Estimation of Multicollinear Statistical Model Using Orthonormalmoore – Penrose Generalized Inverse Matrix

DrM.Pushpalatha[1] , Dr K.Sreenivasulu[2]
[1, 2]Lecturer, Dept of Statistics
[1, 2] S.P.W.Degree and P.G.College,Tirupati

Abstract*- Non-experimental sciences rely heavily on data that is generated indirectly. Multicollinearity is a problem that arises when there aren't enough data points or when the data used to draw conclusions isn't clear. Unfortunately, the estimate of linear statistical models is complicated by the problem of multicolliearity. Extremely multicollinear explanatory variables might produce parameter estimates that are extremely sensitive to changes in the model's specification and pattern coverage. Statistics experts are aware that multicollinearity significantly increases the likelihood of improper model specification and that suboptimal model specification weakens the high linear independence of parameter estimates over multicollinear explanatory variable units.*

This paper develops a method for predicting the parameters of a linear statistical model while addressing the multicollinearity issue using orthonormal and morepenrose generalised inverse matrices. In addition to the new estimator, the generalised mean squared errors for the estimate have also been determined.

I. INTRODUCTION

The majority of the information used in the non-experimental sciences is generated through other means. The multicollinearity problem is a term that refers to what happens when there is insufficient information as well as ambiguity in the statistical results that are based on the information. Sadly, the issue of multicollinearity contributes to the difficulties that is involved in the estimation of linear statistical models. When regression techniques are attempted to be applied to explanatory variables that are highly multicollinear, the resulting parameter estimates are extremely sensitive to changes in the model specification as well as variations in the amount of data collected.

The concept of multicollinearity refers to a condition of interdependence that can take place completely independently of the nature, or even the existence of, or the reliance that exists between X and Y.

The presence of multicollinearity poses a risk, and in many cases, a significant risk, to the accurate and efficient estimation of the kind of structural relationship that is typically sought through the application of regression techniques. Multicollinearity is a threat.

II. MULTICOLLINEARITY IN STATISTICAL MODEL

Estimating the parameters of a dependency connection, as opposed to the parameters of an interdependency relationship, is the goal of regression analysis. First, we establish that Y and X are connected to one another in a linear fashion.

$$Y = X\beta + \in \qquad \ldots..\ (2.1)$$

Y, X as a sample of N observations on one dependent and n independent variables, each of which is normalized to unit length.
β as a vector of true (structural) coefficients
$\in$ as a true error term, with distributional properties specified by the general linear model.
Least squares regression analysis leads to estimates

$$\hat{\beta} = (X^1X)^{-1}\ X^1Y$$

with variance – covariance matrix

$$V(\hat{\beta})\ = \sigma_{\in}^2\ (X^1X)^{-1}$$

As interdependence among explanatory variables x grows, the correlation matrix (X^1X) approaches singularity, and elements of the inverse matrix $(X^1X)^1$ explode.

III. REASONS FOR MULTICOLLINEARITY

A cause for multicollinearity is the employment of lagged values of several explanatory components as separate explanatory variables inside the courting process.

1. This is among the most important factors that contribute to multicollinearity. As a consequence of this, the existence of multicollinearity inside the dispersed lag models can be seen to be absolutely guaranteed.
2. The connections that exist between the many different aspects of the economy typically get stronger over the course of time. For instance, during periods of economic expansion, wages, savings, intake, investment, rate levels, and employment levels, amongst other things, tend to increase. On the other hand, during times of economic contraction, these same factors normally tend to decrease. As a direct consequence of this, the aspects of growth and fashion in time series are the ones that contribute to multicollinearity the most significantly.
3. As a result of the imprecision of the dimensioning, there is the possibility of multicollinearity.
4. It is possible, but not guaranteed, that the model will consist of a higher number of explanatory variables than there are observations. However, this is the case.

IV. ADVERSES OF MULTICOLLINEARITY

The main adverses of multicollinearity are :

i) As a result of the decline in estimating precision, it becomes exceedingly challenging to arrive at accurate estimations of the regression coefficients. There are three aspects to the reduction in precision.

A) Precise estimates are prone to extremely significant error.
B) There is a strong possibility that these errors are intertwined with one another.

C) The coefficients' sampling variances can sometimes be quite large in magnitude.

I) Estimates of regression coefficients become extremely sensitive to particular sets of sample records, and the addition of some greater observations or the deletion of a few observations can sometimes produce dramatic shifts in some of the coefficients. This sensitivity is due to the fact that regression coefficients tend to be positively correlated with the number of observations used in the analysis.

Ii) If there is a high level of multicollinearity, one may also incur a high cost in terms of R2, and only a select fraction of the calculated regression coefficients will be statistically significant.

Iii) The possibility of falling for a false hypothesis rises as a direct result of the prevalence of huge preferred blunders. To put it another way, the significance of the examination will be diminished.

