Prediction Analysis on Student Library Usage: A Case Study at Eastern Mediterranean University (EMU), Gazimagusa, Turkey.

Ademola E. Babalola, M.Sc.

Eastern Mediterranean University, Gazimagusa, TRNC, Turkey.

E-mail: babalola.ea@gmail.com

ABSTRACT

This paper highlights the importance of statistics in everyday life in the 21st Century. Statistics, which is based on collecting, managing, processing and disseminating information, and it applications are seen in banks, airports, information technology, and schools.

This research presents a simple linear regression model and how it is used to analyses data obtained from the school library of Eastern Mediterranean University. The data defines the average population of students who use the library yearly. The model obtained is used to predict future use of the library by students, as well as estimate students that used the library prior to the year 2007. A statistical software application, called SPSS was used in the analysis. Discussions will also be made on the scatter plot to know outliers effect.

The result obtained from this research shows about 33% of data can be analyze which makes our model somewhat a good fit for the data; however, a non-linear model would be best to describe the library data**.**

(Keywords: linear regression models, sample data, correlation coefficient, influential point, SPSS, EMU, Eastern Mediterranean University, library usage)

INTRODUCTION

Scientific inquiry is an iterative learning process. Its aims, pertaining to the explanation of a social or physical phenomenon, must be specified and tested by gathering, organizing, numerating, and analyzing data.

"Population" is defined as a set of objects or units that are of interest to study, such as bottles filled up in a day a cola drink company.

"Sample" is simply some or subset of the actual population such as bottles filled for the first hours at a cola company.

A study of elements in a population is called a census; however the study of a census is almost impossible because of time, cost, and method/means of measurement. Data in a population are studied by taking samples from the population and every property observed from the samples is generalized for the population. Samples not reflecting the population are referred to as biased.

"Variables" are often mentioned in statistical analysis; they define or are seen as characteristic changes between objects in a population. They are often regarded as data [2]. They occur in two categories: qualitative variables/data and quantitative variables/data.

Qualitative data are data variables are assigned values such as name, color, or labels. They are also regarded as categorical variable. Quantitative data are numeric. They simply are used in measurable quantities such as the population of students at the university. The number of students is a measurable attribute which makes the population a quantitative variable.

Statistics basically deals with Data analysis obtained from a given population or randomly selected samples. It uses some data analysis parameter called descriptive statistics and inferential statistics.

According to [10] descriptive statistics is different from inferential statistics. Descriptive statistics describes what the data shows, they are quantitative in nature and are used in a manageable form. It is sometimes called exploratory data analysis, and hence can also be define as investigating the measurements of the variables in data sets. These are sometimes relationship between variables or individual variables.

Processed data refers to information, or raw data, obtained from an experiment or an historical record. If we are analyzing student ID cards, for example, descriptive statistics describes the percentage of ID cards issued to students at the University each semester, or the birth certificate issued at the General Hospital Magusa. Any random numbers chosen for computation are descriptive statistics for the data from which the statistic is computed. Descriptive statistics are used at one time to give a full picture of the data.

Inferential Statistics, on the other hand, are concerned with making conclusions about a population from a sample. This is done by random sampling, followed by inferences made about the number of distribution.

METHODOLOGY

A careful study was carried out to determine the average population of students using the EMU library, yearly, starting from the 2007/2008 session to 2014/2015 session. A histogram chart as already been constructed, to highlight the level of improvement according to records obtained from the library managements [histogram not included]. The information was obtained through a clocking system which all students must pass through with a valid identification to properly ascertain number of students entering the library. Data were obtained every day except weekends (i.e., Saturday and Sunday). This was done to generate data monthly and then yearly (session) for an average value yearly (Table 1).

Development of Regression a Model

A simple regression seeks to find the relationship between variables 'Y' and 'X'. It is very important in every area of life as it is used for future predictions.

