CONCEPTS IN REGRESSION ANALYSIS

C. Dwarakanatha Reddy[1], K. Naga Vihari[2], K. Murali[3], R. Abbaiah[4] and N. Ramachandra[5*]

Department Of Statistics, Sri Venkateswara University, Tirupati-517502.

*Corresponding Author : **nakkalaramachandra@gmail.com**

Abstract

Regression analysis is a set of statistical techniques that may be used to make judgements about the relationships between variables. Regression analysis is now the most widely used data analysis tool, with applications in practically every field of study, including social, physical and Biological Sciences, business and engineering. As a result, the purpose of this text is to develop the underlying theory of this key statistical method and to illustrate it with examples drawn from economics, demography, engineering and biology. We incorporated fundamental statistics, linear algebra and numerical analysis topics to make the text somewhat self-contained. In addition, we have sought to convey details of the theory rather than only giving computational and interpretive features, in contrast to other publications on the subject.

KeyWords: Linear Regression Model, BLUE for Linear Regression Model, OLS Estimators, Kronecker Product and Restricted Least Squares.

INTRODUCTION :

Sir Francis Galton (1822-1911), a well-known British anthropologist appears to be the first to use the term “Regression” in his research of heredity. He discovered that on average children’s heights do not tend towards the heights of their parent’s, but rather towards the average when compared to their parent’s. Galton termed this “regression to mediocrity in hereditary stature”. The experiments demonstrated further that the filial regression towards mediocrity was exactly proportionate to the parental divergence from it, according to the journal of the anthropological institute, Vol. 15 [1885], PP. 246-263. Galton then explained how to use parent's heights to determine the link between children's heights. Today, Galton's approach would be referred to as a "correlation analysis", a term he coined. There are no aspects of "regression" in the original meaning in most model-fitting scenarios now a days. Nonetheless, the term has become so ingrained in our lexicon that we continue to use it. We recommended the statistical encyclopaedia (or) the history of Regression for more related stories about Regression.

$Y=X\beta + \in$ is a well-known basic linear regression model. It is actually a system of n-linear equations, each of which is one linear equation for a single observation. One may argue that we are dealing with a system of equation systems when we have numerous behavioural (or) technological equations that need to be approximated at the same time. In general, the kronecker product is a useful tool for dealing with such complex systems. These systems have become increasingly important in data analysis not only in economics but also in other sectors such as industry, geography, social science and biology. Various fields of economic theory use sets of linear regression models. We assume a series of linear regression equations, each of which is an equation of the type $Y=X\beta + \in$ and where the disturbances in distinct equations may be correlated. When estimating a number of related economic functions, such as demand equations for a variety of commodities, such specifications is likely to be realistic, consumption functions for subsets of the population or investment functions for a number of enterprises. In these instances, the disturbances for distinct functions at a given point in time are likely to represents some shared unmeasurable or ignored causes, therefore some connection is expected. Contemporaneous correlation is the relationship between different disturbances at a specific point in time.

BASIC CONCEPTS IN REGRESSION ANALYSIS

STATISTICAL LINEAR REGRESSION MODEL

Consider a linear relationship between a dependent variable Y and a set of k explanatory variables $X_1 , X_2 , \ldots\ldots\ldots X_k$ and a disturbance variable $\in$ as

$$Y_j = \beta_1 X_{i\,1} + \beta_2 X_{i\,2} + \ldots\ldots\ldots\ldots + \beta_k X_{i\,k} + \in_j$$

Where, Y_j = jth observation on the dependent variable Y

X_{ij} = jth observation on the ith explanatory variable (regressor).

β_i = regression coefficient corresponding to the ith regressor

$\in_j$ = jth observation on the disturbance (error) term

n = Number of observations

k = Number of Regressors

The above model can be written in matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} \; . \; . \; . \; . \; . \; . \; . \; . \; . & X_{1k} \\ X_{21} & X_{22} \; . \; . \; . \; . \; . \; . \; . \; . \; . & X_{2k} \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ X_{n1} & X_{n2} \; . \; . \; . \; . \; . \; . \; . \; . \; . & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \in_1 \\ \in_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \in_n \end{bmatrix}$$

(or)

$$Y = X\,\beta + \in$$

Here, Y is of dimension n×1

X is of dimension n×k

β is of dimension k×1

∈ is of dimension n×1

The crucial assumptions about the classical linear regression model (1.4.2) are

1. ∈ has a zero mean vector $0_{n\times1}$

   i.e., E [∈] = $0_{n\times1}$

2. ∈ has the variance – covariance matrix $\sigma^2 I_n$

   i.e., $E[\in\in^{|}] = \sigma^2 I_n$

   Thus, the random disturbances are uncorrelated with each other and have same unknown variance σ^2.

3. X is a non-stochastic matrix of rank k< n, and has the property that

   $\lim_{n\to\infty}\left[\frac{X^{|}X}{n}\right]$ = Q exists as a finite non-singular matrix.

4. The explanatory variables are independent of measure mental error.
5. ∈ follows a multivariate normal distribution. The assumptions (1), (2) and (5) may be combined stated as

$$\in \sim N[0, \sigma^2 I_n]$$

(or)

$$\in_i \sim \text{i.i.d. } N[o, \sigma^2], \qquad i = 1,2,\dots,n$$

The linear model Y=Xβ + ∈ with the above five assumptions is known as ‘Standard General Linear Model’ or ‘Gauss-Markoff Linear Model’. Anscombe and Tukey (1963) termed the above assumptions as the full ideal conditions for the linear model.

BLUE FOR LINEAR PARAMETRIC FUNCTION IN LINEAR REGRESSION MODEL

In the Statistical linear regression model $Y=X\ \beta\ +\in$, the Best Linear Unbiased Estimator (BLUE) of a linear parametric function $P^{|}\beta$ is given by $P^{|}\widehat{\beta}$ ·

Where $\hat{\beta}$ is the Ordinary Least Squares (OLS) estimator of β which can be obtained by solving the system of normal equations $X_1^{|}X\ \hat{\beta}\ X^{|}Y$.

Here P is a (K×1) vector of known constants.

For $P^{|} = [0\ 0\ ......1\ 0........0]$, where unity is in i^th position and zero elsewhere, the OLS estimators $\hat{\beta}_i$, $i=1, 2, ... , K$ are the best linear unbiased estimators of the parameters β_i , $i=1,2,....., K$ Thus, the OLS estimator $\hat{\beta}$ is the BLUE of β.

Further, an unbiased estimator of error variance σ^2 is given by

$$S^2 = \frac{e^{|}e}{n-K}$$

Where, $e = (Y-X\hat{\beta})$ is the $(n\times1)$ vector of OLS residuals.

This criterion was first given by Markoff based on Gaussian Law of Errors.

PROPERTIES OF OLS ESTIMATORS

1. The OLS estimators $\hat{\beta}$ and S^2 are unbiased estimators of β and σ^2 respectively.
2. The covariance matrix of $\hat{\beta}$ is given by $v(\hat{\beta}) = \sigma^2\ (X^{|}\ X)^{-|}$
3. $\widehat{\beta}$ and S^2 are consistent estimators of β and σ^2 respectively.
4. If one of the Regressors is a constant term then coefficient of multiple determination is given by

$$R^2 = \frac{\hat{\beta}X'Y}{Y'Y}$$

5. $\hat{\beta}$ has a multivariate normal distribution with mean vector β and covariance matrix $\sigma^2\ (X'\ X)^{-1}$

6. $\frac{(n-K)S^2}{\sigma^2}$ has a χ^2- distribution with (n-K) degrees of freedom.

7. The variance of S^2 is $\frac{2\sigma^4}{n-K}$

8. $\hat{\beta}$ and $\frac{(n-K)S^2}{\sigma^2}$ are independently distributed as $N[\ \beta,\ \sigma^2\ (X'\ X)^{-1}]$ and χ^2 with (n-K) degrees of freedom respectively.

9. $\hat{\beta}$ and S^2 are efficient estimators of β and σ^2 respectively.

10. $\hat{\beta}$ and $(n\text{-}K)S^2$ are the joint sufficient statistics of β and σ^2.

11. The Cramer - Rao lower bounds for the variance of $\hat{\beta}$ and σ^2 are

    $\sigma^2\ (X'\ X)^{-1}$ and $\frac{2\sigma^4}{n}$ respectively.

12. The maximum likelihood estimators of β and $\sigma^2$are respectively given by $\hat{\beta}$ and $\frac{(n-K)S^2}{n}$

13. The asymptotic distribution of $\sqrt{n}\ (\hat{\beta}\text{-}\beta)$ is given by

    $$N\ [0,\ \sigma^2\{\lim_{n\to\infty}(\frac{X'X}{n})^{-1}\}]$$

14. The asymptotic distribution of $\sqrt{n}\ (S^2 - \sigma^2)$ is given by $N[0,\ 2\sigma^4]$

15. $\hat{\beta}$ and S^2 are asymptotically efficient estimators and their asymptotic

    Variances are respectively given by $\sigma^2\ (X'\ X)^{-1}$ and $\frac{2\sigma^4}{n}$.

THE VIOLATION OF THE CRUCIAL ASSUMPTIONS CAUSES PROBLEMS

The numerous issues that may occur as a result of violations of the critical assumptions one at a time, in the sense that it will consider a violation of one of the assumptions at a time. While assuming that everything else remains the same.

1) Suppose that ∈ has a non zero mean vector, say μ i.e., E (∈) = μ.

Then, (i) $E(\hat{\beta}) = \beta + (X^{|}X)^{-|}\, X^{|}\, \mu$;

(ii) Plim $(\hat{\beta}) = \beta + Q^{-|} \lim_{n\to\infty}[\frac{X^{|}\mu}{n}]$;

(iii) $E(s^2) = \sigma^2 + \left(\frac{\mu^{|}M\mu}{n-k}\right)$, Where $M = [1-X\,(X^{|}\,X)^{-|}\,X]$

(iv) Plim $(S^2) = \sigma^2 + \lim_{n\to\infty}\left(\frac{\mu^{|}M\mu}{n}\right)$

Thus, $\hat{\beta}$ is biased unless $X^{|}\mu = 0$ and is inconsistent unless $\lim_{n\to\infty}(\frac{X^{|}\mu}{n}) = 0$. Further, S^2 is biased unless $\mu^{|}M\mu = 0$ and is inconsistent unless $\lim_{n\to\infty}(\frac{\mu^{|}M\mu}{n})=0$

Hence, the usual least squares estimators $\hat{\beta}$ and S^2 may not have the desirable properties of unbiasedness and consistency. Also, they do not lead to valid tests of hypothesis.

2) Suppose that the rank of the regressor matrix X is less than K, the number of regressors.

i.e., ρ (X) < K

In this case, the columns of X are linearly dependent and there may arise the problem of multicollinearity.

Then the OLS estimator $\hat{\beta} = \left(X^{|}X\right)^{-|} X^{|}Y$ does not exist. Since, normal equations are always consistent, it is possible to solve the set of least squares normal equations $X^{|}X\,\hat{\beta} = X^{|}Y$ by using the concept of generalised inverse of a matrix.

3) Suppose that the regressor matrix X is random (atleast in part) rather than non-stochastic. In this case, the OLS estimates fail to give the best linear unbiased estimates for the parameters of the linear model and there may arise the problem of Stochastic Regressors. The OLS estimator $\hat{\beta}$ will be consistent as long as Plim $\left[\frac{X^{\prime}\in}{n}\right]$=0.

4) Suppose that the observations on explanatory variables involve measure mental error. Then it causes the problem of errors in variables in this context, the OLS estimator $\hat{\beta}$ is inconsistent. Instrumental variables estimation. The maximum likelihood estimation and grouping of observations method of estimation may give consistent estimators for the parameters of the linear model.

5) Suppose that $\in$ follows a multivariate probability distribution other than multivariate normal distribution. Then there may arise the problem of non-normal disturbances. In this case, the OLS estimator $\hat{\beta}$ is unbiased, consistent and BLUE for β. Also, S^2 is unbiased and consistent estimator for σ^2. Further, $\hat{\beta}$ does not have a normal distribution, $\frac{(n-K)S^2}{\sigma^2}$ does not have a χ^2–distribution, and either $\hat{\beta}$ or S^2 are not efficient or asymptotically efficient. In this context, some Robust estimation methods may be applied to get efficient estimators for β.

6) Suppose that the regression co-efficients of the linear model are governed by some probability law. Then it may cause the problem of Random Co-efficients Regression (RCR) models.

TESTING THE LINEAR RESTRICTIONS ABOUT PARAMETERS OF LINEAR REGRESSION MODEL

Linear hypothesis is a statement about the parameters of a linear model which is in the form of linear function of parameters. General linear hypothesis consists of a set of linear hypothesis about the parameters of a linear model.

Consider a classical linear regression model as

$$Y_{n\times1}= X_{n\times K}\,\beta_{K\times1} + \epsilon_{n\times1}$$

A general linear hypothesis consists of a set of q ($\leq K$) linear restrictions about the elements of β can be expressed in the matrix notation as

$$H_0: R_{q\times K}\,\beta_{K\times1}= r_{q\times1}$$

Where, R :(q×K) matrix of known constants with full row rank

r: (q×1) vector of known constants

To test H_0, first of all the unknown β in may be replaced by the OLS estimator $\hat{\beta}$ and obtaining $R\hat{\beta}$. The sampling distribution of $R\hat{\beta}$ is derived as follows:

$$E[R\,\hat{\beta}\,] = RE[\,\hat{\beta}\,] = R\beta$$

$$V[R\hat{\beta}\,] = E[R\,(\hat{\beta} - \beta)][R\,(\hat{\beta} - \beta)]'$$

$$=RE[\hat{\beta} - \beta][\hat{\beta} - \beta]'\,R'$$

$$= \sigma^2\,[R\,(X'X)^{-1}\,R'] \qquad [\because V\,(\hat{\beta}) = \sigma^2(X'X)^{-1}]$$

Since, $\hat{\beta}$ follows multivariate normal distribution,

$$R\hat{\beta} \sim N\,[R\beta, \sigma^2 R\,(X'X)^{-1}\,R']$$

(or)

$$R\,[\hat{\beta} - \beta] \sim N[0, \sigma^2 R(X'X)^{-1}\,R']$$

If the hypothesis is true, one can replace Rβ in by r obtaining

$$[R\hat{\beta} - r] \sim N[0, \sigma^2 R\,(X'X)^{-1}\,R']$$

$$\Rightarrow (R\hat{\beta} - r)' \; [\sigma^2 R \left(X'X\right)^{-1} R']^{-1} \, (R\hat{\beta} - r) \sim \chi^2_q$$

Here, $[R\left(X'X\right)^{-1} R']$ is positive definite matrix.

Since, $\frac{(Y-X\hat{\beta})'\,(Y-X\hat{\beta})}{\sigma^2} = \frac{e'e}{\sigma^2} \sim \chi^2_{(n-k)}$ independently with $\hat{\beta}$ and hence independently with $R\hat{\beta}$.

Thus to test the general linear hypothesis H_0: $R\beta = r$, the F- statistic is given by

$$F = \frac{(R\,\hat{\beta} - r)'[R\,(X'\,X)^{-1}R']^{-1}\,(R\hat{\beta} - r)/q}{e'e/n-K} \sim F_{q,\,n\text{-}K}$$

CONCEPT OF KRONECKER PRODUCT

If $A = ((a_{ij}))$ and $B = ((b_{ij}))$ be two matrices of order $(m\times n)$ and $(p\times q)$ respectively, then the kronecker product of A and B is a matrix of order $(mp\times nq)$ and is denoted by $A\otimes B$. It is defined as

$$A\otimes B = \begin{bmatrix} a_{11}B & a_{12}B \,.\,.\,.\,.\,.\,.\,.\, a_{1n}B \\ a_{21}B & a_{22}B \,.\,.\,.\,.\,.\,.\,.\, a_{2n}B \\ . & . \qquad\qquad\quad . \\ . & . \qquad\qquad\quad . \\ . & . \qquad\qquad\quad . \\ . & . \qquad\qquad\quad . \\ . & . \qquad\qquad\quad . \\ a_{m1}B & a_{m2}B \,.\,.\,.\,.\,.\,.\,.\, a_{mn}B \end{bmatrix}$$

From the definition of the kronecker product, it is clear that there are no restrictions on the numbers m, n, p and q for the product to be meaningful.

PROPERTIES OF KRONECKER PRODUCT OPERATOR

1. If A, B and C are any three matrices, then $(A\otimes B)\otimes C = A\otimes(B\otimes C)$.

It is Customary to denote $(A\otimes B)\otimes C$ by $A\otimes B\otimes C$.

2. If A, B and C are three matrices with B and C being of the same order then

$A \otimes (B+C) = (A \otimes B) + (A \otimes C)$

3. If K is a scalar and A is any matrix then

$K \otimes A = KA = A \otimes K$

4. If A, B, C and D are four matrices such that each pair A and C, B and D is

Conformable for the usual multiplication then

$(A \otimes B)(C \otimes D) = AC \otimes BD$

5. If A and B are any two matrices then

$(A \otimes B)^{|} = A^{|} \otimes B^{|}$

6. The equality A⊗B = B⊗A rarely holds just as it is the case for the multiplication of matrices. However, the kronecker product has one distinctive feature. The matrix B⊗A can be obtained from A⊗B by interchanging rows and columns of A⊗B **.**

7. If A and B are non-singular matrices not necessarily of the same order, then

$(A \otimes B)^{-|} = A^{-|} \otimes B^{-|}$

8. If A and B are square matrices not necessarily of the same order, then

r (A⊗B) = [r(A)] [r(B)]

9. If A and B be two square matrices with eigen values

λ_1 , λ_2 , ...λ_m and μ_1 , μ_2 ,.............. μ_n respectively, then the

$\lambda_i \mu_j$, i =1,2,.....m and j = 1,2, ...n are the eigen value of A⊗B.

10. If A and B are two matrices not necessarily square, then

Rank [A⊗B] = [Rank (A)] [Rank (B)]

11. If the matrices A and B are of the same order and so are the matrices C and

D, then (A+B)⊗(C+D) = (A⊗C) + (A⊗D) + (B⊗C) + (B⊗D).

ESTIMATION OF RESTRICTED LEAST SQUARES FOR β

Consider a general linear hypothesis as in (1.3.3), $H_0 : R\beta = r$

If H_0 is not rejected, then one may re-estimate the model, incorporating the linear restrictions in the estimation process. This estimation process improves the efficiency of the estimators. The Restricted Least Squares (RLS) estimator β^*satisfying the set of q (≤ K) restrictions embodied in $R\beta^* = r$ can be obtained by minimizing the residual sum of squares with respect to β^* subject to the constraints $R\beta^* = r$. It is given by

$$\beta^* = \hat{\beta} + (X'X)^{-1} R' \left[R\,(X'\,X)^{-1}R'\right] (r-R\hat{\beta})$$

The mean vector and covariance matrix of β^* are given by

$$E\,[\beta^*] = \beta + (X'\,X)^{-1}\, R' \left[R\,(X'\,X)^{-1}R'\right] (r-R\beta) \quad \text{and}$$

$$V\,[\beta^*] = \sigma^2\,[(X'\,X)^{-1}\text{-}\ (X'\,X)^{-1}R' \left[R\,(X'\,X)^{-1}R'\right]^{-1} R(X'\,X)^{-1}]$$

Some alternative expressions for F-statistic for testing the general linear hypothesis H_0: $R\beta = r$ using RLS estimator is given by

$$\text{(i) } F = \frac{(\beta^* - \hat{\beta})'\,(X'\,X)(\beta^* - \hat{\beta})/q}{e'\,e/n-K} \sim F_{q,\,n\text{-}K}$$

$$\text{(ii) } F = \frac{[\,e^{*'}e^* - e'e]/q}{e'e/n-K} \sim F_{q,\,n\text{-}K}$$

(iii) $F = \dfrac{[R^2_{OLS} - R^2_{RLS}]/q}{[1 - R^2_{OLS}]/n - K} \sim F_{q,\, n\text{-}K}$

Where, $e^{*\prime} e^{*}$ = Restricted least squares residual sum of squares

$e^{\prime} e$ = unrestricted least squares residual sum of squares

R^2_{OLS} and R^2_{RLS} are respectively the R^2- values obtained from the OLS and RLS regressions. It should be noted that

(i) $R^2_{OLS} \geq R^2_{RLS}$ and

(ii) $e^{*\prime} e^{*} \geq e^{\prime} e$

Summary and Conclusions :

The literature on Regression analysis rapidly fully-fledged used in forecasting models. The Present research paper assesses the regressors in regression analysis and their functional forms with respective time series modeling.

In the present paper, a Regression analysis, intent the inferential techniques for forecasting and validations problems in forecasting problems.

BIBLIOGRAPHY :

1. Amemiya, T.(1980), "Selection of Regressors", International Economic Review, 21, 331-354.
2. BASIC ECONOMETRICS, Fifth Edition, Damodar N. Gujarati & Dawn C. Porter
3. Chalapathi Rao, M.V. (1998), "Some Aspects of Interference in Linear 165 Models with Autocorrelated Disturbances", published Ph.D., Thesis, Dept of Statistics, S.V. University, Tirupati, India.
4. Connifee, D. (1982b), "A Note on Seemingly Unrelated Regressions."

Econometrica, 50, 229-233.

5. DRAPER N.R. and H. Smith (1981): "Applied Regression Analysis" John Wiley & Sons.
6. David A. Belsley Edwin Kuh Roy E. Welsch (1980): "Regression Diagnostics", Chapman and hall, London.
7. Harvey, A.C. (1981), The Econometric Analysis of Time Series, Philip Allan, Oxford.
8. Johnston, J. (1984), "Econometric Methods", Third Edition, Mc Graw-Hill, New York.
9. Lindley, D.V. (1962), "Discussion on Professor Stein's paper". Journal of the Royal Statistical Society, Series B, 24, 285-287.
10. Narayana, P. (2001), "Statistical Inference in Sets of Linear Regression Models", published Ph.D., Thesis, S.V. University, Tirupathi, India.
11. Nafeez Umar, Sk. (2004), "Statistical Inference on Model Specification in Econometrics", published Ph.D., Thesis, S.V. University, Tirupathi, India.