_1 + a_2 + 2a_2x$. The fact that the change in yield depends on *x* is what makes the relationship nonlinear. Point to be noted here that this must not be confused with saying that this is nonlinear regression; on the contrary, this is still a case of linear regression. In general, we can model the expected value of *y* as an *n*th order polynomial, yielding the general polynomial regression model

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_m x^m + \varepsilon. \text{..........} (5)$$

Conveniently, these models are all linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters a_0, a_1, ... Therefore, for least squares analysis problems of polynomial regression can be completely addressed using the techniques of multiple regression. This is done by treating x, x^2, ... as being distinct independent variables in a multiple regression model.

EXPERIMENTS AND RESULTS

Regression analysis on Edu-data is performed using IBM PASW Statistics 18. PASW can perform variety of data analysis and presentation functions, including the statistical analysis and graphical representation of data. Before using PASW on the data set it is recommended to run a Scatter plot. Before performing a regression analysis to determine if there is a linear relationship between the variables. If there is no linear relationship, no need to run a simple regression. The following figure clearly shows that points on a graph are clustered in a straight line. This clearly indicates that there is a linear relationship between the variables and the simple regression can be run in PASW. The dataset referred to show the scatter plot is edu-data. The data set description is shown in table 1.

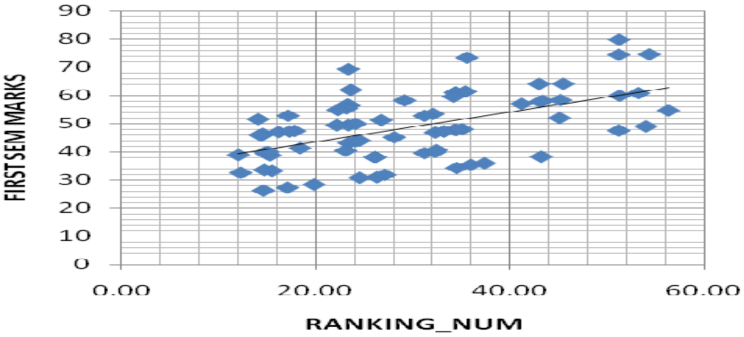


Fig 4.Scatterr plot to check the linearity

The student data set in Edu-data comprises of different attributes as shown in the following table 1.

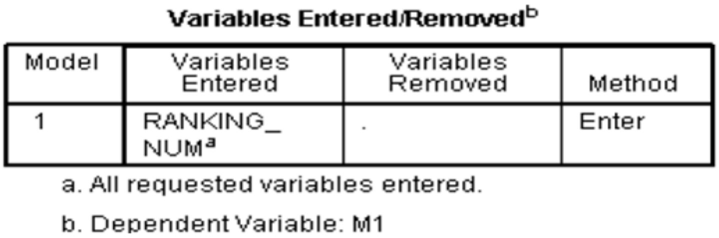
Table 1. Student Data set description in Edu-data

Attribute	**Description of attributes**
NAME	Name of the student
USN	University seat number
GENDER	Gender
MODE	Mode of entry whether diploma or Regular
SEAT_TYPE	Type of seat obtained
TENTH_ MARKS	Tenth std marks
PU_MARKS	12 th std marks
RANKING_NUM	Ranking no. of a student
M1	Marks of semester one
M2	Marks of semester two
ATTENDENCE	Overall attendance of the student
AGGREGATE_UPTO7TH SEM	Aggregate % of student from 1st semester to 7th semester
EXPECTED RESULT	RESULT AFTER APPLYING MULTIPLE AND CUBIC POLYNOMIAL REGRESSION

RESULTS OF LINEAR REGRESSION

The following four tabulations namely Table 2(a) ,2(b) and 2(c) depict the results obtained from PASW-18 for student result predictions for second semester of their engineering course[4]. The last column of table 3 represents the predictions of results. It is found that the predictions were almost nearing to the actual values.

Table 2.(a)



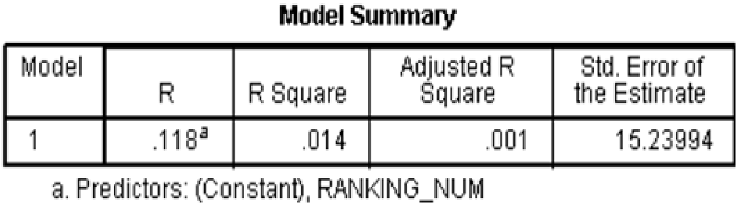
Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	RANKING_NUM[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: M1

Table 2.(b)

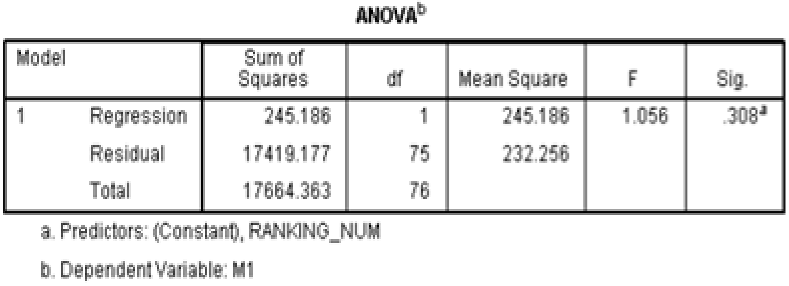


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.118[a]	.014	.001	15.23994

a. Predictors: (Constant), RANKING_NUM

Table 2.(c)