"Simple" means that we are working in two dimensions. Mathematically a straight line equation is given [9]:

$$Y = ß_0 + ß_1X\ldots. \quad \textbf{(1)}$$

Where,
$ß_0$ = intercept on the response axis,
$ß_1$ = slope of the straight line, $\varepsilon = 0$ = error

Then; Y= random variable whose value change with X, with errors
and X = is an independent variable with negligible error.

Hence our regression equation becomes:

$$E(y) = ß_0 + ß_1 X\ldots. \quad \textbf{(2)}$$

$$Y = E(Y) + \varepsilon \ldots\ldots \quad \textbf{(3)}$$

Equation (2) is a true regression line with data's scattered around the line, for our estimated regression:

$$y = b_0 + b_1 x \ldots\ldots \quad \textbf{(4)}$$

$$b_0 = \Sigma Y_i / n - b_1 \Sigma X_i / n$$

$$b_1 = n\Sigma Y_i x_i - \Sigma Y_i \Sigma x_i \ / \ (n\Sigma x_i^2 - \Sigma x_i \Sigma x_i)$$

Data Collection and Analysis

The collection of data actually began in year 2007/2008 through to 2014/2015. An average value is obtained for each session and recorded. Table 1, presents the variables and Figures 1 and 2 gives the SPSS output.

Table 1: Data Obtained from Library.

Year/session	Year since 2007 (X)	Average population (Y)
2007/2008	7	1742
2008/2009	8	1750
2009/2010	9	1568
2010/2011	10	1393
2011/2012	11	1697
2012/2013	12	1791
2013/2014	13	1744
2014/2015	14	2611

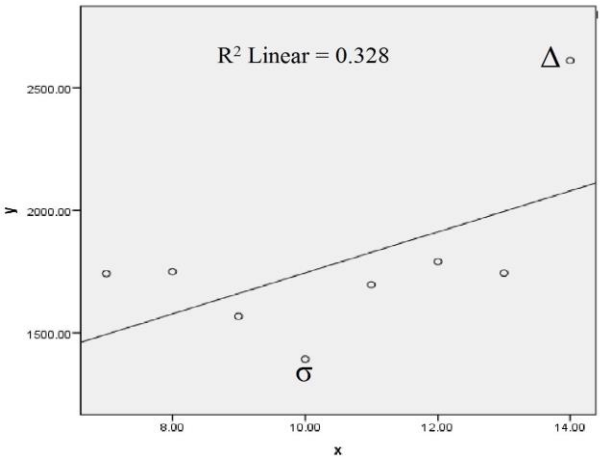

Figure 1: SPSS Output on Data Obtained from Library.

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	908.750	525.144		1.730	.134
x	83.643	48.864	.573	1.712	.138

Figure 2: Regression Coefficient

Hence regression equation becomes:

$\underline{Y} = 908.750 + 83.643\underline{x}$ …. **(5)**

This equation can be used to predict future average population to be expected in the next four years as well as four years prior to 2007 session, anything outside this range becomes void. This is as a result of a parameter called Quartiles. Quartiles helps to know limit which interpolation must not exceed, it also helps to identify outliers but for this paper, the data collated does not go beyond its extreme value.

Table 2, shows variable X and Y, and the estimates of $\underline{Y}$ using equation (4), for sessions 2015/2016, 2016/2017, and 2017/2018 are respectively estimated to have 2163, 2247 and 2331 students for the coming sessions. This however just shows a light increase in library population. Notice Fgure 1 with two symbols "σ" and "Δ", this are observed outliers. Figure 3 is reproduced with SPSS without "σ" and a better "R^2" was obtained.

"R^2" is the correlation coefficients which explains the degree which variables are related, comparing Figure 1 and Figure 3, there is a difference in "R^2" value.

Hence, one may conclude that the fourth variable may be recorded in error as it had a bad effect on the regression line.

Table 2: Data Comparison between Real "Y" Variables and Estimated "Y".

