Inspira-Journal of Commerce, Economics & Computer Science (JCECS)
ISSN : 2395-7069, Impact Factor: 4.964, Volume 06, No. 03, July-September, 2020, pp. 05-08

LINEAR REGRESSION ANALYSIS (A GIFT OF STATISTICS TO RESEARCHERS)

Dr. Brahmdeo Modi*
Dr. Suresh PD Barnwal**

ABSTRACT

Statistical techniques used worldwide for making calculations and predictions in almost all of the areas of the profession or world. In addition to researchers making research worldwide, this is also used in medical science, engineering, research and development in Mathematics, Mining, Defence, Aviation industry, Navel services and many more. The list is not exhaustic. It could include a variety of areas. This perhaps found to be the most useful statistical tool used everywhere. In Regression analysis, a relationship between two variables is set out in such a way that values of Dependent variable is calculated from the independent variables. In linear regression analysis measurement of association between two different variables is measured. In statistics although number of different Statistical Techniques are available, but still most of the analyst uses linear regression analysis as the most safe and reliable statistical technique and uses it in variety of paradigm. However this is not only the feature of Linear Regression analysis. The another feature which is perhaps not known or researchers do not want to know the same, is Linear regression analysis too suffers from number of Deficiencies or drawbacks. Various errors have been noticed by the researchers during the use of Linear Regression analysis, which makes sense to doubt on the reliability of the techniques while using the same, especially when it is to be used for those researchers which have worldwide impact for example medical Science and CORONA VACCINE is the latest and best example in today's scenario. Hence this is always advised to use the technique with the utmost care and with the known exception. This article is enlighten the all aspect of Linear regression analysis.

***KEYWORDS**: Linear Programming, Regression Analysis, Statistical Techniques, Medical & Engineering.*

Introduction

The concept of Linear Regression Analysis toward the means of statistical Technique was first introduced by Sir Galton in 1894. Linear Regression Analysis toward the mean may be a statistical test applied to information set to define and measure the relation between the considered variables. Univariate statistical tests like Chi-square, Fisher's exact test, t-test, and analysis of variance (ANOVA) don't allow taking under consideration the effect of other covariates/con-founder during analyses. However, correlation and regression are the tests that allow the researcher to manage the effect of confounders within the understanding of the relation between two variables.

Generally in researches, the researcher often tries to know or relate or find out two or more independent (predictor) variables to predict an outcome or variable. This might be understood as how the danger factors or the predictor variables or independent variables account for the prediction of the

* Assistant Professor & Head, Department of Commerce, RNYM College, Barhi VBU, Hazaribag, Jharkhand, India.
** Assistant Professor & Head, Department of Mathematics, UCET VBU, Hazaribag, Jharkhand, India.

prospect of a disease occurrence, i.e., variable. Risk factors (or dependent variables) go together with biological (such as age and gender), physical (such as body mass index and Blood Pressure), or lifestyle (for example smoking and alcohol consumption) variables with the disease. Both correlation and regression provide this chance to grasp the "risk factors-disease" relationship as mentioned in their study by Gaddis and Gaddis in 1990. While correlation provides a quantitative way of measuring the degree or strength of a relation between two variables, multivariate analysis mathematically describes this relationship. On the other hand multivariate analysis allows predicting the worth of a variable supported the worth of a minimum of one experimental variable. In correlation analysis, the coefficient of correlation "r" may be a dimensionless number the value of which ranges from −1 to +1. A worth toward −1 indicates inverse or negative relationship, whereas towards +1 indicate a positive relation. The statistical regression analysis uses the mathematical equation, which is generally (y = mx + c), that describes the road of best acceptable the link between y (dependent variable) and x (independent variable). The parametric statistic, i.e., r2 implies the degree of variability of y thanks to x.

The Value of Data in Today's World

In today's world, data is generated everywhere so rapidly and there is a need to analyze the data and interpret it for future prediction. Many linear and non-linear models were developed for analysis of the data and estimation. Linear Regression Analysis is perhaps the most usable statistical model for making predictions the response or outcome variable. Statistical literature is rich with a large number of research articles on various regression models. One reason for the large number of publications is the great variety of regression models. Even a casual literature survey can provide information on the variety of different types of regression models. However the model gives various benefits such as Linear Regression Analysis indicates significant relationships between the predictor variables and the response variable and Linear Regression Analysis also indicates the strength of the impact that each predictor variable has on the response variable.

