Volume 9, Issue 03, March 2024

ISSN (online): 2456-0006

IJSTMR

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Analyzing Effect of Predictors Variables on Target Variable Using Regression

Vivek Bhargaw
M Tech (CSE) 4th semester
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus Indore M.P.

Mr Nilesh Avinash Joshi
Assistant Professor
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus Indore M.P.

Abstract: ***In the proposed work we have taken data of advertisement and sales. Advertisement on YouTube, Facebook and WhatsApp and we want to know the effect on sales. Based on sales and advertiser various media (YouTube, Facebook and WhatsApp) we found that sales has been increased during advertisement on TV as compared to other two media To predict future sales price. We used Multiple Linear Regression. We have the following problems. How to select following variables Dependent Variable, Independent Variable(s), Intercept, Coefficients. How the model Select data from the past to learn what's the relationship. How minimize the predicted error for regression analysis. We need to check that a linear relationship exists between the dependent variable and the independent variables We want to analyze the linear relationship between independent variables (YouTube, Facebook and WhatsApp) exists or not. Our second objective is to check that a linear relationship either exists: Sales and advertisement on YouTube both have positive are negative relationship, Sales and advertisement on Facebook both have positive are negative relationship, Sales and advertisement on WhatsApp both have positive are negative relationship. Predicted values and observed values should have minimum error. It is also called the residuals.***

Keywords: YouTube, Facebook, WhatsApp, linear Regression , Advertisement, Predicted

I. Predictive Modeling

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables is assumed to be an affine function of those values; less commonly, the conditional median or some other quintile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response[8,10].

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regressions. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

II. Motivation Toward Regression Analysis

Regression analysis is a statistical technique for analyzing and comprehending the connection between two or more variables of interest. The methodology used to do regression analysis aids in understanding which elements are significant, which may be ignored, and how they interact with one another. Regression is a statistical approach used in finance, investment, and other fields to identify the strength and type of a connection between one dependent variable (typically represented by Y) and a sequence of other variables (known as independent variables). Regression is essentially the "best guess" at utilizing a collection of data to generate some form of forecast. It is the process of fitting a set of points to a graph. Regression analysis is a mathematical method for determining which of those factors has an effect. It provides answers to the following questions:

- Which factors are most important
- Which of these may we disregard

- How do those elements interact with one another, and perhaps most significantly, how confident are we in all of these variables

These elements are referred to as variables in regression analysis. We have dependent variable, which is the key aspect attempting to understand or forecast. Then there are your independent variables, which are the elements you assume have an effect on dependent variable. There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the significant relationships between dependent variable and independent variable.
2. It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

1. Simple Linear Regression
2. Multiple Linear Regression
3. Non Linear Regression

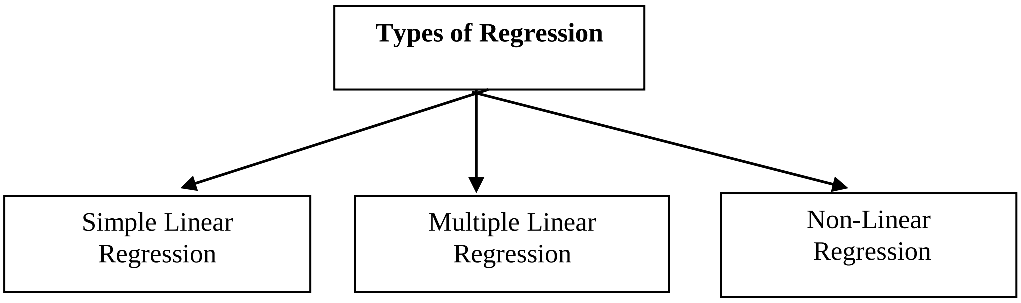

Figure 1 Types of regression

III. R-SQUARED AND THE GOODNESS-OF-FIT

R-square is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

After fitting a linear regression model, we need to determine how well the model fits the data. Does it do a good job of explaining changes in the dependent variable? There are several key goodness-of-fit statistics for regression analysis. Linear regression identifies the equation that produces the smallest difference between all of the observed values and their fitted values. To be precise, linear regression finds the smallest sum of squared residuals that is possible for the dataset. Statisticians say that a regression model fits the data well if the differences between the observations and the predicted values are small and biased. Unbiased in this context means that the fitted values are not systematically too high or too low anywhere in the observation space.