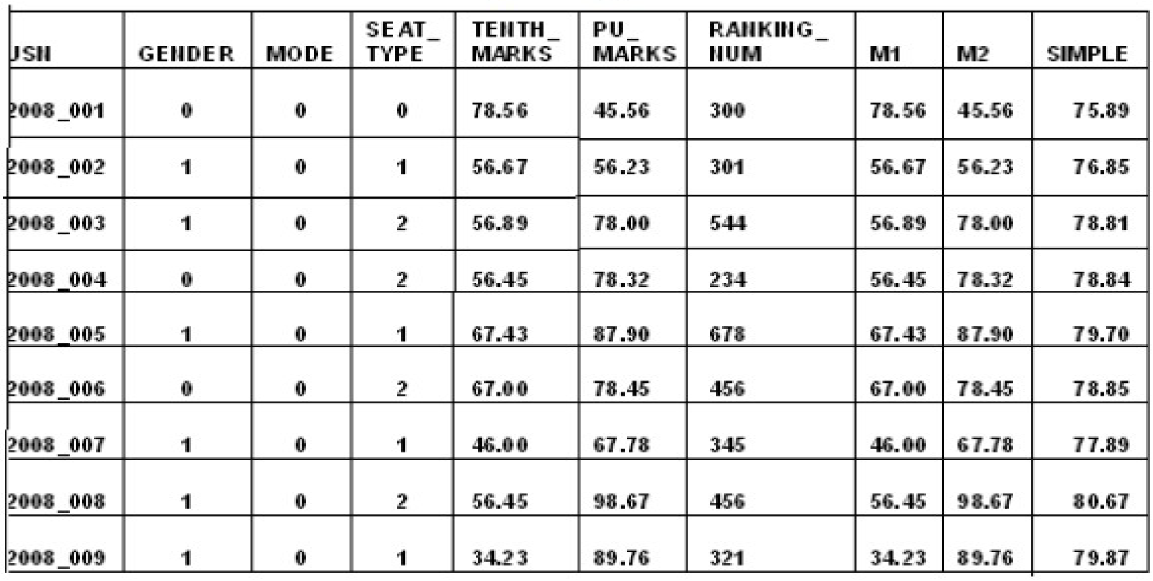
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	245.186	1	245.186	1.056	.308[a]
	Residual	17419.177	75	232.256		
	Total	17664.363	76			