V. ESTIMATION OF MULTICOLLINEAR STATISTICAL MODEL

Consider the linear regression model,

$$Y_{nx1} = X_{nxk}\,\beta_{kx1} + \in_{nx1} \qquad \text{….(5.1)}$$

$$E\lfloor \in \rfloor = 0,\; E\left[\in\in^1\right] = \sigma^2 I_n$$

Rank of $X \le K$

Under the problem of multicollinearity, the Ridge Regression estimator for β is given by

$$\hat{\beta}(\delta) = \left[X^1X + \delta I\right]^{-1} X^1Y, \delta > 0 \qquad \text{….(5.2)}$$

Let $[P:Q]$ be a $(K \times K)$ orthonormal matrix; P and Q are $(K \times r)$ and $[K \times (K-r)]$ submatrices ; T is a $(r \times r)$ diagonal matrix with positive diagonal elements; and r is rank of matrix X. Then the canonical form of (X^1X) can be expressed as

$$X^1X = [P \vdots Q]\begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} P^1 \\ Q^1 \end{bmatrix} = PTP^1$$

….(5.3)

Further one may have $Q^1X^1XQ = 0$ so that XQ=0
Now (5.2) can be written as

$$\hat{\beta}(\delta) = P\left[T + \delta I\right]^{-1} P^1X^1Y \qquad \text{….(5.4)}$$

we have , $\hat{\beta}(0) = \text{Lt}\ \hat{\beta}(\delta) = (X^1X)^g X^1Y\ \delta \to 0$

….(5.5)where

$(X^1X)^g = PT^{-1}P^1$ is the More – Penrose generalized inverse of matrix X^1X .

$\hat{\beta}(0)$ gives the OLS solution of βsatisfied the normal equations

$$X^1X\hat{\beta}(0)=X^1Y$$

….(5.6)

For some d, if C= Pd, an estimable function of β is given by C¹β. It can be shown that the necessary and sufficient condition for the

$$MSE\left[C^1\hat{\beta}(\delta)\right]<MSE\left[C^1\hat{\beta}(0)\right]$$ is that

$$\beta^1\left[\tfrac{2}{\delta}I+(X^1X)^g\right]^{-1}\beta<\left[\sigma^2+\tfrac{\delta}{2}\eta^1\eta\right]$$

….(5.7)

where, $\eta=Q^1\beta$

The condition (5.3.7) is satisfied if, $\delta\le\frac{2\sigma^2}{\Lambda^1\Lambda}$

….(5.8)

or $\left[\frac{2}{\delta}+\frac{1}{\lambda}\right]^{-1}\Lambda^1\Lambda<\sigma^2$

….(5.9)

or $\beta^1X^1X\beta\le\sigma^2$

….(5.10)

Where $\Lambda=P^1\beta \text{ and } \lambda$ is the largest given value of (X^1X).

By substituting $pp^1+QQ^1=I$ and $XQ=0$ in (5.1), one may obtain the model as

$$Y=Z\Lambda+\in$$ ….(5.11)

$$E\left[\in\right]=0,\ E\left[\in\in^1\right]=\sigma^2I$$

where $Z=XP,\Lambda=P^1\beta$

The OLS estimator of Λ is given by

$$\hat{\Lambda}(0)=(Z^1Z)^{-1}Z^1Y$$

….(5.12)

and $\hat{\Lambda}(\infty)=0,$ Here, $\hat{\Lambda}(\delta)$ is the Ridge Regression estimator of $\Lambda=P^1\beta$. Also, $\hat{\Lambda}(0)=P^1\hat{\beta}(0) \text{ and } \hat{\Lambda}(\infty)=P^1\hat{\beta}(\infty).$

The equivalence condition of (5.7) is given by

$$\left\{\hat{\Lambda}\left[\tfrac{2}{\delta}I_r+T^{-1}\right]^{-1}\Lambda+\tfrac{\delta}{2}\eta^1\eta\right\}<\left[\sigma^2+\tfrac{\delta}{2}\eta^1\eta\right]$$

….(5.13)

$$\Rightarrow\ \Lambda^1\left[\tfrac{2}{\delta}I_r+T^{-1}\right]^{-1}\Lambda<\sigma^2$$

….(5.14)

We have, $MSE\left[C^1\hat{\beta}(\delta)\right]<MSR\left[C^1\hat{\beta}(0)\right]$

….(5.15)

$$\forall C=Pd\neq0 \text{ and } \forall\delta>0$$

iff $MSE\left[d^1\hat{\Lambda}(\infty)\right]\le MSE\left[d^1\hat{\Lambda}(0)\right],\forall d\neq0$ ….(5.16)

By using Internally studentized residual sum of squares, the necessary and sufficient condition for satisfying (5.15) and (5.16) is given by

$$\tilde{\delta}\le\left[\frac{2\tilde{\sigma}^2}{\hat{\Lambda}^1(0)\hat{\Lambda}(0)}\right]$$

….(5.17)

