iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Multicollinearity Problem and Some Hypothetical Tests in Regression Model

R.V.S.S. Nagabhushana Rao[1], T. Gangaram[2], C.Narayana[3], K.Vasu[4], G. Mokesh Rayalu[5]
[1]*Assistant professor, Department of Statistics, Vikrama Simhapuri University, Nellore*
[2]*Lecturer, Department of Statistics, PVKN Govt College, Chittoor*
[3]*Assistant professor, Department of Mathematics, Sriharsha Institute of P.G.Studies, Nellore*
[4]*Assistant Professor, Vidya Educational Institutions, Yadamari , Chittoor*
[5]*Assistant Professor, Department of Mathematics, School of Advanced sciences ,VIT, Vellore*

Abstract: In the non-experimental sciences, much of the data used is passively generated. The lack of sufficient information and the ambiguity of statistical results based on information leads to commonly called the multicollinearity problem. Multicollinearity, unfortunately, contributes to difficulty in the specification as well as the estimation of economic relationships. Attempts to apply regression techniques to highly multicollinear independent variables result in parameter estimates that are markedly sensitive to changes in model specification and to sample coverage. Econometricians recognize that multicollinearity imports a substantial bias toward incorrect model specification, and that poor specification undermines the "best linear unbiased" character of parameter estimates over multicollinear independent variable sets.

I. INTRODUCTION

Multicollinearity may have several adverse effects on estimated coefficients in a multiple regression analysis. Consequently it is important that researchers be trained in detecting its presence. Examination of a data for the existence of multicollinearity should always be performed as an initial step in any multiple regression analysis. The statisticians and researchers of many disciplines that employ regression analysis should be aware of the adverse effects of multicollinearity and that may exist in the defection of linear dependencies.

A. Detection

Let C denote the correlation matrix for k regressor variables. Let l_j be the j^{th} latent root of C with corresponding latent vector v_j. Let the matrix of latent vectors be

$$V = [v_1, .v_2 \ldots\ldots v_k]$$

The degree of multicollinearity among the regressor variables is often determined by using one or more of the following measures

1) Extreme pair wise correlation two regressor variables

$|r|_{\max} = \max|C_{ij}|, whereC = VLV^1$ .Small determinant of the correlation matrix where $|C| = \pi_{j=1}^{k} l_j$

2) One or more small latent roots of the correlation matrix. If l_j=0, then an exact linear dependence exists.

3) Large variance inflation factors, VIF(j), are the diagonal elements of the inverse of the correlation matrix, that is, $VIF(j) = C_{ij}^{-1}, C^{-1} = vL^{-1}v^1$

4) Large R_j^2, where x_i is predicted using the reaming regressor variables;

$R_j^2 = 1 - VIF(j)$ *or* $tolerance(j) = 1 - R_j^2 = 1/VIF(j)$

Consider a KxK correlation matrix of the from

$$C = (1-\rho)I + \rho J$$

where I is identity matrix of order K and J is a square matrix of order n with every element equal to –1/k-1<p< 1.0

Graybill (1969), has inverse $C^{-1} = \frac{1}{1-\rho}\left[I - \frac{\rho}{1+(k-1)\rho} J \right]$, consequently, the VIF for the j[th] regressor variable

$$VIF(j) = [1+(K-2)\rho]/[(1-\rho)(1+(k-1)\rho)]$$

The determinant of C is $|C| = [1+(k-1)\rho](1-\rho)^{k-1}$

Linear relationships among regressor variables are best detected by examination of the latent roots and latent vectors of the correlation matrix. For orthogonal data each latent root has a value of 1.0 In general, multicollinearity is a problem when one or more of the latent roots are near zero. The latest vectors corresponding to small latent roots indicate how the standardized regressor variables are involved in the linear dependencies. A large value within a latent vector signifies that the corresponding regressor variable is contributing to the multicollinearity problem. The major problem with multicollinearity is that the least squares estimators of coefficients of variables involved in the linear dependencies have large variances. The VIF is an indicator that provides the user with a measure of how many times larger the var $(\hat{\beta}_j)$ will be for multicollinear data than for orthogonal data advantage of knowing the VIF for each variable is that it gives users a tangible feel of how much the variances of estimated coefficients are affected by the multicollinearity. Hence, one is able to determine how strongly a vaiable, if added to those regressor variables already in the model, will be linearly related to those variables. Examination of the latent roots and latent vectors of the correlation matrix provides the user with a necessary and a sufficient measure of detecting multicollinearity.