a. Predictors: (Constant), RANKING_NUM

b. Dependent Variable: M1

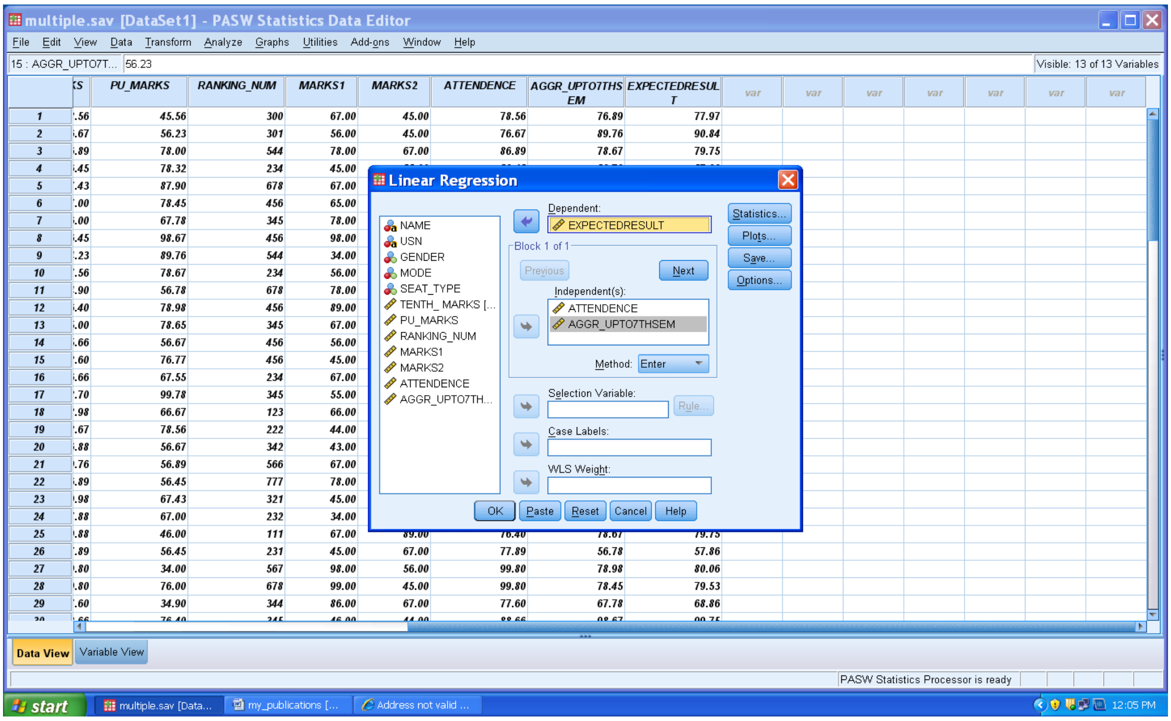
Table 3: Simple Regression Table



JSN	GENDER	MODE	SEAT_TYPE	TENTH_MARKS	PU_MARKS	RANKING_NUM	M1	M2	SIMPLE
2008_001	0	0	0	78.56	45.56	300	78.56	45.56	75.89
2008_002	1	0	1	56.67	56.23	301	56.67	56.23	76.85
2008_003	1	0	2	56.89	78.00	544	56.89	78.00	78.81
2008_004	0	0	2	56.45	78.32	234	56.45	78.32	78.84
2008_005	1	0	1	67.43	87.90	678	67.43	87.90	79.70
2008_006	0	0	2	67.00	78.45	456	67.00	78.45	78.85
2008_007	1	0	1	46.00	67.78	345	46.00	67.78	77.89
2008_008	1	0	2	56.45	98.67	456	56.45	98.67	80.67
2008_009	1	0	1	34.23	89.76	321	34.23	89.76	79.87

RESULTS OF MULTIPLE REGRESSION

The following snap shots present the results of multiple regression. In the practical scenario of TES a student's result can be expected based on his overall attendance and aggregate marks up to seventh semester. From fig 5 it is clear that expected result is dependent variable and independent variables are *overall attendance(*ATTENDENCE*)* and *aggregate marks up to seventh semester(*AGGR_UPTO7THSEM).



multiple.sav [DataSet1] - PASW Statistics Data Editor

15 : AGGR_UPTO7T... 56.23

Visible: 13 of 13 Variables

	KS	PU_MARKS	RANKING_NUM	MARKS1	MARKS2	ATTENDENCE	AGGR_UPTO7THSEM	EXPECTEDRESULT
1	.56	45.56	300	67.00	45.00	78.56	76.89	77.97
2	.67	56.23	301	56.00	45.00	76.67	89.76	90.84
3	.89	78.00	544	78.00	67.00	86.89	78.67	79.75
4	.45	78.32	234	45.00				
5	.43	87.90	678	67.00				
6	.00	78.45	456	65.00				
7	.00	67.78	345	78.00				
8	.45	98.67	456	98.00				
9	.23	89.76	544	34.00				
10	.56	78.67	234	56.00				
11	.90	56.78	678	78.00				
12	.40	78.98	456	89.00				
13	.00	78.65	345	67.00				
14	.66	56.67	456	56.00				
15	.60	76.77	456	45.00				
16	.66	67.55	234	67.00				
17	.70	99.78	345	55.00				
18	.98	66.67	123	66.00				
19	.67	78.56	222	44.00				
20	.88	56.67	342	43.00				
21	.76	56.89	566	67.00				
22	.89	56.45	777	78.00				
23	.98	67.43	321	45.00				
24	.88	67.00	232	34.00				
25	.88	46.00	111	67.00	89.00	76.40	78.67	79.75
26	.89	56.45	231	45.00	67.00	77.89	56.78	57.86
27	.80	34.00	567	98.00	56.00	99.80	78.98	80.06
28	.80	76.00	678	99.00	45.00	99.80	78.45	79.53
29	.60	34.90	344	86.00	67.00	77.60	67.78	68.86

