Research Journal of Mathematical and Statistical Sciences ________________________________ISSN 2320–6047
Vol. **1(6),** 14-20, July **(2013)** *Res. J. Mathematical & Statistical Sci.*

Review Paper

Applications of Modeling and Statistical Regression Techniques in Research

Ajay S. Singh and Micha B. Masuku
Department of AEM, University of Swaziland, Swaziland, Luyengo M205, SWAZILAND

Available online at: www.isca.in
Received 6th June 2013, revised 20th June 2013, accepted 2nd July 2013

Abstract

Applied statistics in research have an important role to play in the collection, compilation, analysis and interpretation of the data. In view of the day to day rapid changes in the research spectrum, the scenario is becoming interesting for a researcher. A Model is defined as abstraction of real situations, which aim to give the empirical content to relationships of variable and their interpretation. Modeling techniques are very common in basic as well as multidisciplinary research. This paper discusses the modeling and regression techniques for specific circumstances. Regression analysis technique explains the importance of variables and amount of change in exogenous variables if explanatory variables change with one unit. In this paper also describes the multiple and limited dependent variables especially logistic regression

Keywords: Statistical Model, Linear Regression, Logistic Regression, Odds Ratio.

Introduction

The model is a simplified design to describe the complex process using mathematical and statistical techniques. In the 18th century, simple mathematical models based on probability theory were used to understanding of the economics associated with insurance theory marked by Adam Smith[1]. This theory was associated with an estimation and chance theory, and played important role in development of probability theory and in development of vital statistics[2]. In 1730, De Movire described these problems in the 3rd edition of Doctrine of chances[3,4]. In 1709, Nicolas Bernouli studied problems related to saving and interest[5]. In 1730, Daniel Bernouli studied logarithmic utility of money. These developments were summarized in by Laplace in 1812[6-9]. Many mathematicians, during the last decade, have contributed in this field. Heckman and Synder[10] studied the probability models basede on linear assumption of the demand for attributes. Recently, Itzhak et al.[11] studied the probability and uncertainty in economic modeling. There are different types of statistical models in applied research which have been used for forecasting and modifying the future activity, efficient planning and allocation of funds, trading especially investment and speculation analysis. The statistical models are also used for long term risk management.

Understanding of Model

Modeling concepts are based on abstraction of real situations. Different types of models are applied in research according to the nature of variables and circumstances.

Qualitative model: These models involve some concepts of mathematical or quantitative approaches, some circumstances qualitative models are used because of their lack of precision.

Generally mathematical models are of classified as non-stochastic and stochastic types, depending upon the nature of the variables. Mathematical models classified as quantitative or qualitative according to the nature of proposed model and according to the characteristics of data.

Quantitative Model: In this model relationship of the associated quantitative variables expressed in terms of mathematical expressions.

Probabilistic Model: In the quantitative mathematical model outcomes associated with chance is defined as probability or probabilistic model.

Stochastic Model: The probabilistic model depending on time is known as stochastic model or stochastic process. The observations of variables are observed over time.

Non Stochastic Model: Non-stochastic mathematical models in some situations purely qualitative or quantitative. Non stochastic models involve same aspect of social qualitative concepts or rationalization of quantitative variables.

In the practical situation, quantitative model is applicable to many areas of economic, demographic and many methods have evolved independently. Many statistical models have been developed to test the significance of differences among means of different types of data.

Applications of Modeling Techniques

As mentioned in the introductory paragraphs, various models have been designed and employed in pure statistics and in diverse types of social, economic, health and scientific researches. A detailed review of them is beyond the scope of this article; selected researches from literatures are described in the succeeding paragraphs.

