OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Regression techniques to study the student performance in post graduate examinations - A case study

Ananth. Y. N[1], Narahari. N. S[2]

[1]Associate Professor &Ph.D student- Dept of Computer Science- School of Graduate Studies, Jain University, Bangalore-India

[2]Professor and Head, Dept of Industrial Engineering & Management, R.V.College of Engineering Autonomous-VTU - Bangalore-India

***Abstract**: Aptitude of students entering into Post graduate courses in INDIA is an aspect to be studied. Entrance Examinations do test the aptitude but to a certain extent. Post graduate students are expected to have a certain level of aptitude and this aptitude should be sustained till the end of their course and beyond. Therefore, it is a necessity to examine to what extent their aptitudes are getting tested. The marks scored by the students in the entrance examination is an indicator of the aptitude but does not speak of the ability of the students in all the aspects or the subject of their specialization in the post graduate courses.It has been observed that a causal dependency exists between the degree marks and the entrance test marks. This paper tries to investigate this dependency with linear regression techniques. On the whole , categories of students are identified by studying the distribution of marks. A mapping between the marks and the questions are being studied. Linear regression technique is being used to identify the groups of students and to predict the expected marks that the future students would score.*

***Keywords**: Aptitude, Dependent variable, Performance, Predictor, Regression*

I. INTRODUCTION

Testing students' aptitude is a challenging area – given the wide variety of the questions in the degree and the entrance test examinations. Students' abilities are also varied – Some will be good in quantitative techniques, some in reasoning and analytical abilities and so on. The undergraduate degree examinations test mainly the knowledge level of the students-habitual learning methods directed at memorizing ,learning the facts and so on .The reasoning power and the spirit of enquiry at a higher level blooms by the finishing years of graduation. Hence it is imperative to study the progression of students. In this context, not all the students will have the same sustainable ability to persue post graduate courses. In this regard, setting the right kind of questions across all kinds of students is a difficult task .Therefore it is good to consider the marks in the undergraduate examinations and study the dependency of the entrance test marks on that. Taking this as an indicator, sectors of students could be identified .This information could be used to correlate to what kind of questions that would be suitable for students' groups-Doing statistical studies among the groups, the right kind of questions can be set for the right kind of groups. With this perspective, this paper discusses linear regression and its use. Although this techniques has its own limitations, this technique is an useful one. The particular case study that has been considered is the Karnataka –Post Graduate Common Entrance Test examinations

II. LITERATURE SURVEY

Many techniques do exist in order to study the distribution and patterns of marks in examinations. Reliability issues of the questions has been studied by S. O. Bandele and Dr. A. E. Adewale in [1**]**.Requirement of prerequisites for Masters programmes can be very clearly seen as in [2]

What can be inferred from this data is that the stream of the undergraduate degree is quite important for the choice of the subjects in post graduate courses.

Regression techniques are quite a powerful technique of analysis in many areas like market survey; understanding of customer behavior .The main use of regression techniques is that they can be used for investigating the causal dependencies between variables. These techniques are being used in education to study the causal factor analysis in many situations. For example regression techniques are used in analyzing the influence of peer pressure in secondary schools, upon the scores of the students. Multiple regression can be to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendance, and gender. This has also been used in analyzing the advantages of high ranked schools in moulding the students in an indirect way, by considering the marks scored by the students of such schools. The study of the effect of good teachers, good teaching aids upon student performance have also employed to a good extent the technique of regression. On the whole it can be said that wherever there is a study of causal factors relating to student performance in schools /colleges, regression techniques can be employed.

III. Brief Methodology

Linear regression is a modelling technique in which the relationship between a scalar dependent variable y and one or more explanatory variable X is being studied. The case of a single explanatory variable is called as linear regression and when there are more than one explanatory variable, it is called as multiple linear regression.

In linear regression, datasets are modelled by using linear predictor functions and unknown parameters are modelled using "predictor variables". More commonly, the conditional dependency of Y given X is what is being depicted here.The technique can be used for prediction where in a model to describe the behavior of Y given X can be developed. This model can be used in order to predict the data values Y for a new set of X.