Linear Regression

Dependent: EXPECTEDRESULT

Block 1 of 1

Independent(s): ATTENDENCE, AGGR_UPTO7THSEM

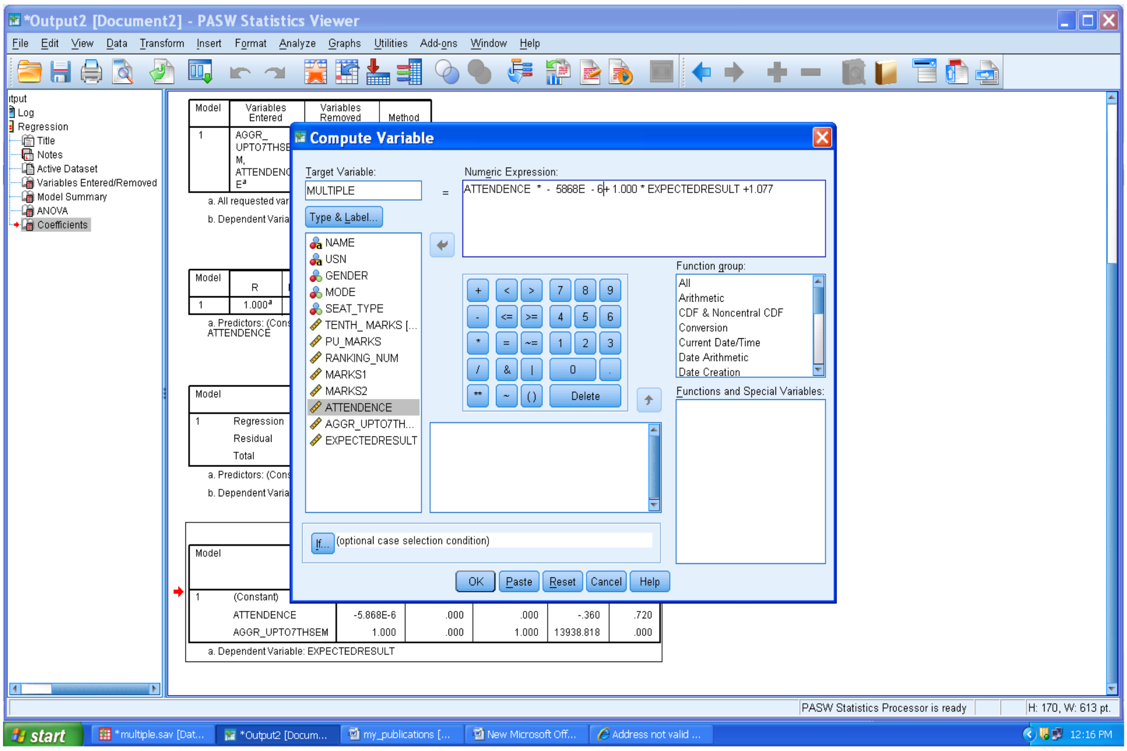
Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Fig 5. Linear regression dialog box for selecting two independent variables



*Output2 [Document2] - PASW Statistics Viewer

Model	Variables Entered	Variables Removed	Method
1	AGGR_UPTO7THSEM, ATTENDENCE^a		

a. All requested var
b. Dependent Varia

Model	R
1	1.000^a

a. Predictors: (Cons ATTENDENCE

Model	
1	Regression
	Residual
	Total

a. Predictors: (Cons
b. Dependent Varia

Model						
1	(Constant)					
	ATTENDENCE	-5.868E-6	.000	.000	-.360	.720
	AGGR_UPTO7THSEM	1.000	.000	1.000	13938.818	.000

a. Dependent Variable: EXPECTEDRESULT

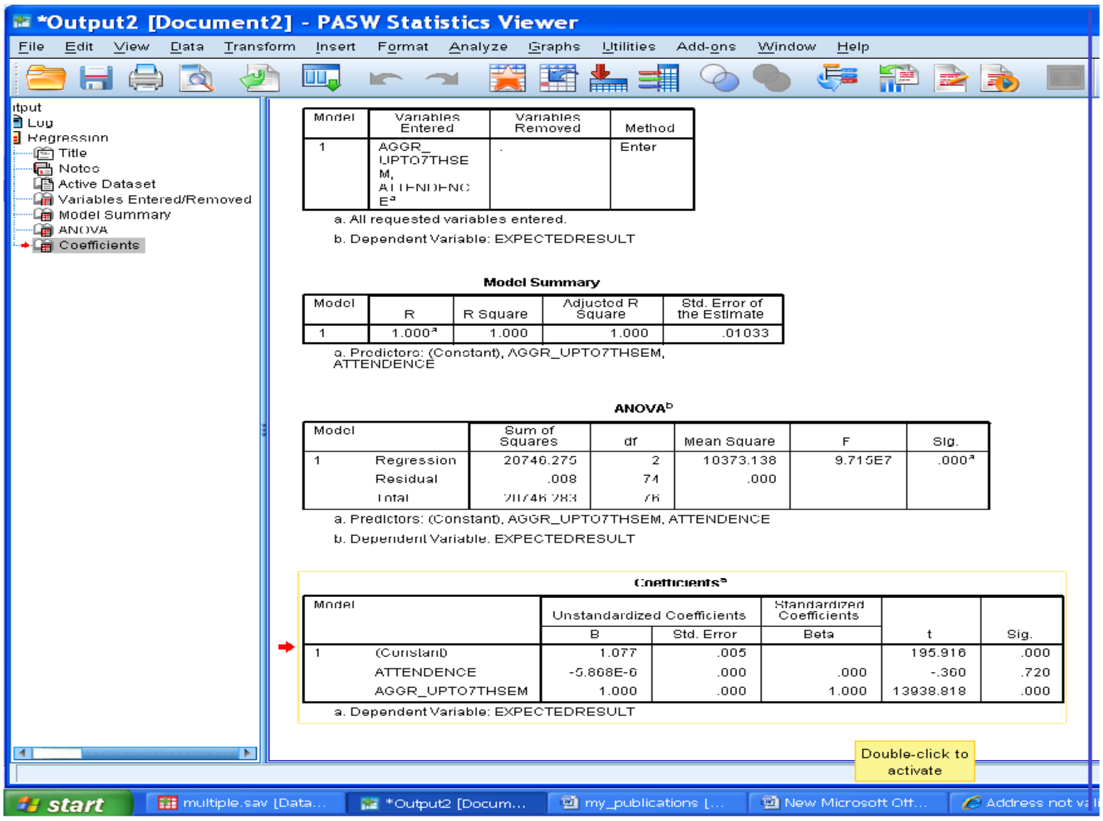
Compute Variable

Target Variable: MULTIPLE

Numeric Expression: ATTENDENCE * - 5868E - 6+ 1.000 * EXPECTEDRESULT +1.077

(optional case selection condition)