However, before assessing numeric measures of goodness-of-fit, like R-squared, we should evaluate the residual plots. Residual plots can expose a biased model far more effectively than the numeric output by displaying problematic patterns in the residuals. R squared (R^2) value in machine learning is referred to as the coefficient of determination or the coefficient of multiple determination in case of multiple regression. R squared in regression acts as an evaluation metric to evaluate the scatter of the data points around the fitted regression line. It recognizes the percentage of variation of the dependent variable.

R-squared is the proportion of variance in the dependent variable that can be explained by the independent variable.

$$R^2 = \frac{Variance\ explained\ by\ the\ model}{Total\ Variance}$$

- *Interpret R squared*

  We have a visual demonstration of the plots of fitted values by observed values in a graphical manner. It illustrates how R-squared values represent the scatter around the regression line.

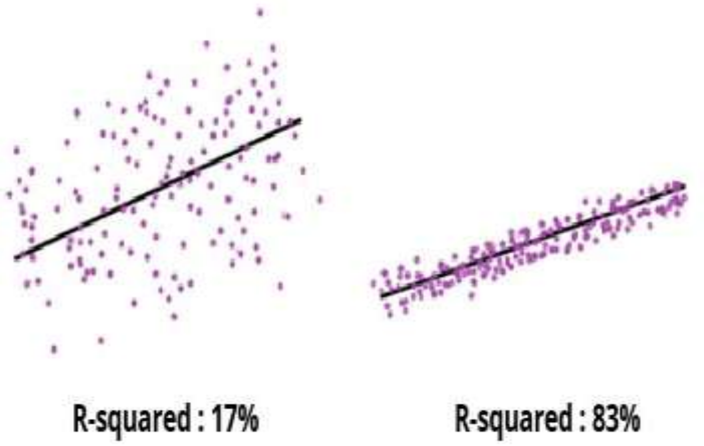

Figure 2 represent the scatter around the regression line.

R^2=1 (100%) All the variation in the y values is accounted for by the x values

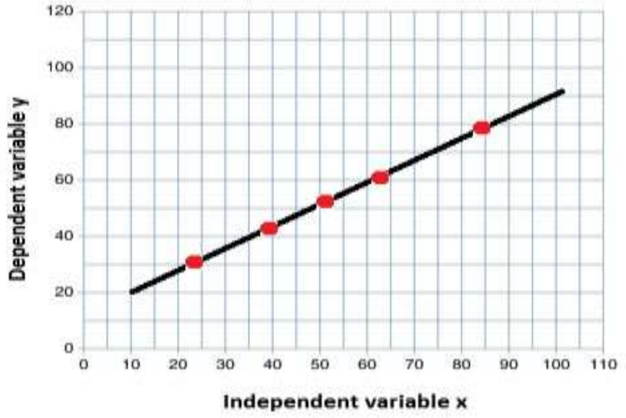

Figure 3 100% variation in the y values is accounted for by the x values.

R^2=0.83(83%) of the variation in the y values is accounted for by the x value

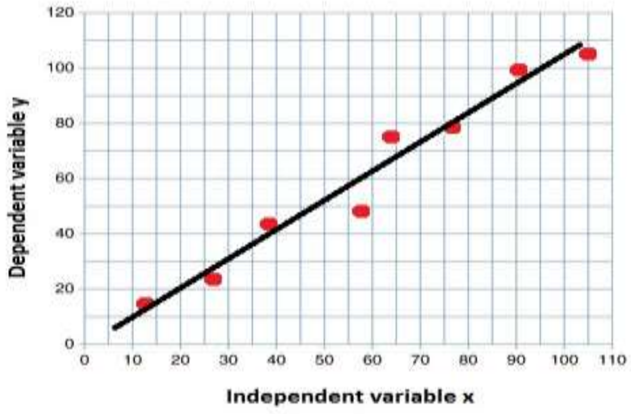

Figure 4 80% of variation in the y values is accounted for by the x values.