Bergstrand, prove the relationship of demand, supply as well as role of income through regression analysis[12]. Davidson et al. studied the relationship between income and expenditure in the United Kingdom through modeling based on econometric and time series[13]. A quantitative modeling technique has been employed by Bennett in analysis of livestock health and disease control[14]. Ruerd and Arjan used modeling to estimate the technical coefficient bio- economic farm household for Southern Mali[15]. Power et al derived a modeling technique with participation framework and integrated observation from interviews and discussion with farmers and consultants, with the help of dynamic bio-economic models to answer complicated questions on the allocation of resources at the business level[16]. Harville studied the estimation of component and related problems with biological, agricultural and behavioral sciences[17]. Emanuelson et. al. derived the model and estimate the important factors for clinical mastitis, somatic cell counts, and milk production[18]. Lui et al. presented the proportion of infected men and incubation period for AIDS in specific groups through model with estimation procedure[19]. Mishra, et al. and Singh estimate the fertility as well as bio-social parameters through statistical birth interval model[20, 21]. Allison uses modeling techniques to estimate the means and presented in very gentile fashion for missing data in the social science through maximum likelihood method[22].

Recently, some researchers studied the association of climate change on economic development of the nation through economic modeling techniques. Economic modeling has been employed to assess the economic growth with an independent factor[23, 24]. Insurance against catastrophic climate change has also been discussed by economic modeling[25].

Stochastic Model and Regression Analysis Technique

Stochastic model are widely used in the economic and econometric models, and generalized by Tinbergen. If the stochastic process satisfies the relationship between present and past observations, the econometric model is known as autogressive moving average model and later defined as autogressive heteroskedasticity and heteroskedasticity[26- 28].

Regression analysis is important statistical techniques to explain the relationships between the independent variables and amount of increase in one variable if another variable will be changed. The coefficient of parameters are estimated with the help of mathematical technique least of square. Regression methods are important techniques because of simplicity and fruitful interpretation.

Simple Linear Regression

Regression analysis is a mathematical measure of the average relationship between two are more variables in terms of the original unit of data. The variable whose value is influenced is called dependent variable and the variable which influences the values is defined as independent variable. In regression analysis independent variable is also known as regress or predictor while the dependent variable is known as regressed or explained variable. Simple linear regression analysis is a technique used for estimating the unknown value of a dependent variable from the known value of independent variable. In other words, X and Y are two related variables, then linear regression techniques helps to estimate the value of Y for a given value of X. Similarly, estimate the value of X for given value of Y. For example Market price, demand and supply related variables. In other hand linear regression analyses predict the market price based on demand and supply. It means that simple linear regression, one explanatory and one exogenous variable.

Lines of regression: There are two lines of regression first Y on X and second X on Y. The line of regression of Y on X is given by Y = α + β X where α and β are unknown constants known as intercept and slope of the equation. This is used to predict the unknown value of variable Y when value of variable X is independent.
$Y = \alpha + \beta X$

Similarly, the line of regression of X on Y is given by $X = \tau + \gamma Y$

Calculation of Linear Regression Coefficient if Y = α + β X: Regression coefficient β is the slope of the line of regression Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X.

First: calculate Slope (**α**) and Intercept (**β**)
Slope (**α**) = $[(\Sigma Y_i)(\Sigma Xi^2) - (\Sigma X_i)(\Sigma X_i Y_i)] / /\{ n \Sigma X_i^2 - (\Sigma X_i)^2 \}$
Intercept (**β**) = $[n(\Sigma X_i Y_i) - (\Sigma Y_i)(\Sigma Xi)] / /\{ n \Sigma X_i^2 - (\Sigma X_i)^2 \}$

Second: To calculate deviation and standard error for the regression line.
S = SQRT $[1/n\text{-}2 \{(\Sigma Y_i^2) - I \Sigma Y_i - \text{Slope} (\Sigma Y_i X_i)\}]$

Third: To calculate the error related with the slope and intercept
S_α = S x SQRT $[n / \{n \Sigma X_i^2 - (\Sigma X_i)^2\}]$
S_β = S x SQRT $[\Sigma (X_i^2) / \{ n (\Sigma X_i^2) - (\Sigma X_i Y_i)^2 \}]$
$Y = \alpha + \beta X$
b = Slope ± Δ slope ≈ Slope ± $t^* S_\alpha$
α = Intercept ± Δ Intercept ≈ Intercept ± Δ $t^* S_\beta$

Assumption: Linear regression does not test whether data is linear. It finds the slope and the intercept assuming that the relationship between the independent and dependent variable can be best explained by a straight line. It means that linear relationship between the true response and the independent variable. The observed values of independent variable are measured without error.