B. Multicollinearity And The Mean Square Error

The problem of multicollinearity in regression analysis is essentially a lack of sufficient information in the sample to permit accurate estimation of the individual parameters. More specifically,

Let $Y = \beta X + \gamma Z + u$

Where X and Z are non stochastic, each variable has mean zero, and u is a serially independent random disturbance with mean zero and constant variance. The equation is first estimated by ordinary least squares. In the coefficient of the "nuisance" variable (z) has a low t statistic, the equation is re-estimated with z omitted; if not, the original ordinary least squares (OLS) estimate is retained. Bancroft concentrated on deriving estimates of the bias when γ does not equal zero but did not calculate the MSE of the estimate of β. Toro Vizcarrondo and wallace suggest that to test the null hypothesis that the true t_γ is less than one and, if the null hypothesis is not rejected, omit the collinear nuisance variable. Since this null hypothesis is tested by using the sample t_γ statistic, their procedure differs from that Bancroft only in the critical level of the sample t_γ required to omit z. A variety of mean square error loss functions is presented to indicate the potential gains and losses of different COV estimators. Each COV estimator is defined by an essentially arbitrary test statistic for choosing in any sample between the OLS estimator and an unconditional omitted variable (OV) estimator.

II. THE LOSS FUNCTIONS OF COV ESTIMATORS

Consider the basic model

$Y_t = \beta X_t + \gamma Z_t + u_t$, (t=1, 2 ….T)

E (u_t) = E ($u_t u_{t-s}$) = 0 and $E\left(u_t^2\right) = \sigma^2$, …. (3.1)

where X and z are nonstochastic variables with mean zero. The conditional omitted variable estimator with parameter ξ is defined by

$$\beta\,[\mathrm{cov}(\xi)] = \begin{cases} \hat{\beta} \text{ if } \hat{t}_\gamma \geq \xi \\ \hat{b} \text{ if } \hat{t}_\gamma < \xi \end{cases} \quad \text{…. (3.2)}$$

where $\hat{\beta}$ is the OLS estimator of β in equation (3.2) $\hat{t}_r$ is the ratio of $\hat{r}$ to its sample standard error and $\hat{b}$ is the (OV) estimator of β i.e; the OLS estimator of b in the omitted variable equation.

$$Y_t = bX_t + v_t \qquad \text{…. (3.3)]]}$$

The ratio at the mean square errors of the OV and OLS estimators provides insight into the potential gains and losses of using a COV estimator. The mean square error of the OV estimator is

$$MSE\ (\hat{b}) = E\ (\hat{b} - \beta)^2 = \sigma^2\ (x^1x)^{-1} + (x^1x)^{-2}\ (x^1z)^2\ \gamma^2 \qquad \text{……(3.4)}$$

The mean square error of the OLS estimator is

$$MSE\ (\hat{\beta}) = E\ (\hat{\beta} - \beta)^2 = \sigma^2\ (z^1z)[\ (X^1X)\ (z^1z) - (X^1z)^2\ ]^{-1} \qquad \text{….. (3.5)}$$

The ratio of (3.4) to (3.5) simplifies to

…. (3.6)

where r_{xz} is the correlation between x and z, and t_γ is the absolute value of the ratio of γ to its true standard error i.e; $r_{XZ}^2 = (X^1z)\ (X^1X)^{-1}(z^1z)^{-1}$ and $t_\gamma^2 = \gamma^2\ [\ (X^1X)\ (z^1z) - (X^1z)^2\ ] \big/ \sigma^2\ (X^1X)$.

Equation (3.6) shows that if the regressors are correlated $\left(r_{xz}^2 \neq 0\right)$, the omitted variable estimator has a smaller mean square error than the OLS estimator whenever $t_\gamma^2 < 1$ and a larger mean square error whenever $t_\gamma^2 > 1$.