The actual mathematical model for regression can be given as follows. Given a data set $\{y_i,x_{i1},x_{i2},\ldots\ldots x_{ip}\}$i=1to n,the regression model assumes that the relationship between the dependent variable y and the regressors x is linear.This relationship is modelled with a set of regression coefficients and an error term which is also called as noise. The regression coefficients give the extent to which a particular regressor x_i affects y.The relationship may not be perfect always and the difference between the actual relationship and the observed relationship is modelled using the error term or the noise term. Thus the model takes the form:

$Yi= \beta_1 x_{i1}+ \beta_2 x_{i2} +\ldots\ldots\beta_i x_{ip}+ \varepsilon_i=[X_i]^T\beta + \varepsilon_i i= 1$to n. where T denotes the transpose so that $[X_i]^T\beta$ is the inner product between x_i and β

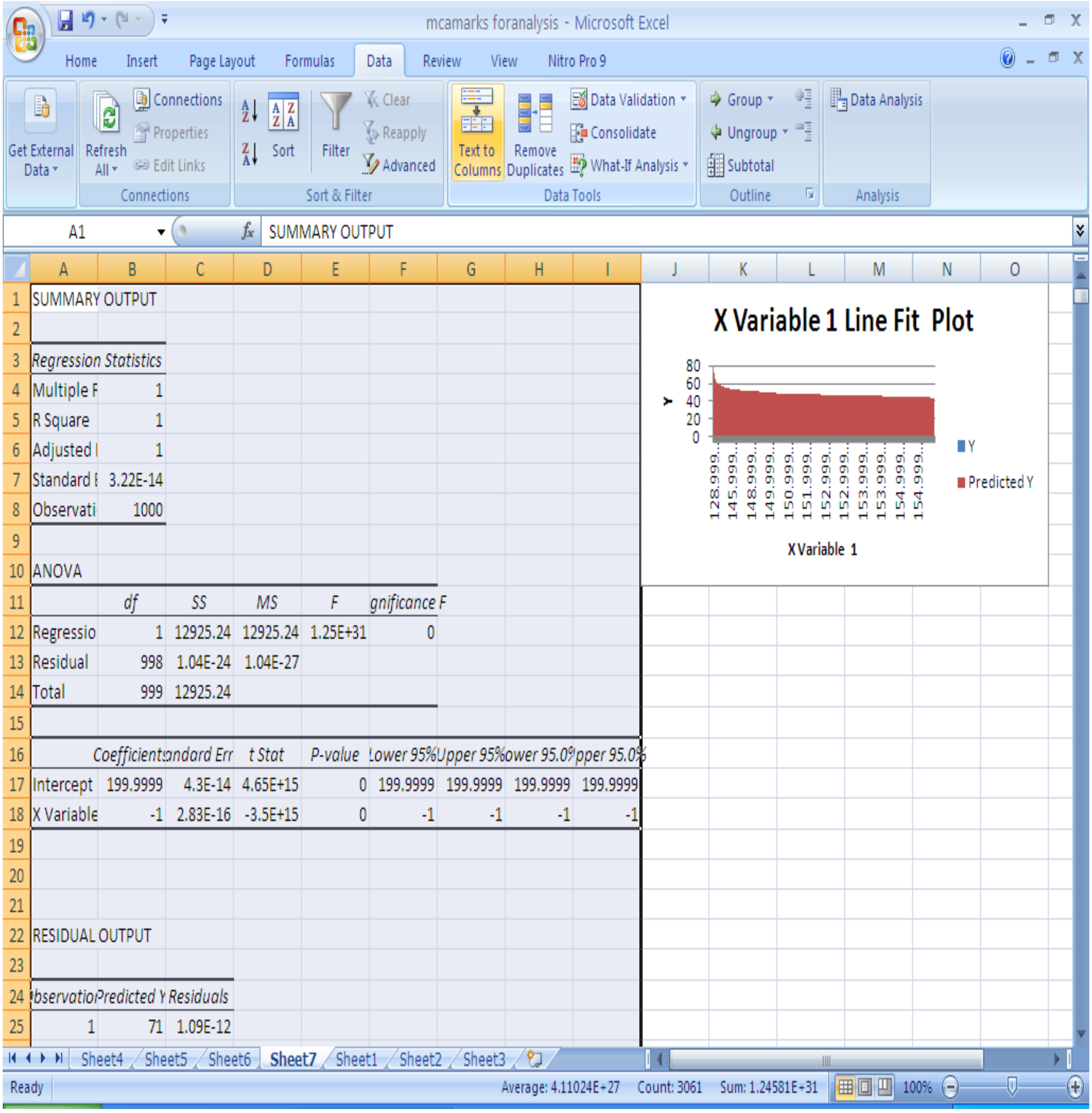
This model can take the compact form

$Y = x\beta + \varepsilon$ where,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

IV. The Context

Karnataka -PGCET examination –The area of study for this paper is the study of the marks distributions and considerations of the dependent and independent factors in the Karnataka PGCET examination. InKarnataka , seats for the Post graduate Courses in Engineering, MCA and MBA are conducted by Karnataka examination Authority-(This year it has been changed).The examination is taken by nearly a lakh students and the rank is allocated to them based on their score in the entrance examination –Previously the ranks were decided by considering the students' marks in their respective degrees also-equal weightage was being given to both the marks. Now a study of the dependency of the ranks upon these is made .For regression analysis , the entrance examination marks field is considered to be the dependent variable and the degree marks is considered to be the independent variable. A sample output from Excel for this case is given below. The data is taken from the MCA results and the number of data points is 1000.The actual data consists of only the entrance test marks but for the analysis, the degree marks is derived by taking into consideration the equal weightage given to both the marks.This has been generated by running a formula involving the rank and the entrance test marks.