Multiple Regression Technique

Multiple regression analysis is a technique used for estimating the unknown value of a dependent variable from the known value of two or more than two variables (independent variables).

Multiple regression approach helps to predicting the value of Y for given values of $X_1, X_2, \ldots, X_k$. For example, the yield of maize per acre depends upon quality of seeds, fertility of soil, fertilizer used, temperature and rainfall.

Multiple Regression Model: In the multiple regression equation of Y on $X_1, X_2, \ldots, X_k$ is given as
$Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots\ldots + \beta_k X_k$ + Random Term (ε)

Where α_0 is the intercept or constant and $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are the slope in linear regression equation and also known as regression coefficients. Regression coefficient can be interpreted the same manner as slope. Thus if $\beta_i = 5.5$, indicates that Y will increase by 5.5 units if X_i increased by 1 unit. The significances of the multiple regression coefficients in the model can be tested by F-test through in the analysis of variance. Significant value of F indicates the linear relationship between Y and at least one of the independent variable's.

Multiple regression has been defined and for their goodness in terms estimating or predicting ability by coefficient of determination (R^2). The R^2 always lies between 0 and 1. Statistical significance of regression coefficient indicates that the variable influences Y significantly.

Assumptions: Multiple regression technique does not test whether data are linear related or not. It proceeds by assuming that the relationship between the Y and each of X_i's is linear. If non linearity present, one may use a suitable transformation to the linearity. Another important assumption is without multicollinearity. It means that independent variables are not correlated.

Multiple regression technique is used when researcher is interested in estimating a continuous dependent variable from a number of independent variables in case of dependent variable. If variable is dichotomous in this case logistic regression should be used.

Logistic regression

Logistic regression is used for estimating the results of a predictor or independent variable. Variables can take a limited number of classified variables. Logistic regression is classified as binomial or multinomial logistic regression. Binomial logistic regression is the instance in which the observed results have only two possibilities. Multinomial logistic regression refers the outcome have three or more possibilities. Generally, the outcome is coded as "0" and "1" in binary logistic regression.[29] Binomial distribution has a mean equal to the proportion of cases or events, denoted by P, and a variance equal to the product of cases and non- cases (Q) or (1 –P) [30.] In other words logistic regression is used to estimate the odds based on the predictor(s). The odds are defined as the probability of a positive case divided by the probability of a negative case.

Concepts of Function: The logistic function relates the independent variable, X. $P = e^{\alpha + \beta X} / (1 + e^{\alpha + \beta X})$ Or $P = 1 / 1 + e^{-(\alpha + \beta X)}$ P is the probability of a "1", "*e*" is the base of the natural logarithm and **α** and **β** are the parameters of the model.

In logistic regression, the dependent variable is a logit, which is the natural log of the odds, that is, log (odds) = logit (P) = In (P/ 1 – P)

So that logit is a log of odds and odds are functions of P, the probability of a 1. In logistic regression, we find logit (P) = $\alpha + \beta X$,

Which is assumed to be linear, that is, the log odds are assumed to be linearly related to X. First have to take the log out of both sides of the equation. Then we have to convert odds to a simple probability:

$\text{In} (P/ 1 - P) = \alpha + \beta X$

$P / (1 - P) = e^{\alpha + \beta X}$

$P = e^{\alpha + \beta X} / 1 + e^{\alpha + \beta X}$

If log odds are linearly related to X, then the relation between X and P is nonlinear.

Logistic Regression Model: Logistic regression begins with the logistic function which always values between zero and one [29]

$$\pi (x) = e^{(\beta 0 + \beta 1X1 + e)} / e^{(\beta 0 + \beta 1X1 + e)} + 1 = 1 / e^{-(\beta 0 + \beta 1X1 + e)} + 1 \quad (1)$$

and

$$g (x) = \text{In} [\pi (x) / 1 - \pi (x)] = \beta_0 + \beta_1 X_1 + e \quad (2)$$

and

$$\pi (x) / 1 - \pi (x) = e^{(\beta 0 + \beta 1X1 + e)} \quad (3)$$

In this equation input is $\beta_0 + \beta_1 X_1 + e$ and the output is π (x).
g(*X*) is the logit function of the predictor *X*, while π(x) is the probability of a case, β_0 is the intercept from the linear regression equation, $\beta_1 X_1$ is the regression coefficient multiplied by some value of the predictor, base *e* denotes the exponential function and e in the linear regression equation denotes the stochastic term.