Fig.6. Dialog box from PASW for equation generated for multiple regression



Model	Variables Entered	Variables Removed	Method
1	AGGR_UPTO7THSEM, ATTENDENCE[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: EXPECTEDRESULT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	1.000	.01033

a. Predictors: (Constant), AGGR_UPTO7THSEM, ATTENDENCE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20746.275	2	10373.138	9.715E7	.000[a]
	Residual	.008	74	.000		
	Total	20746.283	76			

a. Predictors: (Constant), AGGR_UPTO7THSEM, ATTENDENCE

b. Dependent Variable: EXPECTEDRESULT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.077	.005		195.916	.000
	ATTENDENCE	-5.868E-6	.000	.000	-.360	.720
	AGGR_UPTO7THSEM	1.000	.000	1.000	13938.818	.000

a. Dependent Variable: EXPECTEDRESULT

Fig 7. Results of model summary ANOVA etc in multiple regression

The slopes and the y-intercept should be substituted in the following linear equation to predict the results of students in the equation **Z = aX + bY + c**. From the model summary obtained the above equation is written as follows

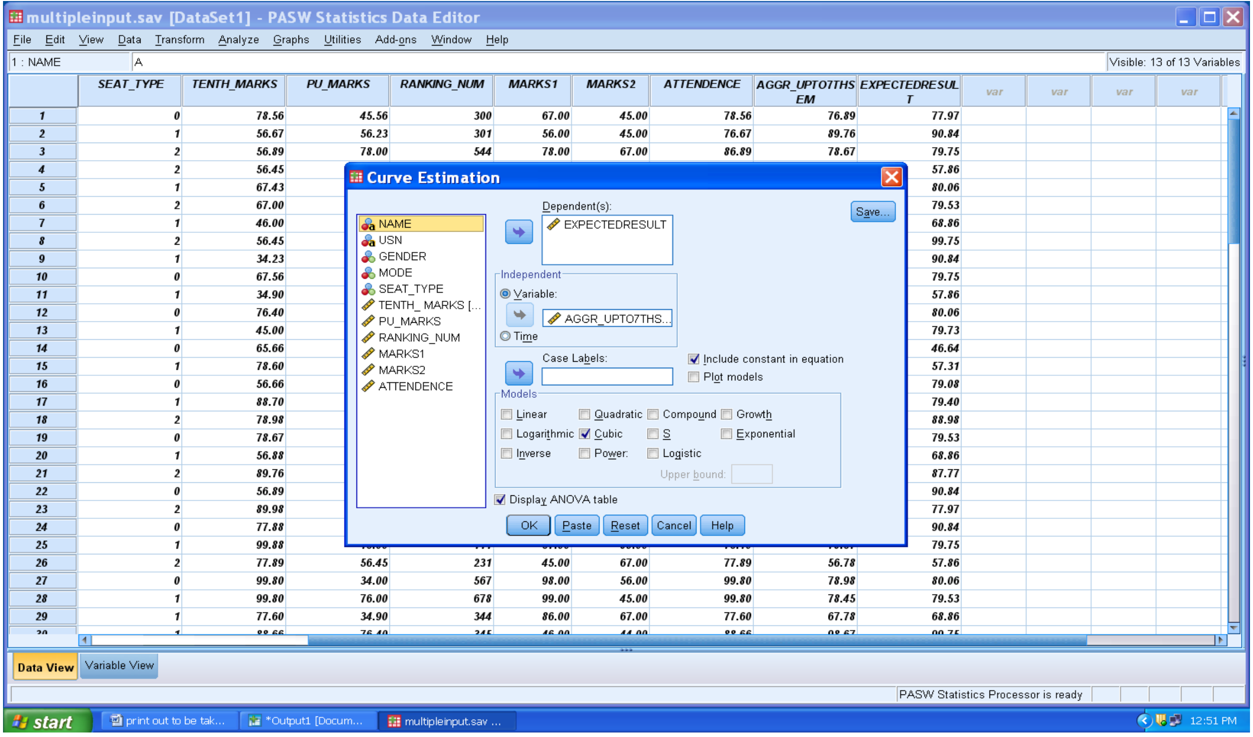
Expected result(Z) =(-5.868e-6) x ATTENDANCE + 1.000 x AGGREGATE UPTO SEVENTH SEM +1.077 ....(6)

Which is entered in to the compute variable dialog box as shown in figure 6. From the model summary obtained as shown in fig 7 it is clear that the R^2 Value is 1.000 and hence the results obtained using multiple regression are accurate. The results are shown in table 8.

