International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

Survey on College Admission using Regression Analysis in Excel

G.Hari Priya, G.Sivaranjani, N.Suruthi, R.Susithra
Department Of Information Technology,
Thiagarajar College of Engineering, Madurai.

Abstract- *Statistics plays important role in education domain. Every year student's area of interest is varying in choosing branches of engineering .This change is caused mainly due to the placements happened previous year in the college. This paper uses linear regression model to find how well the placements affect the interest shown to the departments by the students .This paper uses visualization techniques to support the aim. This is done with the help of previous year placement details given by the college .The ultimate aim of the paper is order branches based on expected priority. Microsoft excel with its various facilities is used to achieve this .Our survey will be helpful for the students to understand the trend especially those who are in rural areas and has no idea about which branch to take.*

I. INTRODUCTION

Regression -Regression is a statistical measure used in finance, education, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. In all cases, a function of the independent variables called the **regression function** is to be estimated. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution

It is used for prediction .Classification is for predicting class labels. While regression is for predicting the values. It is done with the help of regression equation.It is the function of independent variable with the dependent variable on the left hand side. There are many methods to find this equation. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

. The two basic types of regression are linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis. Linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple regression uses two or more independent variables to predict the outcome .This paper proves that admission in each department can be proved by using linear regression model with the dependent variable as admission and independent variable s placement in each department.

For this scatter plot is used to prove that regression can be used for this purpose.

Simple linear regression-Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. One variable, denoted *x*, is regarded as the predictor, explanatory, or independent variable. The other variable, denoted *y*, is regarded as the response, outcome, or dependent variable. For that the two variables must be correlated. Positive correlation is one in which the dependent increases as independent variable increases and vice versa. Negative correlation is that in which the dependent variable decreases as the independent variable increase and vice versa.

Best fitting Line-A line of best fit (or "trend" line) is a straight line that best represents the data on a scatter plot. This line may pass through some of the points, none of the points, or all of the points.

$y = ax + b$

- **y** is the response variable.
- **x** is the predictor variable.
- **a** and **b** are constants which are called the coefficients.

Method-Least squares method is one of the method to find this equation.It uses normal equations to find the coefficients a and b. It reduces the sum of squared residuals. Residuals is nothing but the difference between the observed and expected value.

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

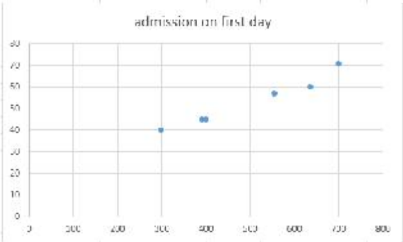
$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i$$

These are the two normal equations to find the coefficients. This paper also use this method.

II.PROCEDURE

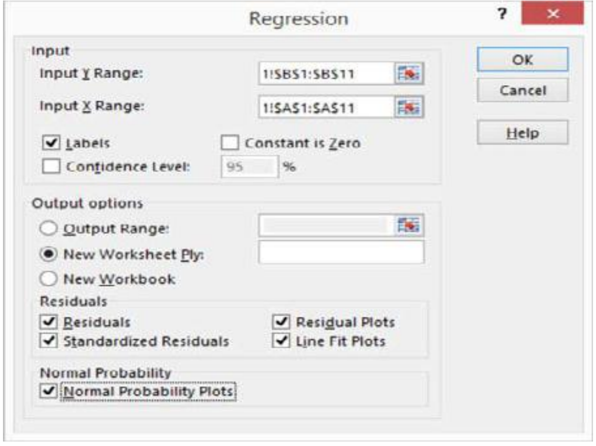
Tools in Excel –Some tools for data analysis available in Microsoft Excel is discussed here:

Scatter plot - Scatter charts and line charts look very similar, especially when a scatter chart is displayed with connecting lines. However, there is a big difference in the way each of these chart types plots data along the horizontal axis (which is also known as the x-axis) and the vertical axis (which is also known as the y-axis). The following is the scatter plot for admission on first day .The points are almost in straight line proves that the linear regression model can be used to predict admission based on placement details



Regression analysis – The data analysis option is there in excel. On clicking this a dialog box opens with various types of topics in statistics. In that one of the option is

Regression. On clicking that the following dialog box
opens.



Input
Input Y Range: 1!$B$1:$B$11
Input X Range: 1!$A$1:$A$11
☑ Labels
☐ Constant is Zero
☐ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

Residuals
☑ Residuals
☑ Residual Plots
☑ Standardized Residuals
☑ Line Fit Plots

Normal Probability
☑ Normal Probability Plots

Data analysis is added as extra-plugins if not available.

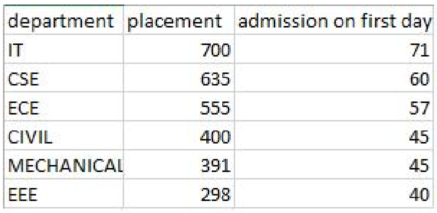
Use the Input Y Range text box to identify the worksheet range holding your dependent variables. Then use the Input X Range text box to identify the worksheet range reference holding your independent variables. Each of these input ranges must be a single column of values.