R^2=0(0%) None of the variation in the y values is accounted for by the x values

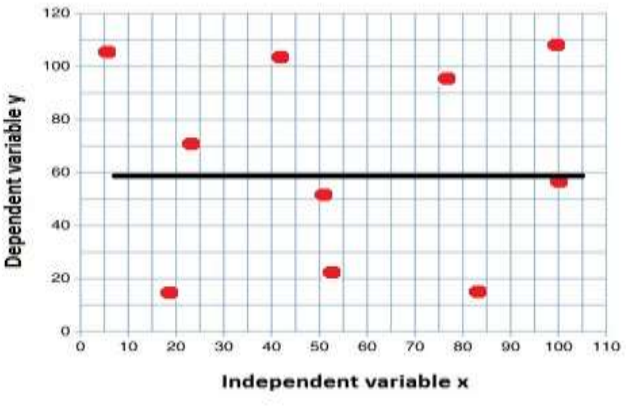

Figure 5 None of variation in the y values is accounted for by the x values

IV. PROBLEM STATEMENT

Multiple Linear Regression analysis is the study of more than two variables to find a relationship, or correlation. A regression line is a straight line that attempts to predict the relationship between two points, also known as a trend line or line of best fit. Multiple Linear Regression is a prediction when a variable (*y*) is dependent on more than one independent variable (*x*) based on the regression equation of a given set of data.

In the proposed work we have taken data of advertisement and sales. Advertisement on YouTube, Facebook and WhatsApp and check the effect on sales.. Based on sales and advertiser various media (YouTube, Facebook and WhatsApp) we found that sales has been affected during advertisement on YouTube as compared to other two media To predict price by these two factors we used Multiple Linear Regression.

We have the following problems [16,17,18].

1. It is difficult to decide following variables.
   - Predictors Variable
   - Response Variable
   - Intercept
   - Coefficients
2. It is difficult to Select data from the past to learn the relationship.
3. It is difficult to minimize the predicted error for regression analysis.
4. We need to check that a linear relationship exists between the dependent variable and the independent variable/s
5. We need to check that a linear relationship exists between the:
   Sales and advertisement on YouTube both have positive are negative relationship.
   Sales and advertisement on Facebook both have positive are negative relationship.
   Sales and advertisement on WhatsApp both have positive are negative relationship.

V. LITERATURE SURVEY

In 2015 Supichaya Sunthornjittanon et al proposed “Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand.” They analyze the ABC Company’s data and verify whether the regression analysis methods and models would work effectively in the ABC Company based in Bangkok, Thailand. After the data are collected, models are created to examine the contribution of each of the company’s financial factors to the net income of the company. The final model is selected using Stepwise Regression Methods. A linear regression line and equation for the model are generated to help observe and predict future trends[2].

In 2016 Sandhya Jain et al proposed “Regression Analysis – Its Formulation and Execution in Dentistry”. Prediction and estimation is the mainstay in the treatment planning in dentistry. With variations being common is many events of the oral cavity, it becomes important to have a methodology which can help us predict the happenings of the region in relation to each other. Regression analysis is one such concept which explores the relationship between two or more quantifiable variables so that one variable can be predicted from other. They provide a simple yet holistic approach to the understanding of the concepts of Regression Analysis along with its use and misuse, advantages and disadvantages pertaining to the art and science of dentistry in the formulation and execution of a dental treatment plan, the variables involved in the decision making are often poorly characterized and incompletely validated. For these reasons we have to rely on the mean values or go for a wild guess[3].