III. FARRAR AND GLAUBER TEST

Farrar and Glauber (1967) used three test statistics Viz, χ^2, F *and* t for testing the multicollinearity in the given data. χ^2 test is used for the detection of the existence and severity of multicollinearity in a function including several explanatory variables. F-test is used for the location of multicollinearity that is which variables are multicollinear.t-test is used for the pattern of multicollinearity that is which variables are responsible for the appearance of multicollinear variables. Since multicollinearity is a departure from orthogonality we state the null hypothesis as

$H_o : X^1s$ are Orthogonal

i.e, there is no multicollinearity

$\left|X^1X\right| \neq 0\ or\ r_{X_iX_j} = 0\ \forall\ i \neq j$

where is the determinant of correlation matrix i.e.,

$$\text{i.e., } \left|X^1X\right| = \begin{bmatrix} 1 & r_{12} & r_{13} & . & . & . & . & r_{1k} \\ r_{21} & 1 & r_{23} & . & . & . & . & r_{2k} \\ . & . & . & & & & & \\ . & . & . & & & & & \\ . & . & . & & & & & \\ r_{k1} & r_{k2} & r_{k3} & . & . & . & . & 1 \end{bmatrix} \qquad \text{…. (4.1)}$$

In case of perfect multicollinearity $\left|X^1X\right|$ = 0.On the other side in the case of orthogonal $\left|X^1X\right| = 1$ Since $r_{xixj} = 0$; $i \neq j = 1, 2, ...$

If the value of $\left|X^1X\right|$ lies between 0 and 1 there exists some degree of multicollinearity. For detecting the degree of multicollinearity over the whole set of explanatory variables "Farrar and Glauber" suggest the following χ^2 test.

To test the Ho, χ^2 test statistic is given by

$$X^2 = -\left[(n-1)-(\frac{2k+5}{6})\right]\log 1X^1X1 \sim \chi^2_{\frac{K(K-1)}{2}} \quad \ldots\ldots (4.2)$$

when n= number of observations
K= number of explanatory variables
X^1X= correlation matrix

Compare χ^2 calculated value with χ^2 central value. If $\chi^2 cal > \chi^2_{\frac{k(k-1)}{2}}$ we reject Ho. Otherwise accept H_O. If we accept Ho, then we say that there is no significant multicollinearity in the function. The higher value of χ^2 calculated indicates the more severe multicollinearity.

IV. F – TEST

To locate the variables which are multicollinearity Farrar and Glauber computed the coefficients of multiple determination among the explanatory variables in the model. Let R_i^2 denotes the coefficient of Multiple determination of i^{th} explanatory variable regressed on the remaining (K-1) explanatory variables. To test the significance of R_i^2, we use the F_i statistic. The F_i statistic is given by

$$F_i = \frac{R_i^2/(k-1)}{(1-R_i^2)(n-k)} \sim F_{(k-1,n-k)} \quad \ldots\ldots (5.1)$$

Where n = number of observations

k = number of explanatory variables

If F_i cal $\leq F_{(k-1,\ n-k)}$, we accept H_o. Then we can infer that variable X_i is not multicollinear. If we reject H_0, then F_i is significant, we say that the i^{th} explanatory variable X_i is most affected by multicollinearity.

V. T-TEST

This test helps to detect the variables which are responsible for multicollinearity. We compute the partial correlation coefficients among explanatory variables and then test for their significance by the student t- test.

$$Ho: r_{x_i x_j . x_1 x_2 \ldots x_k} = 0$$

i e; X_i and X_j are not multicollinear

The t-test statistic for the significance of partial correlation coefficient between X_i and X_j; $i \neq j$ is given by

$$t = \frac{r_{XiXj.X_1X_2----X_k}\sqrt{n-k}}{\sqrt{1-r^2}_{XiXj.X_1X_2-----X_k}} \sim t_{n-k}; i \neq j = 1,2,\ldots k \quad \ldots\ldots (6.1)$$

where n = number of observations
k= number of explanatory variables

If t cal $\leq$ t cri value we accept Ho. Then we infer that variables X_i and X_j are not responsible for the multicollinearity. Otherwise we reject Ho, so t is significant, then we say that the variable X_i and X_j are responsible for the multicollinearity in the function.