Table 8: Results of multiple regression

USN	GENDER	MODE	SEAT_TYPI	TENTH_M.	PU_MARK	RANKING_	MARKS1	MARKS2	ATTENDEI	AGGR_UP	EXPECTEDRESULT
2008_001	0	0	0	78.56	45.56	300	67	70	78.56	76.89	77.97
2008_002	1	0	1	56.67	56.23	301	56	59	76.67	89.76	90.84
2008_003	1	0	2	56.89	78	544	78	81	86.89	78.67	79.75
2008_004	0	0	2	56.45	78.32	234	45	48	56.45	56.78	57.86
2008_005	1	0	1	67.43	87.9	678	67	70	67.43	78.98	80.06
2008_006	0	0	2	67	78.45	456	65	68	98	78.45	79.53
2008_007	1	0	1	46	67.78	345	78	81	78	67.78	68.86
2008_008	1	0	2	56.45	98.67	456	98	101	56.45	98.67	99.75
2008_009	1	0	1	34.23	89.76	544	34	37	34.23	89.76	90.84
2008_010	0	1	0	67.56	78.67	234	56	59	67.56	78.67	79.75
2008_011	0	0	1	34.9	56.78	678	78	81	34.9	56.78	57.86
2008_012	0	0	0	76.4	78.98	456	89	92	76.4	78.98	80.06
2008_013	1	0	1	45	78.65	345	67	70	98	78.65	79.73
2008_014	1	0	0	65.66	56.67	456	56	59	78.56	45.56	46.64
2008_015	1	0	1	78.6	76.77	456	45	48	56.67	56.23	57.31
2008_016	1	0	0	56.66	67.55	234	67	70	56.89	78	79.08
2008_017	1	0	1	88.7	99.78	345	55	58	56.45	78.32	79.4
2008_018	1	0	2	78.98	66.67	123	66	69	67.43	87.9	88.98

RESULTS OF POLYNOMIAL REGRESSION. The polynomial regression dialog box is presented in figure 8 for the cubic polynomial case.



multipleinput.sav [DataSet1] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : NAME A

Visible: 13 of 13 Variables

	SEAT_TYPE	TENTH_MARKS	PU_MARKS	RANKING_NUM	MARKS1	MARKS2	ATTENDENCE	AGGR_UPTO7THSEM	EXPECTEDRESULT	var	var	var	var
1	0	78.56	45.56	300	67.00	45.00	78.56	76.89	77.97				
2	1	56.67	56.23	301	56.00	45.00	76.67	89.76	90.84				
3	2	56.89	78.00	544	78.00	67.00	86.89	78.67	79.75				
4	2	56.45							57.86				
5	1	67.43							80.06				
6	2	67.00							79.53				
7	1	46.00							68.86				
8	2	56.45							99.75				
9	1	34.23							90.84				
10	0	67.56							79.75				
11	1	34.90							57.86				
12	0	76.40							80.06				
13	1	45.00							79.73				
14	0	65.66							46.64				
15	1	78.60							57.31				
16	0	56.66							79.08				
17	1	88.70							79.40				
18	2	78.98							88.98				
19	0	78.67							79.53				
20	1	56.88							68.86				
21	2	89.76							87.77				
22	0	56.89							90.84				
23	2	89.98							77.97				
24	0	77.88							90.84				
25	1	99.88							79.75				
26	2	77.89	56.45	231	45.00	67.00	77.89	56.78	57.86				
27	0	99.80	34.00	567	98.00	56.00	99.80	78.98	80.06				
28	1	99.80	76.00	678	99.00	45.00	99.80	78.45	79.53				
29	1	77.60	34.90	344	86.00	67.00	77.60	67.78	68.86				

Curve Estimation

NAME
USN
GENDER
MODE
SEAT_TYPE
TENTH_ MARKS [...
PU_MARKS
RANKING_NUM
MARKS1
MARKS2
ATTENDENCE

Dependent(s): EXPECTEDRESULT

Independent
● Variable: AGGR_UPTO7THS...
○ Time

Case Labels:

☑ Include constant in equation
☐ Plot models

Models
☐ Linear ☐ Quadratic ☐ Compound ☐ Growth
☐ Logarithmic ☑ Cubic ☐ S ☐ Exponential
☐ Inverse ☐ Power: ☐ Logistic
Upper bound:

☑ Display ANOVA table

PASW Statistics Processor is ready

Fig 8. Polynomial regression dialog box

The results of polynomial regression are presented in the following tables 9(a) to 9(f). This cubic model has R^2 value of 100% (From Table 9(d)). The F ratio (from Table 9(e)) indicates a highly significant fit. The best fitting cubic polynomial is given by the following equation (7).

$$Y_i=1.081 + 1.000X_i + 9.501E - 7X_i^2 - 2.074E - 9X_i^3 \quad (7)$$

Table 9(a) Model Description

Model Name		MOD_3
Dependent Variable	1	EXPECTEDRESULT
Equation	1	Cubic
Independent Variable		AGGR_UPTO7THSEM
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified
Tolerance for Entering Terms in Equations		.0001

Table 9(b) Case Processing Summary

	N
Total Cases	78
Excluded Cases[a]	1
Forecasted Cases	0
Newly Created Cases	0

Table 9(b) Case Processing Summary

	N
Total Cases	78
Excluded Cases[a]	1
Forecasted Cases	0
Newly Created Cases	0

a. Cases with a missing value in any variable are excluded from the analysis.