The first equation is the probability of being a case is equal to the odds of the exponential function of the linear regression equation. The second equation shows that the logit is equal to the linear regression equation. Equation third explains the odds of a case are equivalent to the exponential function of the linear regression equation. The logit varies from (-∞, + ∞) and it is an adequate criterion to conduct linear regression and the logit is easily converted into the odds. [29,]

A log-linear model: In this case, the logit of the probabilities p_i as a linear predictor. Linear predictor classified into two, one for each of the possibilities:

$\text{In } p (Y_i = 0) = \beta_0 X_i - \text{InZ}$

$\text{In } p (Y_i = 1) = \beta_1 X_i - \text{InZ}$

The linear predictor and InZ is the normalizing factor.

$\mathbf{P (Y_i = 0) = (1/Z) e^{\beta 0.Xi}}$

$\mathbf{P (Y_i = 1) = (1/Z) e^{\beta 1.Xi}}$

This means that Z is the sum of all non-normalized probabilities, and by dividing each probability by Z. Therefore,

$Z = \mathbf{e^{\beta 0.Xi} + e^{\beta 1.Xi}}$

After that resulting equations are
$P(Y_i = 0) = e^{\beta 0.Xi} / e^{\beta 0.Xi} + e^{\beta 1.Xi}$
$P(Y_i = 1) = e^{\beta 1.Xi} / e^{\beta 0.Xi} + e^{\beta 1.Xi}$
or generally, $P(Y_i = c) = e^{\beta c.Xi} / \sum e^{\beta h.Xi}$ where h = 1, 2, …

Generalization of linear model: $\text{logit}(E[Y_i \mid x_{1,i}, \ldots, x_{m,I}] = \text{logit}(p_i) = \text{In}(p_i / 1-p_i) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_m x_{m,i}$

By using the more compact notation described above, $\text{logit}(E[Y_i \mid X_i]) = \text{logit}(p_i) = \text{In}(p_i / 1- p_i) = \beta . X_i$

This logistic regression as a type of generalized linear model, which predicts variables with various types of probability distributions by fitting a linear predictor function and both the probabilities p_i and the regression coefficients are determined by maximum likelihood estimation.

Meaning of R^2s in regression techniques: In linear regression the multiple correlations, R^2 is used to assess goodness of fit and it represents the proportion of variance in the criterion that is explained by the predictors.[30] In logistic regression analysis, there is no agreed upon analogous measure, but there are several competing measures each with limitations. [30] There are three of the most commonly used indices are examined with the likelihood ratio R^2_L [30].
$R^2_L = [D_{null} - D_{model}] / D_{null}$
R^2_L also represents the proportional reduction in the deviance and the deviance is considered as a measure of variation but it is not identical to the variance in linear regression.[31]

Some applications of regression techniques in applied research

Singh analyzed the fertility behavior through regression model and Mishra also discussed stochastic regression model and analyzed the impact of couple protection rate on birth[32, 33]. Bade et al.[34] studied the development of metropolitan area with the help of this technique. Connolly analyzed the employment discrimination through multiple regression and Amoako-Abu et al. used regression analysis technique for evaluating the effect on stock price reaction to the introduction and reduction of capital gains tax exemption[35,36]. Pilotte[37] estimated the stock price response to new financing with regression analysis approach. Bartov[38] studied the nature of information conveyed by open market stock repurchase announcements by regression techniques. Dayhoff[39] presented the estimates of the determinants and regression analysis proved that higher displacement rates had a positive significant effect on enrollments for all groups. Edwards[40] analyzed the seafood consumption and prices based on time series data and suggested that seafood had strengthened in response to medical preferences through regression techniques. Kang and Reichert[41] used multiple regression techniques to generate the adjustment coefficients used in the grid adjustment method. Maki and Ignace[42] used regression analysis to examine the effect of trade unions and the earnings gap between male and female employees in Canada. Singh and Singh analyzed the effect of some important variables on birth rate and infant mortality rate[43, 44].