VI. SIMPLE CORRELATIONS AMONG REGRESSORS

The detection of multicollinearity by Farrar and Glauber (1967), actually involves three aspects I) determining its presence, 2) its severity and 3) its location or form in a set of data. Some commonly suggested detection measures and procedures are appraised. A commonly quoted rule is that if a simple descriptive measure in the form of a correlation coefficient between two regressors is greater than 0.8 or 0.9, then multicollinearity is a serious problem. A more elaboration version of this rule compares the simple correlation coefficients to R^2, the coefficient of determination; multicollinearity is deemed harmful if the simple correlation is greater than R^2. An intuitive explanation of this rule is given by Farrar and Glauber (1967). Unfortunately, the only certain information simple correlations can provide occurs when one of them is unity. Then the observation matrix will be singular and

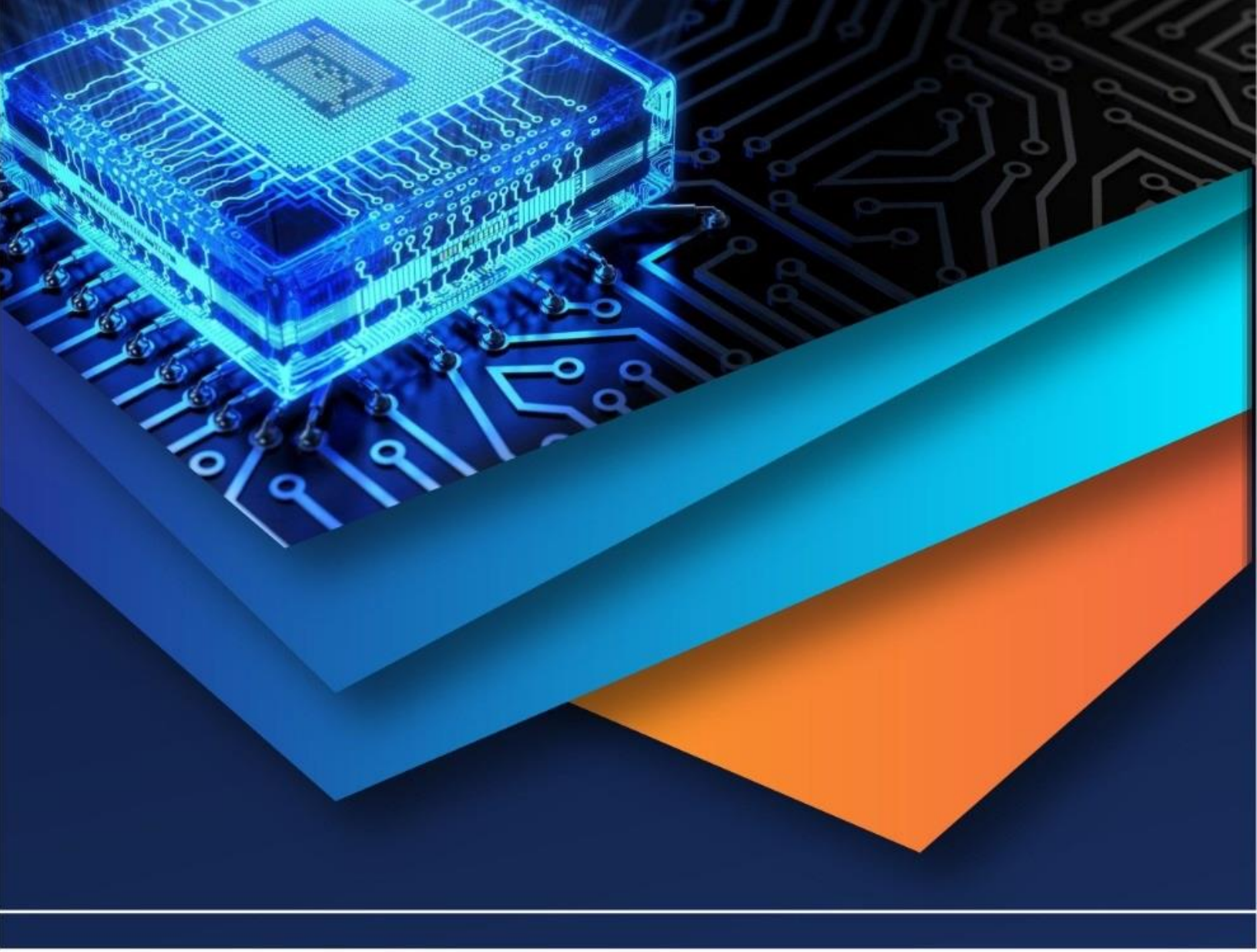
unbiased, least squares estimates are not available for all parametric functions. Furthermore, it is clear that if a linear dependency involves more than two regressors, pair wise sample correlation coefficients provide no information about them.