Table 9(c) Variable Processing Summary

		Variables	
		Dependent	Independent
		EXPECTEDRESULT	AGGR_UPTO7THSEM
Number of Positive Values		77	77
Number of Zeros		0	0
Number of Negative Values		0	0
Number of Missing Values	User-Missing	0	0
	System-Missing	1	1

EXPECTEDRESULT
Cubic

Table 9(d) Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
1.000	1.000	1.000	.010

The independent variable is AGGR_UPTO7THSEM.

Table 9 (f) Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
AGGR_UPTO7THSEM	1.000	.001	1.000	1538.074	.000
AGGR_UPTO7THSEM ** 2	9.501E-7	.000	.000	.070	.944
AGGR_UPTO7THSEM ** 3	2.074E-9	.000	.000	.025	.980
(Constant)	1.081	.011		101.509	.000

CONCLUSION

Regression analysis on Edu-data is performed using IBM PASW Statistics 18. PASW can perform variety of data analysis and presentation functions, including the statistical analysis and graphical representation of data. No doubt the size of the Edu-data is quite big (since it comprises data with regard to the three stakeholders namely Student, faculty and Management) the present study focuses on the student data only. The same analysis can be effectively employed for the other cases also. The authors have made the thorough analysis classification technique on Edu-data. Present results are explored using statistical method namely multiple regression and polynomial regression. Some of the important results of the present study are: (i) Scatter Plot for checking the linearity (ii) Prediction of the results pertaining to the simple linear regression analysis. Accurate predictions observed are 80%. (iii) Prediction of the results pertaining to the multiple regression analysis. It is amazing to note that 100% accuracy is achieved in this case and (iv) detailed results pertaining to polynomial regression show that the predictions are 100% accurate and accordingly the error is nil. Finally it is concluded that the results are found to be interesting and provide an excellent platform for future study. The results help the management to analyze the Edu-data by considering student as a Stake holder. Further the results could be effectively used effectively in the accreditation process with regard to the overall growth of the system.

ACKNOWLEDGEMENT

One of the authors Mrs. Malini M. Patil acknowledges J.S.S Academy of Technical education, Bangalore, Karnataka and Bhartiyaar University, Coimbatore, Tamilnadu, India for providing the facilities for carrying out the research work.

REFERENCE

[1] P.K Srimani, and Malini M Patil, Edu-mining:A Machine learning approach, AIP. Conf. Proc , 1414,pp,61-66, 2011.

[2] P.K Srimani, and Malini M Patil, A classification Model for Edu-mining, PSRC Proc,Dubai, UAE, pp, 35-40. Jan 2012

[3] P.K Srimani, and Malini M Patil, A Comparative Study of Classifiers for Student Module in Technical Education System(TES)", IJCR, Vol 4 Issue 01, pp 249-254, Jan 2012.

[4] P.K Srimani, Malini M Patil, and P.K Srivatsa Performance evaluation of Classifiers for Eduata : An integrated approach, IJCR, Vol 4 , Issue 02, pp 183-190,February, 2012.

[4] P.K Srimani, and Malini M Patil, Linear regression model for Edu-Mining in TES, Proceedings of ICCIT-2012, July 13-14,Tirupati, AP, India. 2012.

[5] D Freedman, R Purves Statistics 4th Edition, Newyork.

[6] A. Wu and C. Leung, Evaluating learning behavior of web-based training (WBT) using web log in *Proc. Int. Conf. Comput. Educ.*, New Zealand, pp. 736–737,2002.

[7] A. Ingram, Using web server logs in evaluating instructional web sites, *J. Educ. Technol. Syst.*, vol. 28, no. 2, pp. 137–157, 1999.

[8] J. Gibbs and M. Rice, "Evaluating student behavior on instructional Web sites using web server logs," in *Proc. Ninth Sloan-C Int. Conf. Online Learn.*, Orlando, FL, pp. 1–3, 2003.

[9] H. L. Grob, F. Bensberg, and F. Kaderali, Controlling open source intermediaries—A web log mining approach, in *Proc. Int. Conf. Inf. Technol. Interfaces*, Zagreb, Croatia, pp. 233–242, 2004.

[10] G. J. Hwang, P. S. Tsai, C. C. Tsai, and J. C. R. Tseng, A novel approach for assisting teachers in analyzing student web-searching behaviors, *Comput. Educ. J.*, vol. 51, pp. 926–938, 2008.

[11] D. Monk, Using data mining for e-learning decisionmaking, *Electron. J. E-Learning*, vol. 3, no. 1, pp. 41–54, 2005.

[12] C. Pahl and C. Donnellan, Data mining technology for the evaluation of web-based teaching and learning systems, in *Proc. Congr. E-Learning.*, Montreal, Canada, pp. 1–7, 2003.