Conclusion

Research investigation and their analysis is the part of a wider development of any nation with regard to finance, education, public health, and agriculture, etc. that are indicators of better life of human beings. Any social phenomenon and especially those that can be characterized by numerical facts are the results of one or more causes of action. The concept of models is simplified to describe the complex economic, social, biological etc. processes. The regression analysis is a powerful technique of applied research and it becomes efficient through the skill of the researcher to replace a complex system of causes with a simple system where causative situations change with time. In this paper, the described the role of model in applied research as well as applications of regression analysis techniques such as relationship, amount of increase in dependent variable, have been emphasized. Linear, Multiple, Logistic Regressions analyze the factors in efficient manner and researchers use and interpret the results scientifically. A researcher can calculate the coefficients according to the design of model and regression techniques mentioned and interpret the result. A careful consideration of techniques will hopefully result in more meaningful studies whose results and interpretations will eventually receive a high priority for publication.

References

1. Smith A., An inquiry in to the nature and causes of the wealth of nation, Edwin Cannan, ed. 1904, Library of Economics and Liberty (www.ecolib.org/library) (**1904)**

2. Frees Edward W. and Chow L. Martin, Stochastic life contingencies with solvency considerations, Transaction of the Society of Actuaries, XLII, 91-148 (**1990**)

3. Hald Andres and De Movire, The Doctrine of changes, 1718, 1738, 1756, History of Probability and Statistics and their Applictions before 1750, Wiley Series in Probability and Statistics, Wiley Inter-science, 397 (**1990**)
4. Schneider Lvor and A. De Moivre, The Doctrine of chances (1718, 1738, 1756) in Grattan, Guinness, L., Landmark Writings in Western Mathematics, 1640-1940, Amsterdam: Elsevier, 105-120 (**2005**)
5. Fleckenstein J.O. and Bernouli N.I., Dictionary of Scientific Biography 2, New York: Charles Scribner's Sons, 56-57 (**1980**)
6. Lange O., The scope and Method of Econometric, *Review of Economic Studies*, **13(1),** 19-32 (**1994**)
7. Holcombe R., Economic Models and Methodology, Greenwood Press, New York (**1989**)
8. Samuelson Paul A., The simple mathematics of income determination. in Metzeler L. loyed., Income, Employment and Public Policy; easy in honor of Alvin Hansen, New York (**1948**)
9. Stephen M. Stigler, The history of statistics, Harvard University Press, Chapter 3 (**1986**)
10. James J. Heckman and James M. Synder, Linear probability models of the demand for attributes with an empirical applications to estimating the preferences of legislators. National Bureau of Economic Research, WP. No. 5785 (**1996**)
11. Itzhak Gilboa, Andrew W. Postlewaite and David Schmeidler, Probability and uncertainty in economic modeling, *Journal of Economic Perspective,* **22(3),** 173-188 (**2008**)
12. Bergstrand Jeffrey H., The Heckscher-Ohlin-Samuelson Model, the Linder Hypothesis and the Determinants of Bilateral Intra-industry Trade, *Economic Journal,* **100(403),** 1216-29 (**1990**)
13. Davidson J.E.H., Hendry D.F., Siba F. and Yeo J.S., *Economic Journal,* **88,** 661-692 (**1978**)
14. Bennett R.M., Preventive Veterinary Medicine**, 13(1),** 63-76 (**1992**)
15. Ruerd R. and Arjan, R., Technical coefficients for bio-economic farm household: a Meta modeling approach with application for Southern Mali. *Ecological Economics,* **36(3),** 427-441 (**2001**)
16. Power B., Rodriguez D. deVoll P., Harris G. and Payero J., *Field Crops Research,* **124,** 171-179 (**2011**)
17. Harville D.A., Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association,* **72(358),** 320-338 (**1977**)
18. Emanuelson U., Danell B. and Philipsson J., Estimate of the genetic parameters for clinical mastitis, somatic cell counts, and milk production estimated by multiple-trait restricted maximum likelihood, *Journal of Dairy Science, Elsevier*, **71(2),** 467-476 (**1988**)
19. Lui K.J., Darrow W.W and Rutherford G.W., A model-based estimate of the mean incubation period for AIDS in homosexual men. 3rd– Science, New York, USA **(1988)**
20. Mishra R.N., Srivastava P.K. Singh A.S. and Bhardwaj S.D., A probability model of first conceptive delay and its application, Proc. of 10th annual convention of ISMS (**1992**)
21. Singh A.S., Some analytical models for human fertility behavior and its applications Ph. D. Thesis, IMS, BHU, Varanasi, India (**1992**)
22. Allison P.D., Missing data: Quantitative applications in the social sciences, British Journal of Mathematical and Statistical Psychology, 2002 - Wiley Online Library (**2002**)
23. Stem N., The economics of climate change: The Stem Review, Cambridge University Press, Cambridge, (**2007)**
24. Eboli F., Parrado R. and Roson R., Climate change feedback on economic growth: exploration with a dynamic general equilibrium model. Paper presented at XIth annual conference on global economic analysis, Helsinki, June. (**2008**)
25. Adams P.D., Insurance against catastrophic climate change: how much will an emissions trading scheme cost Australia?, The *Australian Economic Review,* **40(4),** 432-452 (**2007**)
26. Hyashi F., Econometrics, Princeton University Press (**2000**)
27. Russel D. and MacKinnnon J.G., Econometric Theory and Methods. New York: Oxford University Press, (**2004**)
28. Wooldridge J., Introductory Econometrics: A modern approach, Thomson South-Western (**2003**)
29. Hosmer David W., Lemeshow, Stanley. Applied Logistic Regression (2^{nd} ed.). Wiley **(2000)**