SUMMARY OUTPUT

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	1							
5	R Square	1							
6	Adjusted I	1							
7	Standard E	3.22E-14							
8	Observati	1000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressio	1	12925.24	12925.24	1.25E+31	0			
13	Residual	998	1.04E-24	1.04E-27					
14	Total	999	12925.24						
15									
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	199.9999	4.3E-14	4.65E+15	0	199.9999	199.9999	199.9999	199.9999
18	X Variable	-1	2.83E-16	-3.5E+15	0	-1	-1	-1	-1
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	*bservatio*	*Predicted Y*	*Residuals*						
25	1	71	1.09E-12						

Ready Average: 4.11024E+27 Count: 3061 Sum: 1.24581E+31

V. Regression Output Explained

The regression output shows both the R^2 and the adjusted R^2 values are equal to 1- which shows that there is perfect correlation between the two sets of marks. In reality this would not be possible because of the different kinds of scoring patterns practiced in different universities and colleges and they cannot be reduced to one single uniformscale. The observed results are due to the fact that the degree marks is generated from the program. The degree marks is generated taking into consideration that there is a 50% component of the degree marks and the entrance test marks and evolving a suitable formula to generate them by using the rank and the entrance test marks. But the marks is indeed following the given pattern which can be verified by observing the rank field. There is a difference of not more than 1 mark in the entrance test for the current and the next rank. Therefore, effectively the marks distribution follows the applied pattern. The same applies to the regression equation which can be taken from the output. It reads

$$Y = 199.99 - X$$

The confidence interval is set to 95% which means that the rejection region is only 5% and the predicted Y value falls within that range only.

This technique can be further extended to contain more explanatory variables where in multiple regression can be used. The analysis done hitherto considers the dependency only upon the degree marks but other causal factors like peer pressure, test anxiety and so on.

VI. Deciding About The Questions In The Entrance Examination

When this output is subjected to a more thorough analysis going into details of why a particular student group has scored a particular set of ranks- the questions in the entrance examination have to be analyzed. These have to be analyzed with a perspective of the aptitude of the students taking up the examination. The nature of the questions in the degree examinations and those in the entrance examination has to be matched in order to bring about a suitable testing of the aptitude. The regression outputs give us a quantitative measure of the impact of the marks of the students – which can be used to derive to what extent the marks is determining the aptitude of the student. The same degree marks in the future examinations can be used with these coefficients to predict what could be the marks in the entrance examination that a particular student community would score-Of course , a stringent study of the types of questions has to be done to determine the aptitude of the students in a particular area of the questions.

The questions in the entrance examination can also be categorized, each one testing a particular aspect of the student ability. The post graduate entrance exams are expected to test the students' ability in quantitative ability, reasoning and so on. By studying the pattern of marks in the entrance examination and the groups given by the regression output, one can decide what kind of questions best measure the aptitude of the student groups.With a perspective of what is required to pick up the right kind of candidates , this information could be used to decide about the questions .The analysis of marks , combined with the comparisons of the questions and answers in both the examinations , this could be used as a predictive technique in deciding the right question needed for the current year.

VII. Limitations

Linear regression, as has been explained here is testing only one aspect of the scenario. The causal factors for a student scoring a certain marks are many, including peer pressure, test anxiety, future expectations , motivation and so on. Some of these are easy to be quantified some and some are not.Linear regression is best useful when there is a linear relationship between the dependent and the independent variables.Better techniques like genetic algorithms and partial least square methods can be used for better analysis.

VIII. Conclusions

This paper discusses a mathematical model to study the marks distributions in the Karnataka PGCET examinations. The technique of linear regression can be applied to a large sets of records also and this is where the usefulness of the method can be seen. As the number of students taking up the examination is very large , deriving a knowledgeable pattern of marks distribution becomes essential when a change of question paper is the issue. Other probabilistic techniques and statistical techniques can be used along with linear regression to better know of the clusters of the students and devise the question papers accordingly.

Acknowledgements

The authors thank the support of Jain University and R.V.College Autonomous in producing this paper.

REFERENCES

[1] http://www.mcser.org/journal/index.php/jesr/article/viewFile/186/171.
[2] http://www20.csueastbay.edu/ecat/graduate-chapters/g-stat.html