[13] A. Anaya and J. Boticario, A data mining approach to reveal representative collaboration indicators in open collaboration frameworks, in *Proc. Int. Conf. Educ. Data Mining*, Cordoba, Spain, pp. 210–218, 2009.

[14] M. Rahkila and M. Karjalainen, Evaluation of learning in computer based education using log systems, in *Proc. ASEE/IEEE Front. Educ. Conf.*, San Juan, Puerto Rico, pp. 16–21, 1999.

[15] J. Sheard, J. Ceddia, J. Hurst, and J. Tuovinen, "Inferring student learning behaviour from website interactions: A usage analysis," *J. Educ. Inf. Technol.*, vol. 8, no. 3, pp. 245–266, 2003.

[16] M. Feng and N. Heffernan, Informing teachers live about student learning: Reporting in the assistment system, *Technol., Instruction, Cognition, Learn. J.,* vol. 3, pp. 1–8, 2006.

[17] R. Mazza, *Introduction to Information Visualization*. NewYork: Springer-Verlag, 2009.

[18] L. Burr and D. H. Spennemann, Pattern of user behavior in university online forums, *Int. J. Instruct. Technol. Distance Learn.*, vol. 1, no. 10, pp. 11–28, 2004.

[19] M. Pechenizkiy, N. Trcka, E. Vasilyeva,W. Aalst, and P. De bra, Process mining online assessment data, in *Proc. Int. Conf. Educ. Data Mining*, Cordoba, Spain, pp. 279–288, 2009.

[20] L. Zoubek and M. Burda, Visualization of differences in data measuring mathematical skills, in *Proc. Int. Conf. Educ. Data Mining*, Cordoba, Spain, pp. 315–324, 2009.

[21] J. Mostow, J. Beck, H. Cen, A. Cuneo, E. Gouvea, and C. Heiner, An educational data mining tool to browse tutor-student interactions: Time will tell!, in *Proc. Workshop Educ. Data Mining*, pp. 15–22, 2005.

[22] R. Shen, F. Yang, and P. Han, Data analysis center based on e-learning platform, in *Proc.Workshop Internet Challenge: Technol. Appl.*, Berlin, Germany, pp. 19–28, 2002.

[23] R. Mazza and D. Vania, The design of a course data visualizator: An empirical study, in *Proc. Int. Conf. New Educ. Environ.*, Lucerne, Switzerland, pp. 215–220, 2003.

[24] R. Mazza and C.Milani, GISMO: A graphical interactive student monitoring tool for course management systems, in *Proc. Int. Conf. Technol. Enhanced Learn.*, Milan, Italy, pp. 1–8, 2004.

[25] D. Ben-naim, N. Marcus, and M. Bain, Visualization and analysis of student interaction in an adaptive exploratory learning environment, in *Proc. Int. Workshop Intell. Support Exploratory Environ. Eur. Conf. Technol. Enhanced Learn.*,Maastricht, The Netherlands, pp. 1–10, 2008.

[26] N. Avouris, V. Komis, G. Fiotakis, M. Margaritis, and E. Voyiatzaki, Why logging of fingertip actions is not enough for analysis of learning activities, in *Proc. AIED Conf. Workshop Usage Anal. Learn. Syst.*, Amsterdam, The Netherlands, pp. 1–8, 2005.

[27] J. Yoo, S. Yoo, C. Lance, and J. Hankins, Student progress monitoring tool using tree view, in *Proc. Tech. Symp. Comput. Sci. Educ., ACMSIGCESE*, pp. 373–377, 2006.

[28] H. Jin, T. Wu, Z. Liu, and J. Yan, Application of visual data mining in higher-education evaluation system, in *Proc. Int. Workshop Educ. Technol. Comput. Sci.*, Washington, DC, pp. 101–104, 2009.

[29] J. Jovanovic, D. Gasevic, C. Brooks, V. Devedzic, and M. Hatala, LOCO-Analyist: A tool for raising teacher's awareness in online learning environments, in *Proc. Eur. Conf. Technol. Enhanced Learn.*, Crete, Greece, pp. 112–126 ,2007.

[30] A. Merceron and K. Yacef, Mining student data captured from a webbased tutoring tool: Initial exploration and results, *J. Interactive Learning Res.*, vol. 15, no. 4, pp. 319–346, 2004.

[31] D. Ben-naim, M. Bain, and N. Marcus, A user-driven and data-driven approach for supporting teachers in reflection and adaptation of adaptive tutorials, in *Proc. Int. Conf. Educ. Data Mining*, Cordoba, Spain, pp. 21–30, 2009.

[32] A. Bellaachia and E. Vommina, MINEL: A framework for mining e-learning logs, in *Proc. Fifth IASTED Int. Conf. Web-Based Educ.*, Mexico, pp. 259–263, 2006.

[33] C. Romero, S. Gutierrez, M. Freire, and S. Ventura, Mining and visualizing visited trails in web-based educational systems, in *Proc. Int. Conf. Educ. Data Mining*, Montreal, Canada, pp. 182–185, 2008.