X "years or session"	Y "students"	Estimates $\underline{Y}$
7	1742	1494
8	1750	1598
9	1568	1662
10	1393	1745
11	1697	1828
12	1791	1912
13	1744	1996
14	2611	2079

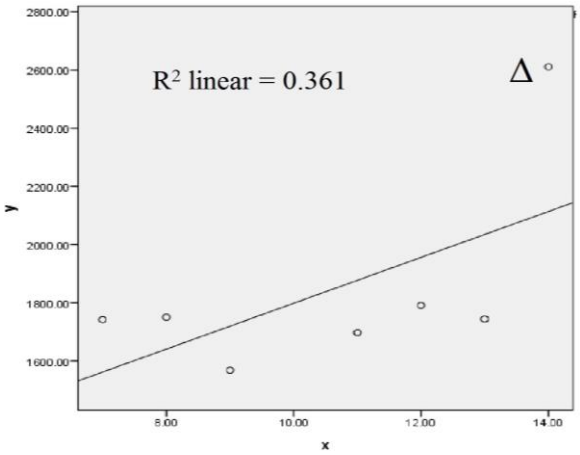

Figure 3: SPSS Output without "σ".

CONCLUSION

Simple regression is an important tool in applied statistics, but it remains pedagogically neglected and controversial. However in light of data analysis from the library, R^2 obtained shows approximately 33% of the data can be explained and hence makes the model somewhat a best fit.

The least line equation in Equation (5), can be used to predict future average population of students that will use the library for the next three session; respective sessions of 2015/2016, 2016/2017 and 2017/2018 shows an estimate of 2163, 2247 and 2331, and no radical increase for each session but a gradual Increase.

In Figure1, point "σ" does have a bad influence on the least line. Figure 3 gives a better R^2 value when the bad influence point was removed. Data analysis in this paper is best analyzed using a non-linear model.

REFERENCES

1. Dallal, G.E. 2014. "Regression Diagnostics". Retrieved from http://www.tufts.edu/gdallal/

2. Faraway, J.J. 2002. Practical Regression and Anova using R. http://www.stat.lsa.umich.edu/˜faraway/book

3. Olive. D. 2006. "Applied Robust Statistics. Preprint M-02-006". Retrieved from http:www.math.siu.edu/

4. Williams, R. 2015. "Outliers Review". Retrieved from http://www3.nd.edu/~rwilliam/

5. Rousseeuw, P.J. and A.M. Leroy. 2003. *Robust Regression and Outlier Detection*. J .Wiley and Sons: Princeton, NJ. ISBN 0-471-48855-0.

6. Singer, T. 2004. *Science*. retrieved from http://www.sciencemag.org (page 6)

7. Wilcox. R.R. 2001. *Fundamentals of Modern Statistical Methods*. Springer: New York, NY. ISBN 0-387-95157-1.

8. Warren, M. 2014. "Regression with SAS". University of California. Retrieved from http://www.ats.ucla.edu/stat/sas

9. Mendenhall, W. and T.T. Sincich. 2011. *A Course in Statistics Regression Analysis*. Seventh Edition. Pearson: London, UK.

10. Trochim, W.M.K. and J.P. Donnelly. 2008. Research Method Knowledge Based". Atomic Dog Publishing: Manson, OH.

ABOUT THE AUTHOR

Ademola Babalola, is a Research Assistant at Eastern Mediterranean University, and a graduate member of IEEE. He holds a B.Sc. in Computer Engineering and M.Sc. in Applied Mathematics and Computer Science. His research interests are in computer networks, network simulation, applied statistics, data mining, and computer architecture.

SUGGESTED CITATION

Babalola, A.E. 2016. "Prediction Analysis on Student Library Usage: A Case Study at Eastern Mediterranean University (EMU), Gazimagusa, Turkey". *Pacific Journal of Science and Technology*. 17(1):152-155.

Pacific Journal of Science and Technology