If the regression line should start at zero — in other words, if the dependent value should equal zero when the independent value equals zero — select the Constant Is Zero check box.

To calculate a confidence level in regression analysis

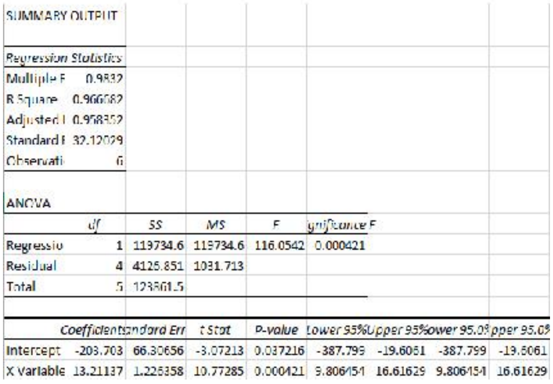
Select the Confidence Level check box and then enter the confidence level you want to use. Use the Output Options radio buttons and text boxes to specify where Excel should place the results of the regression analysis. To place the regression results into a range in the existing worksheet. Select from the Residuals check boxes to specify what residuals results you want returned as part of the regression analysis. Similarly, select the Normal Probability Plots check box to add residuals and normal probability information to the regression analysis results.

Example-Consider the following data.



department	placement	admission on first day
IT	700	71
CSE	635	60
ECE	555	57
CIVIL	400	45
MECHANICAL	391	45
EEE	298	40

Output-



SUMMARY OUTPUT								
Regression Statistics								
Multiple F	0.9832							
R Square	0.966682							
Adjusted I	0.958352							
Standard I	32.12029							
Observati	6							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressio	1	119734.6	119734.6	116.0542	0.000421			
Residual	4	4126.851	1031.713					
Total	5	123861.5						
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-203.703	66.30656	-3.07213	0.037216	-387.799	-19.6061	-387.799	-19.6061
X Variable	13.21137	1.226358	10.77285	0.000421	9.806454	16.61629	9.806454	16.61629

Evaluation of results – Part I –Regression Statistics-There are **‘**goodness of fit measures’.

Multiple R. This is the correlation coefficient. It tells us how strong the linear relationship is. For example, a value of 1 means a perfect positive relationship and a value of zero means no relationship at all. It is the square root of r squared.

R squared. This is r^2, the Coefficient of Determination. It tells us how many points fall on the regression line. For example, 80% means that 80% of the variation of y-values around the mean are explained by the x-values. In other words, 80% of the values fit the model.

Adjusted R square. The adjusted R-square adjusts for the number of terms in a model.

Standard Error of the regression: An estimate of the standard deviation of the error μ. This is *not* the same as the standard error in descriptive statistics! The standard error of the regression is the precision that the regression coefficient is measured; if the coefficient is large compared to the standard error, then the coefficient is probably different from 0.

Observations. Number of observations in the sample.

Part –II-Anova- SS = Sum of Squares.

Regression MS = Regression SS / Regression degrees of freedom.

Residual MS = mean squared error (Residual SS / Residual degrees of freedom).

F: Overall F test for the null hypothesis.

Significance F: The significance associated P-Value.

Part-III- Interpret regression results - Coefficient: Gives you the least squares estimate. Standard Error is the least squares estimate of the standard error .P Value gives us the p-value for the hypothesis test.Lower 95% is the lower boundary for the interval. Upper 95% is the upper boundary for the interval. Finally the aim of this paper the regression equation is found from the output displayed.

The most useful part of this section is that it gives you the linear regression equation-y = mx + b.

y = slope * x + intercept.
For the above table, the equation would be approximately:
y = 13.2113*x-203.703.This equation can be used to predict the current year admission based on placements happened previous year .

Conclusion-We can see that the value of R squared is close to 1.This implies that the generated equation is the best model that can fit the maximum possible points. So that this method can be used with previous to generate the equation. This equation can be used to generate the current year admission details. The values will be close to the original value. The value between the original and predicted value will be less. Hence this paper proved that the admission details can be predicted from placement details.

References-This paper takes help from following references.

[1] http://cameron.econ.ucdavis.edu/excel/ex61multipleregression.htmlJ.

[2] The Relationship Between Summer Weather and Summer Loads - A Regression Analysis

[3] Excel Data Analysis For Dummies, 2nd Edition

[4] “Introduction to data mining” by Tan, Steinbach & Kumar (2006)

[5] Data Mining: Concepts and Techniques, Third Edition by Han, Kamber & Pei (2013)

[6] Data Mining and Analysis Fundamental Concepts and Algorithms by Zaki & Meira (2014)

[7] Data Mining: The Textbook by Aggarwal (2015)

[8] “The Elements of Statistical Learning” by Freidman et al (2009)