Regression analysis also allows comparisons between effects of predictor variables even when they are measured on different scales. It is therefore possible to drop or eliminate variables that are not really useful while identifying the best set of variables for building a predictive model.

The Need for the Study on Linear Regression Analysis

Bio- statistics gives meaningful inference of uncertain and haphazard data, but for that one should have the knowledge of that. Almost all research papers are using statistical methods and techniques. Although Statistics contributes a lot in medical research, there is still a need for the in-depth conceptual understanding of the research problem on hand and the extent of and nature of statistical analysis required to attain desired task of research so that possible errors arising out of lack of conceptual clarity regarding appropriate statistical techniques applicable for carrying out comprehensive statistical analysis of the research problem can be avoided. Various types of the errors which were noticed during the analysis of descriptive statistics Technique such as No mention of the level of measurement for each variable, Used "Mean± SD" for ordinal data, Didn't mention the "±" sign for descriptive statistics for normally distributed data, Calculated mean and SD for ordinal data, Did not mention the sample size calculation and sampling technique selected for this purpose, Only various measures used for calculation (i.e. only mean) and wrong usage of graphs.

Other General Errors which found over the use are Parametric test applying without check distribution assumption of data, Applied ANOVA without considering the assumption of required to apply it, Only the t-test is mentioned, but do not mention about the problem for which it is applicable in that case (i.e. independent, paired), Confusion regarding Parametric and Non parametric test, Use of application t –test for comparison of more than two independent groups(mean), Exact p-values not mentioned, Use of independent t-test for paired data, wirtht the intention of making Comparison of defined p-values, and Use of one or more independent t-test for two or more than two independent groups, or when plotting of the data shows a apparently nonlinear trend, Wrong used t-test for comparing survival times (some of which are censored), For ANOVA no mention of degrees of freedom, Use of chi-square instead of fisher test, Coefficient of agreement not computed, Power of the study not mentioned, Before applying ANOVA directly applied Post –hoc test, Usage of ANOVA for comparison of more than two related group mean, Used linear regression for ordinal data, Mention only p- value but didn't write the name of statistical method, Only mentioned the p- value without calculate confidence interval.

Various Regression Models

	Application	Dependent Variables	Independent Variables
Linear Regression Models	It describes relationship of linear nature in variable	Generally Continuous (for e.g. weight, blood pressure)	Continuous and/or categorical
Logistic regression	It helps in making Prediction of the probability of interrelated groups	Dichotomous	
Proportional hazard regression (Cox regression)	Modeling of survival data	Survival time (time from diagnosis to event)	
Poisson regression	It do Modeling of those processes which are of continuous nature	whole numbers which represents events in temporal sequence	

Various Kind of Models used for Linear Regression Analysis

When it comes to the type of model, three important considerations become the determining factor These three considerations are i) the number of predictor or independent variables in the model , ii) the shape of the regression line or the functional form of regression and iii) the nature of the response or dependent variable in the model. Most of the classical regression models are based on one or more of the above considerations in varying proportions. It is still possible to develop a new type of regression by using a new combination of the above considerations. However, it is necessary that the following seven commonly used regression models are well understood before an attempt is made to develop a new model.

The seven most commonly used regression models are found. The First one is Linear regression is one of the most popular techniques of predictive modeling. In its simplest form, apply known as simple linear regression. The regression line is optimal in the sense that it minimizes the total squared error of prediction. The second one is Logistic regression model is appropriate when the response variable is binary, that is, when there are only two possible responses. In such of the cases, the more required response is success rate so that the other level responses are called failure. Instead of predicting success or failure, the logistic regression model predicts the probability of success. The third one is A regression is said to be polynomial regression if predictor variables are raised to powers higher than 1 in the regression formula. What may be interesting is to note that polynomial regression is non-linear in predictor variables, but is still linear in regression parameters. The fourth one is Ridge regression When predictor variables are highly correlated , the data set is said to be suffering from multi co linearity the condition of multi co linearity does not influence the unbiased nature of the least square estimates of regression coefficients , but sampling variances of these estimates get inflated , resulting in a great loss of precision. The fifth one is Support Vector Regression (SVR) is a very recent development. SVR uses support vector machines as its basis and modifies the problem of classification of observations when the response variable is categorical or discrete. Nevertheless the main feature of maximum margin is maintained in SVR. The Sixth one is Lasso Regression is similar to ridge regression except for the fact that lasso regression results in selection of variables as a result of regularization. The last one is Principle component regression (PCR) is used in presence of multi collinearity or even when the number of predictor variables is too large. Also, since principle components reduce the dimensionality of a data set, PCR also achieves reduction in dimensionality of data.