In 2017 Radek Silhavy et al proposed “Analysis and selection of a regression model for the Use Case Points method using a stepwise approach”. They investigate the significance of use case points (UCP) variables and the influence of the complexity of multiple linear regression models on software size estimation and accuracy. Stepwise multiple linear regression models and residual analysis were used to analyze the impact of model complexity. The impact of each variable was studied using correlation analysis. The estimated size of software depends mainly on the values of the weights of unadjusted UCP, which represent a number of use cases. All other variables (unadjusted actors’ weights, technical complexity factors, and environmental complexity factors) from the UCP method also have an impact on software size and therefore cannot be omitted from the regression model [4].

In 2017 N J Gogtay et al proposed “Principles of Regression Analysis”. Regression analysis is a statistical tool that helps evaluate relationships between a dependent variable and one or more independent or predictor variables. It helps to understand how the dependent variable changes with changes in the independent variable and thus finds its application in forecasting and predicting. The technique must however be used with clear understanding of the assumptions in each type of regression analysis, their limitations and the potential error that can occur when models are applied to a larger population. They apply this equation to the population for making a prediction, and able to predict either the systolic blood pressure perfectly. They need to take into account an “error” or “deviation” that is likely to occur when this equation is used[5] .

In 2018 Ira Sharma et al proposed “Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal”. Their aims to assess the factors associated with carbon stock in Chure forest of Nepal. The data were obtained from Department of Forest Research and Survey (DFRS) of Nepal. A multiple linear regression model and then sum contrasts were used to observe the association between variables such as stem volume, diameter at breast height, altitude, districts, number of trees per plot, and ownership of the forest. 95% confidence interval (CI) plots were drawn for comparing the adjusted carbon stocks with each of the factors and with the overall carbon stock. The linear regression showed a good fit of the model (adjusted $R2$ = 83.75%) with the results that the stem volume (sv), diameter at breast height (dbh), and the number of trees per plot showed statistically significant (p value ≤ 0.05) positive association with carbon stock[6].

In 2019 Anjali Pant et al proposed “Linear Regression Analysis Using R for Research and Development” .The future forecasting opportunities and risks estimation are the most prominent prerequisite for a successful business. Regression analysis can go far beyond forecasting. The linear regression analysis technique is a statistical method that allows examining the linear relationship between two or more quantitative variables of interest. The rationale of the linear regression analysis technique is to predict an outcome based on historical data and finding a linear relationship. They discussed the implementation of linear regression using a statistical computing language R and consider that the suggested approach provides an adequate interpretation of research and business data. Introduction Software[7].

In 2020 Khushbu Kumari et al proposed “Linear Regression Analysis Study”. Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. They explain the basic concepts and explain how we can do linear regression calculations in SPSS and excel. The techniques for testing the relationship between two variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation [8].

In 2020 Samit Ghosal et al proposed “Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th, 2020)”. They analysis tracing a trend related to death counts expected at the 5th and 6th week of the COVID-19 in India. Material and methods: Validated database was used to procure global and Indian data related to coronavirus and related outcomes. Multiple regression and linear regression analyses were used interchangeably. Since the week 6 death count data was not correlated significantly with any of the chosen inputs, an auto-regression technique was employed to improve the predictive ability of the regression model [9].

VI. PRAPOSED APPROACH

Following is a list of 7 steps that could be used to perform multiple regression analysis.

1. Identify a list of potential variables/features; Both independent (predictor) and dependent (response). Gather data on the variables.
2. Check the relationship between each predictor variable and the response variable. This could be done using scatterplots and correlations.
3. Check the relationship among the predictor variables. This could be done using scatterplots and correlations. It is also termed as multi collinearity test.
4. Try and analyze the simple linear regression between the predictor and response variable.
5. Use the non-redundant predictor variables in the analysis. This is based on checking the multi collinearity between each of the predictor variables. If the correlation exists, one may want to one of these variables.
6. Analyze one or more model based on some of the following criteria.
7. Use the best fitting model to make prediction based on the predictor (independent variables). This is done based on the statistical analysis of some of the above-mentioned statistics such as t-score, p-value, R squared, F-value et

VII. COMPARATIVE ANALYSIS

Checking Relationship between Advertising on YouTube and Sales using Scatter

```
data.plot(kind='scatter', x=['YouTube'], y='Sales')
sns.pairplot(data, x_vars=['YouTube'], y_vars='Sales', height=5, aspect=1, kind='reg')
```

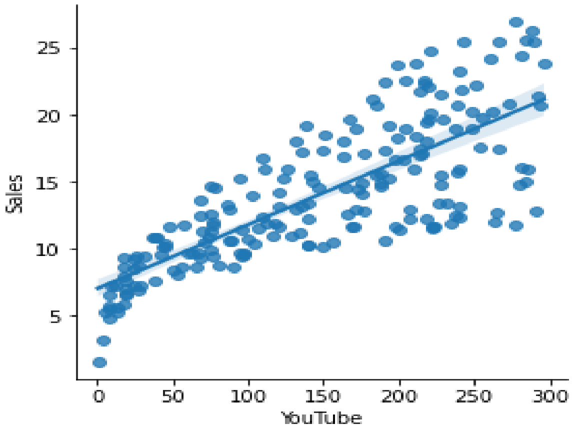

Figure 6 Relationship between advertisement on YouTube and Sales

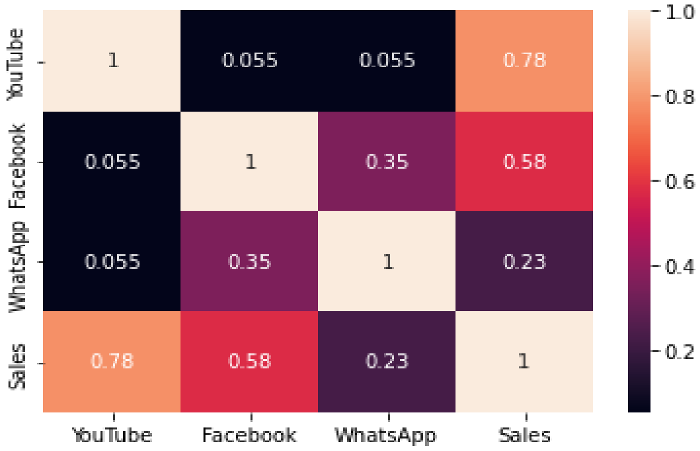

Figure 7 Heatmap to show relationship between advertisement on YouTube and sales

CONCLUSION

Multiple linear regression is a mathematical technique that deploys the relationship among multiple independent predictor variables and a single dependent outcome variable. In the proposed work we have taken data of advertisement and sales. Advertisement on YouTube, Facebook and WhatsApp and check the effect on sales. Based on sales and advertiser various media (YouTube, Facebook and WhatsApp). We found that sales have been increased during advertisement on YouTube as compared to other two media We have the following problems. How to select following variables Dependent Variable, Independent Variable(s), Intercept, Coefficients. How the model Select data from the past to learn what's the relationship. How minimize the predicted error for regression analysis. We need to check that a linear relationship exists between the dependent variable and the independent variables. We want to analyze the linear relationship between independent variables (YouTube, Facebook and WhatsApp) exists or not. Our second objective is to check that a linear relationship either exists: Sales and advertisement on TV both have positive are negative relationship, Sales and advertisement on Radio both have positive are negative relationship, Sales and advertisement on Newspaper both have positive are negative relationship. Predicted values and observed values should have minimum error. It is also called the residuals.

REFERENCE

1. Kosuke Imai“Using the Predicted Responses from List Experiments as Explanatory Variables in Regression” Advance Access publication November 11, 2014 Political Analysis (2015) 23:180–196doi:10.1093/pan/mpu017
2. Supichaya Sunthornjittanon “Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand” Portland State University PDX Scholar University Honors Theses University Honors College. Paper 131.10.15760/honors.137
3. Sandhya Jain et al “Regression Analysis – Its Formulation and Execution” In Dentistry Journal of Applied Dental and Medical Sciences NLM ID: 101671413 ISSN:2454-2288 Volume 2 Issue 1 January - March 2016.
4. Radek Silhavy “Analysis and selection of a regression model for the Use Case Points method using a stepwise approach” The Journal of Systems and Software 125 (2017) 1–14 Contents lists available at Science Direct The Journal of Systems and Software journal homepage: www.elsevier.com/locate/jss.