30. Cohen Jacob, Cohen Patricia, West Steven G. and Aiken Leona S., Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.) Routledge (**2002**)

31. Menard Scott W., Applied Logistic Regression (2nd ed.). SAGE (**2002**)

32. Singh G., Bhardwaj S.D. and Singh A.S., Fertility analysis by linear regression model, Health and Population-Perspective and Issues, **13,** 68, (**1990**)

33. Mishra R.N, Singh A.S., Mohapatra S. C. and Bhardwaj, S. D., Impact of couple protection rate on birth rate: stochastic linear regression model, *Indian J. Community Medicine,* **XV**II, 55, (**1992**)

34. Baade Robert and Dye Rechard F., Studied with help of the regression technique impact of stadiums and professional sports on metropolitan area development, *Growth and Change,* **21(2),** 1-14 (**1990**)

35. Connolly C., The use of multiple regression analysis in employment discrimination cases", *Population Research and Policy Review,* **10(2),** 117-135 (**1991**)

36. Amoako A. Ben, Rashid M. and Stebbins M., Evaluate the differential effect on stock price reaction to the introduction and reduction of capital gains tax exemption, *Journal of Banking and Finance,* **16(2),** 275-287 (**1992**)

37. Pilotte E., Growth opportunities and the stock price response to new financing*, Journal of Bussiness,* **65(3),** 371-374 (**1992**)

38. Bartov E., Open Market Stock Repurchases as Signals for Earnings and Risk Changes, *Journal of Accounting and Economics,* **14(3),** 275-294 (**1991**)

39. Dayhoff Debra A., High School and College Freshmen Enrollments: The Role of Job Displacement, *Quarterly Review of Economics and Business,* **31(1),** 91-103 (**1991**)

40. Edwards, Steven F., Evidence of Structural Change in Preferences for Seafood, *Marine Resource Economics,* **7(3),** 141-151 (**1992**)

41. Kang Han Bin and Reichert Alan K., An Empirical Analysis of Hedonic Regression and Grid Adjustment Techniques in Real Estate Appraisal, *American Real Estate and Urban Economics Association Journal,* **19(1),** 70-91 **(1991)**

42. Maki Dennis, Ng Ignace, Effects of Trade Unions on the Earnings Differential between Males and Females: Canadian Evidence, *Canadian Journal of Economics,* **23(2),** 305-311 **(1990)**

43. Singh A.S., Impact of bio- demographic factors on fertility: Stochastic regression model, *Research Dimension,* **2(1),** 1-6 **(2011)**

44. Singh A.S., Factors affecting Infant mortality in India: Statistical interpretation by regression analysis, *JP Journal of Biostatistics,* **6(2),** 109-120 **(2012)**