Regression line for a multivariable regression

$Y= a + b1 \times X1 + b2 \times X2+ \ldots + bn \times Xn,$

where

Y = variable quantity

Xi = independent variables

a = constant (y-intersect)

bi= parametric statistic of the variable Xi

Example: regression curve for a multivariable regression $Y = –108.07 + 91.81 \times X1 + 0.49 \times X2 + 2.97 \times X3$, where

X1 = height (meters)

X2 = age (years)

X3 = sex (1 = female, 2 = male)

Y = the burden to be estimated (kg)g)

The purpose of describing all these regression models in order to establish a need of unifying the variety of models, so that every model can emerge as the most appropriate case of the unified model. The major advantage of having unified model is uniformity of processing rather than uniformity of model components as can be found in all the models that are in use. The second point of consideration is the fact that linear regression over-emphasizes normality of the distribution of residuals. There is no need for any consideration of this distribution as long as model fitting and predicting values of the response variable are concerned. It is only when some test of significance are to be carried out that the distribution becomes relevant. The objective of this study is aimed at to propose and develop a general linear regression model with the proposed model is linear in parameters but not necessarily in variables. At the same time it is also kept in mind that we have to identify the most appropriate functions of predictor variables for inclusion in the model. To test significance of every regression parameter in the model after estimating it and to identify the most appropriate or most significant general linear regression model for future use, has also been undertaken.

Conclusion

This is undoubtedly clear now that use of Linear Regression analysis is not a need rather it is necessity for today's world of research. The techniques has proven over a period of hundred years that the result obtained is so scientific and calculated that they can be used with the ease and accuracy. In spite of the fact that it do suffer from various errors and randomly occurrence of the error is the part of the use of this Technique, still the varied nature of usages of this technique and benefit obtained has made it compulsory now a days to use it. In statistics although number of different Statistical Techniques are available, but still most of the analyst uses linear regression analysis as the most safe and reliable statistical technique and uses it in variety of paradigm. This is in itself enough to identify the impact of the Linear Regression analysis. However with the development in the techniques with the use in various field including medical, the development has techniques has enriched and grown which made it more convenient to use it in all other fields too. Hence this technique shall be worked out with other scientific tools and Statistical Techniques so that more and optimum benefits can be obtained from it.

References

- ✓ Aram Karalic (1992), "Linear regression in regression tree leaves", John Wiley and Sons Inc.
- ✓ C. J. Leggetter (1995), "Maximum likelihood linear regression analysis for speaker adaptation of continuous density hidden model".
- ✓ Cande V. Ananth (1997) " Linear Regression analysis models for ordinal and special responses ", A review of various methods and applications tools
- ✓ Chan YH. Biostatistics 201: Linear regression analysis. Age (years). Singapore Med J 2004;45:55-61.
- ✓ Freedman DA. Statistical Models: Theory and Practice. Cambridge, USA: Cambridge University Press; 2009.
- ✓ Janhua Z Huang (1998) "Functional ANOVA models for generalized regression", Journal of Multivariate Analysis.
- ✓ Jerome Friedman (2010) "Regularization paths for generalized linear models via co- ordinate descent", Journal of Statistical software.
- ✓ Mendenhall W, Sincich T. Statistics for Engineering and the Sciences. 3rd ed. New York: Dellen Publishing Co.; 1992.
- ✓ Panchenko D. kumar 18.443 Statistical tools for use of Applications, Simple Linear Regression analysis. Massachusetts Institute of Technology: Open Course Ware; 2006.
- ✓ Ramona Maggini, A Lehmann, NE Zimmermann (2006) "Improving generalized regression analysis for the spatial prediction of forest communities", Journal of Biogeography.

❑❑❑