5. N. J. Gogtayetal “Principles of Regression Analysis” Journal of The Association of Physicians of India Vol. 65 April 2017Department of Clinical Pharmacology, Seth GS Medical College and KEM Hospital, Mumbai, MaharashtraReceived: 07.01.2017; Accepted: 15.01.2017.
6. Ira Sharma et al “Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal” Hindawi Scientific Volume 2018, Article ID 1383482, 8 pageshttps://doi.org/10.1155/2018/1383482 .
7. Anjali Pantet al “Regression Analysis Using R for Research and Development Assistant” Professor, Dept. of Applied Science, Dept. of Mathematics Government Polytechnic, G.B. Pant University of Shaktifarm, Uttarakhand, Agriculture and Technology, India Pantnagar, Uttarakhand, India October 2019.
8. Khushbu Kumari “Linear Regression Analysis Study” Downloaded free from http://www.j-pcs.org on Friday, July 17, 2020, IP: 157.34.76.130
9. Samit Ghosalet al “Linear Regression Analysis to predict the number of deaths in Indiadue to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th2020)” Contents lists available at Science Direct Diabetes & Metabolic Syndrome: Clinical Research & Reviews journal homepage: www.elsevier.com/locate/dsx.
10. B. Van Schaeybroecket al “Post-processing through linear regression” Processes Geophys., 18, 147–160, 2011 www.nonlin-processes-geophys.net/18/147/2011/ doi:10.5194/npg-18-147-2011
11. Astrid Schneider et al “Linear Regression Analysis Part 14 of a Series on Evaluation of” Scientific Publications Department of Medical Biometrics, Epidemiology, and Computer Sciences, Johannes Gutenberg University, Mainz, Germany: Dipl. Math. Schneider.
12. Sarimah Omar Ganet al “Multiple Linear Regression to Forecast Balance of Trade” Journal of Fundamental Sciences Vol.7, No.2 (2011) 150-155. | 150 | ISSN 1823-626X Journal of Fundamental Sciences available online at http://jfs.ibnusina.utm.my
13. TuróczyZ suzsannaaet al “Multiple regression analysis of performance indicators in the ceramic industry”2212-6716 © 2012 The Authors. Published by Elsevier Ltd. Selection and peer review under responsibility of Emerging Markets Queries in Finance and Business local organization. Doi: 10.1016/S2212-5671(12)00188-8
14. Gianie Abdu et al “Analysis of Consumer Behavior Affecting Consumer Willingness to Buy in 7-Eleven Convenience Store” Universal Journal of Management 1(2): 69-75, 2013 http://www.hrpub.org DOI: 10.13189/ujm.2013.010205
15. Lin Yu“A study of English reading ability based on multiple linear regression analysis”Hubei University of Technology, Hubei Wuhan, China. Journal of Chemical and Pharmaceutical Research, 2014, 6(6):1870-1877.
16. Marno Verbeek “Using linear regression to establish empirical relationships Using linear regression to establish empirical relationships”. IZA World of Labor 2017: 336 doi: 10.15185/izawol.336 | Marno Verbeek © | February 2017 | wol.iza.org.
17. Shen Rong Zhang et al “The research of regression model in machine learning field MATEC” Web of Conferences 176, 01033 (2018) https://doi.org/10.1051/matecconf/201817601033 IFID
18. Syarifah Diana Permaiaet al “Linear regression model using bayesian approach for energy performance of residential building” Available online at www.sciencedirect.com Science Direct Procedia Computer Science 135 (2018) 671–677 3rd International Conference on Computer Science and Computational Intelligence 2018