Thus, an optimum value for δ in the Ridge Regression estimator $\hat{\beta}\left[\delta\right]=\left[X^1X+\delta I\right]^{-1}X^1Y$ is given by

$$\tilde{\delta}=\frac{2\tilde{\sigma}^2}{\hat{\Lambda}^1(0)\hat{\Lambda}(0)}$$ ….(5.18)

The Ridge Regression estimator of β is given by

$$\hat{\beta}(\tilde{\delta})=\left[X^1X+\tilde{\delta}I\right]^{-1}X^1Y$$ ….(5.19)

The Generalized mean squared error matrix of $\hat{\beta}(\tilde{\delta})$ is given by

$$MSE\left[\hat{\beta}(\tilde{\delta})\right]=\left[(X^1X)+\tilde{\delta}I\right]^{-1}\left[\tilde{\sigma}^2X^1X+\tilde{\delta}^2\hat{\beta}\hat{\beta}^1\right]\left[X^1X+\tilde{\delta}I\right]^{-1}$$

….(5.20)

The mean square error matrix of $p^1\hat{\beta}(\tilde{\delta})$ is given by

$$MSE\left[\hat{\Lambda}(\tilde{\delta})\right]=\left[T+\tilde{\delta}I\right]^{-1}\left[\tilde{\sigma}^2T+\tilde{\delta}^2\hat{\Lambda}(0)\hat{\Lambda}(0)^1\right]\left[T+\tilde{\delta}I\right]^{-1}$$

….(5.21)

The mean square error matrix of $\hat{\Lambda}(0)$ is given by

$$\mathrm{MSE}\left[\hat{\Lambda}(0)\right] = \mathrm{MSE}\left[P^1\hat{\beta}(0)\right] = Z^2T^{-1} \quad \text{….(5.22)}$$

We have,

$$\mathrm{MSE}\left[\hat{\Lambda}(0)\right] - \mathrm{MSE}\left[\hat{\Lambda}(\tilde{\delta})\right] = \tilde{\delta}\left[T+\tilde{\delta}I\right]^{-1} F\left[T+\tilde{\delta}I\right]^{-1} \quad \text{….(5.23)}$$

Which is positive definite for $\tilde{\delta} > 0$ iff

$$F = \left[2\tilde{\sigma}^2 I + \hat{\delta}\tilde{\sigma}^2 T^{-1} - \tilde{\delta}\hat{\Lambda}(0)\hat{\Lambda}^1(0)\right] \quad \text{….(5.24)}$$

is positive definite. But F is positive definite iff

$$\hat{\Lambda}^1(0)\left[\tfrac{2}{\tilde{\delta}} I + T^{-1}\right]^{-1} \hat{\Lambda}(0) < \tilde{\sigma}^2 \quad \text{…. (5.25)}$$

IV. CONCLUSIONS

When trying to estimate the parameters of a linear statistical model, multicollinearity presents a number of challenges, the most significant of which is that the least squares estimators of the coefficients of variables involved in linear relationships have greater variances. In the linear statistical form, multicollinearity is essentially a lack of sufficient records within the sample to permit correct estimation of the individual parameters. In the current study, based on the moore-penrose generalised inverse matrix, a proposal was made for an orthonormal ridge regression estimator for the parameter vector.

REFERENCES

[1] Goldstein, M., and Smith, A.F.M. (1974). Ridge type estimators for regression analysis, Journal of the Royal Statistical Society, series B, 36, pp:284-291.

[2] Gunst, R.F. (1984). “Toward a Balanced Assessment of Collinearity Diagnostics,” The American Statistician, 38, pp:79-82.

[3] Hoerl, A.E. (1962). Application of ridge Analysis to Regression problems, chemical engineering progress, 58, pp:54-59.

[4] Pushpalatha M., Bhupathi Naidu M., Balasiddamuni P (2013), “Statistical Inference in Linear Regression Models Under Multicollinearity”, LAP LAMBERT Academic.

[5] Hoerl, A.E. and Kennard, R.W. (1970). “Ridge Regression: Biased Estimation for Non-orthogonal problems”, Technometrics, 12; pp:55-67.

[6] Johnston, J. (1984). Ecnometric Methods, Third Edition, MaGraw-Hill, New York.

[7] Judge, G.G., Griffiths,W.E. (1985), “Theory and practice of Economics”, Second edition, John Wiley & Sons, New York.

[8] Marquardt, D.W., and Snee, R.D. (1975), “Ridge Regression in practice”, Journal of American statistical Association, 29, pp: 3 – 19.

[9] Mason,R.L., Gunst, R.F.,and Webster, J.T.(1975), “Sources of Multicollinearity in Regression Analysis”, Communications in Statistics, 4,pp:277-292.

[10] McDonald, G.C., and Galarneau, D.I. (1975). A monte carlo Theil, H., (1971), Principles of Econometrics, New York: Wiley evaluation of some ridge type estimators, JASA, 70,