VII. CONCLUSIONS

Smith (1974) points out, by appropriate model transformation and scaling it is always possible to produce a model whose variables are orthogonal and whose characteristic roots are unity. If several near-exact linear dependencies are present do not provide a complete solution to the problem of detecting and identifying structural relations associated with poor implicit sample design. Belsley, Kuh and Welsh (1980) provide a set of condition indexes that identify one or more near dependencies. Furthermore they adopt the silvery regression variance decomposition so that it can used with the indexes to isolate the variables involved and to assess the extent of distortion due to the near dependencies. In this present paper we discuss about various hypothetical tests involved in Multicollinearity problem in criteria for model selection.

REFERENCES

[1] Bacon, R.W., and Hausman, J.A., (1974), The Relationship between Ridge regression and the minimum Mean squared Error estimator of chipman, oxford Bulletin of economics and statistics, 36, pp:115-124.
[2] Belsey, D.A., and Klenma, V.C. (1974), Detecting and Assessing the problems caused by multicollinearity: A use of the singular value decomposition, National Buireau of Economic Research, Cambridge
[3] Belsley, David, A., Edwin Kuh, and Welsch, Roy, E., (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collineartiy, John Willey & Sons, New York.
[4] Breusch, T.S.(1980) "Useful Invariance Results for Generalized Regression Models", Journal of Econometrics, 13, 327-340.
[5] Fomby, T.B., and Johnson, S.R.. (1977). Mean square error evaluation of ridge estimators based on stochastic prior information, communications in dddfddfdskjthe kdstatistics, 6, pp:1245-1258.Gibbons,D.G. (1981), "A simulation study of some Ridge Estimates", 76, pp: 131 – 139.
[6] Giles, D.E.A. (1985), "A Note on Regression Estimation with extraneous information" Journal of Quantitative Economics, 1, pp. 152 – 159.
[7] Goldbergen, A.S. (1968), Topics in Regression Analysis, Macmillan, New York.Hoerl, A.E., Kennard, R.W. and Baldwin, K.T. (1975). "Ridge Regression: Some Simulations, "Communications in Statistics", 2, PP:105-123.Huang, D.S. (1970), Regression and Econometric Methods, John Wiley & Sons, New York.
[8] Jodge, G.G. and Bock, M.E (1978), the statistical Implications of Pre-test and Stein-Rule Estimators in Econometrics; North-Holland Publishing Company; Amsterdam.
[9] Lindley, D.V., and Smith, A.F.M. (1972), Bayes Estimates for the Linear Model, Journal of Royal Statistical Society., 34, pp:1-41.
[10] Maddala, G.S. (1992), Introduction to Econometrics, Macmilla, Second Edition, New York.Mansfield, E.R. and Helms, B.P. (1982). Detecting Multicollinearity, The American Statistician, 36, pp:157-160.
[11] Sarkar, N. (1992), A New Estimator Combining the Ridge Regression and the Restricted Least Squares Methods of Estimation, Communications in Statistics, Indian Statistical Institute, pp:1987-2000.
[12] Schmidt, P. (1976) "Econometrics", Marcel. Dekker, New York.Smith, A.F.M. and Goldstein, M (1975), Ridge Regression : Some Comments on a paper of conniffe and stone, Mathematical Institute, University of Oxford, 24, pp:61-65.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY