Journal of Management
Vol.VI.No.1.October 2010.pp.1 – 12

Collinearity Affects and It′s Analysis in Data

Aboobacker Jahufer
Department of Mathematical Sciences
Faculty of Applied Sciences
South Eastern University of Sri Lanka
Sri Lanka
Email: jahufer@yahoo.com

Abstract

One of the assumptions of the multiple linear regression model is that there is no exact linear relationship between any of the independent variables. If such a linear relationship does exist, it can be said that the independent variables are collinear or multicollinearity.

When collinearity exists among the regressors, a variety of interrelated problems are created. Specially, in the model building process collinearity causes high variance for parameters if ordinary least squares estimator (OLSE) is used. The main objective of this research paper is to analyze and detect the collinearity in the data set and recommend some important dealing methods for collinearity problems. Two collinearity data sets are used to illustrate the methodologies proposed in this research paper. The first data set was generated using Monte Carlo Simulation method with the highest correlation between the regressors and this data set contains five regressors and a response variable. The second data set is also a real collinearity data set of Macroeconomic Impact of Foreign Direct Investment in Sri Lanka form 1978 to 2004 and it contains four regressor and one response variables.

Keywords: Collinearity; Correlation Matrix; Eigen Analysis; Variance Inflation Factor; Conditional Indices; Variance Decomposition; Biased Estimation.

Introduction

In many situations both experimental and non-experimental, the regressors tend to be correlated. Then collinearity or collinearity exists among the regressors. A variety of interrelated problems are created when collinearity presents. Specially, in the model building process collinearity causes high variance for parameters if OLSE is used.

Unfortunately in most applications of regression analysis, the regressors are not orthogonal. Sometimes the lack of orthogonal is not serious. However, in some situations the regressors are nearly perfectly linearly related and in such cases the inferences based on the regression model can be misleading or erroneous. When there are near linear dependences between the regressors, the problem of collinearity is said to be exist.

The collinearity is a form of ill-conditioning in the **X'X** matrix. Furthermore the problem is one of degree; that is, every data set will suffer from collinearity to some extent unless the columns of **X** are orthogonal. As we can see, the presence of collinearity can make the usual OLS analysis of the regression model dramatically inadequate (See: Montgomery and Peck 1992; Quirino 2001; Draper and Smith 1998; Allen 1974; Afifi and Clark 1996).

This paper is composed into six sections. Section 2 derives the affects of collinearity. Section 3 describes the collinearity diagnostics. Section 4explains the methods for dealing with collinearity. In section 5, two data sets are analyzed for numerical illustrations. Comments are given in the last section.

Affects of Collinearity

Affects of Collinearity in OLSE

The presence of collinearity has a number of potentially serious affects on the OLS-estimates of the regression coefficients.Some of these effects may be easily demonstrated. Suppose that there are only two regressor variables, X_1 and X_2. The model, assuming that X_1, X_2 and Y, are scaled to unit length, then regression model is

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{1}$$

and the OLS normal matrix equation is:

$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$, where **X** is an n×2 matrix, $\boldsymbol{\beta}$ is a 2×1 unknown vector and **Y** is an n×1 vector.

The estimates of the regression coefficients are

$$\hat{\beta}_1 = (r_{1y} - r_{12}r_{2y})/(1-r_{12}^2), \quad \hat{\beta}_2 = (r_{2y} - r_{12}r_{1y})/(1-r_{12}^2) \tag{2}$$

where r_{12} is the simple correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ and r_{iy} is the correlation between X_i and Y, i = 1, 2.

If there is strong collinearity between $\mathbf{X}_1$ and $\mathbf{X}_2$, then the correlation coefficient r_{12} will be large. From equation (2) we can say that, when $|r_{12}| \to 1$, $Var(\hat{\beta}_i) = \dfrac{\sigma^2}{(1-r_{12}^2)} \to \infty$ and

$Cov(\hat{\beta}_1, \hat{\beta}_2) = \dfrac{-r_{12}\sigma^2}{(1-r_{12}^2)} \to \pm\infty$ depending on whether r_{12} closed to positive one (+1) or negative one (-1). Therefore strong collinearitybetween $\mathbf{X}_1$ and $\mathbf{X}_2$ results in large variance and covariance for the OLSE of the regression coefficients. This implies that different samples taken at the same **X** levels could lead to widely different estimates of the model parameters.

When there are more than two regressor variables, collinearity produces similar effects. It can be shown that the diagonal elements of the $\mathbf{C}=(\mathbf{X'X})^{-1}$ matrix are

$$C_{jj} = \frac{1}{(1-R_j^2)}, \qquad j = 1, \ldots, p, \qquad (3)$$

where R_j^2 is the coefficient of multiple determination from the regression of X_j on the remaining p-1 regressor variables. If there is strong collinearity between X_j and any subset of the other (p-1) regressors, then the value of R_j^2 will be close to unity. Since the variance of $\hat{\beta}_j$ is $Var(\hat{\beta}_j) = C_{jj}\sigma^2 = (1-R_j^2)^{-1}\sigma^2$, strong collinearity implies that the variance of the least squares estimate of the regression coefficient $\hat{\beta}_j$ is very large. Generally the covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$ will also be large if the regressors$\mathbf{x}_i$ and $\mathbf{x}_j$ are involved in a collinearity relationship.

Indications of Collinearity

An estimated model with high standard errors and low t statistics could be indicative of collinearity, but it could alternative suggest that the underlying model is a poor one. One can test the following methods to detect the presence of collinearity in the data.

1. A relatively high R^2 in an equation with few significant t-statistics is one indicator of collinearity. In fact, it is possible that the F-statistic for the regression equation will be highly significant, while none of the individual t-statistics are themselves significant.
2. Relatively high simple correlation between the regressors may indicate collinearity.
3. A number of formal tests for collinearity have been proposed; here some useful methods are suggested and the popular methods are: Examination of the Correlation Matrix, Variance Inflation Factor (VIF) and Eigen Analysis of **X'X.**

Collinearity Diagnostics

Several techniques have been proposed for detecting collinearity. Here some important and useful diagnostics measures are discussed. Desirable characteristics of a diagnostics procedure are that it directly reflect the degree of the collinearity problem and provide information helpful in determining which regressors are involved.

Examination of the Correlation Matrix

A very simple measure of collinearity is inspection of the off-diagonal elements r_{ij} in **X'X** matrix, when regressors are standardized into unit length scaling system. If regressors$\mathbf{x}_i$ and $\mathbf{x}_j$ are nearly linearly dependent, then $|r_{ij}|$ will be near unity, where r_{ij} is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Variance Inflation Factor (VIF) Analysis

The diagonal elements of the $\mathbf{C} = (\mathbf{X'X})^{-1}$ matrix are very useful in detecting collinearity. Recall from (3) that C_{jj}, the j^{th} diagonal element of **C**, can be written as $C_{jj} = (1-R_j^2)^{-1}$. If $\mathbf{x}_j$ is nearly orthogonal to the remaining regressors, R_j^2 is small and C_{jj} is close to unity, while if $\mathbf{x}_j$ is nearly linearly dependent on some subject of the remaining regressors, R_j^2 is near unity and C_{jj} is large.

Since the variance of the j^{th} regression coefficients is $C_{jj}\sigma^2$. It can be viewed C_{jj} as the factor by which the variance of $\hat{\beta}_j$ is increased due to near linear dependencies among the regressors.

The diagonal elements of the $\mathbf{C} = (\mathbf{X'X})^{-1}$ matrix is called VIF, this terminology is due to Marquardt (1970). The $j^{th}VIF_j$ is

$$VIF_j = C_{jj} = (1-R_j^2)^{-1}. \quad (4)$$

The VIF for each term in the model measures the combined effect of the dependencies among the regressors on the variance of that term. One or more large VIFs indicate collinearity among the regressors. Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of collinearity.

Eigensystem Analysis of X'X

The characteristic root or eigenvalues of **X'X** say $\lambda_1, \lambda_2, \ldots, \lambda_p$, can be used to measure the extent of collinearity in the data. If there are one or more near linear dependencies in the data, then one or more of the characteristic roots will be small. The condition number of **X'X**is defined as

$$K = \frac{\lambda_{\max}}{\lambda_{\min}}. \quad (5)$$

Generally if the condition number K<100, there is no serious problem with collinearity. If 100<K<1000 imply moderate to strong collinearity, and if K>1000, severe collinearity is indicated.

Eigensystem analysis can also be used to identify the nature of the near linear dependencies in the data. The **X'X** matrix may be decomposed as

$$\mathbf{X'X} = \mathbf{T\Lambda T'}, \quad (6)$$

where**Λ** is a p ×p diagonal matrix whose main diagonal elements are the eigenvalues λ_j (j=1, …, p) of **X'X** and **T** is a p×p orthogonal matrix whose columns are the eigenvectors of **X'X**. Let the column of **T** be denoted by $\mathbf{t}_1, \ldots, \mathbf{t}_p$. If the eigenvalue λ_j is close to zero, indicating a near linear dependency in the data, the elements of the associated eigenvector $\mathbf{t}_j$ describe the nature of this linear dependency. Specifically the elements of the vector $\mathbf{t}_j$ are the coefficients $t_1, \ldots, t_p$.

Belsey, Kuh, and Welsch, (1980) propose a similar approach for diagnosing collinearity. The n×p **X** matrix may be decomposed as

$$\mathbf{X} = \mathbf{UDT'} \quad (7)$$

where**U** is an n×p, $\mathbf{U'U} = \mathbf{I}$, $\mathbf{T'T} = \mathbf{I}$, and **D** is a p×p diagonal matrix with nonnegative diagonal elements μ_j, j = 1, …, p. The μ_j are called the *singular-values* of **X** and **X** = **UDT'** is called the *singular-value-decomposition* of **X**. The singular-value decomposition is closely related to the concepts of eigenvalues and eigenvectors, since **X'X** = (**UDT'**)'(**UDT'**) =$\mathbf{TD}^2\mathbf{T}'$ = **TΛT'**, so that the squares of the singular values of **X** are the eigenvalues of **X'X**. Here **T** is the matrix of eigenvectors of **X'X** defined earlier, and **U** is a matrix whose column are the eigenvectors of associated with the p nonzero eigenvalues of **XX'**.

Ill-conditioning in **X** is reflected in the size of the singular values. There will be one small singular value for each near linear dependency. The extent of ill-conditioning depends on how small the singular value is relative to the maximum singular value μ_{max}. Belsey, Kuh, and Welsch, (1980) define the *condition-indices* of the **X** matrix as $\eta_j = \frac{\mu_{\max}}{\mu_j}$, j = 1, …, p.

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1} = \sigma^2\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}' \qquad (8)$$

and the variance of the j^th^ regression coefficient is the diagonal element of this matrix, or $Var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{p} \frac{t_{ji}^2}{\mu_i^2} = \sigma^2 \sum_{i=1}^{p} \frac{t_{ji}^2}{\lambda_i}$. Note also that apart from σ^2, the j^{th} diagonal element of **TΛT'** is the j^{th} variance inflation factor, so $VIF_j = \sum_{i=1}^{p} \frac{t_{ji}^2}{\mu_i^2} = \sum_{i=1}^{p} \frac{t_{ji}^2}{\lambda_i}$. Clearly one or more small singular values (or small eigenvalues) can dramatically inflate the variance of $\hat{\beta}_j$. Belsey, Kuh, and Welsch, (1980) suggest using variance- decomposition proportions, defined as $\pi_{ij} = \frac{t_{ji}^2 / \mu_i^2}{VIF_i}$, i, j = 1, …, p, as measures of collinearity. Suppose, if π_{32} and π_{34} are large, the third singular value is associated with a collinearity that is inflating the variances of $\hat{\beta}_2$ and $\hat{\beta}_4$. The variance-decomposition proportions greater than 0.5 is recommended guidelines for collinearity.

Methods for Dealing with Collinearity

Several statistical techniques have been proposed for dealing with the problems caused by collinearity. Some important methods are: (i) Collecting additional data, (ii) Model respecification and (iii) The use of estimation methods other than least squares that are specifically designed to combat the problems induced by collinearity that is called biased estimation technique.

Biased Estimation Technique

When collinearity presents among regressors the biased estimates are more reliable than OLS estimates in that they have smaller mean square error. This means that on average they will come closer to estimating the true model parameters than the OLS-estimates. Because of this property, biased estimation often applied to problems where there is a large amount of collinearity among the predictor variables and the OLS-estimates are unstable.

The ridge regression estimator (RRE) was originally introduced by Hoerl (1964) and Hoerl and Kennard (1970a, b) is given by

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X'X} + k\mathbf{I})^{-1}\mathbf{X'Y}, \quad (9)$$

where k>0 is the ridge estimator biasing parameter. It is a *biased*estimator, however, the variances of its elements are lessthan the variance of the corresponding elements of the OLSE forsuitable k. By accepting some bias to reduce variance, the meansquared error (MSE) might thus be improved.

In this research paper biased estimator ordinary ridge regression is used to fit multiple linear regression model for the collinearity data and the stochastic properties of ridge regression is compared with OLSE stochastic properties.

Stochastic Properties Analysis

The following stochastic properties of OLSE and RRE important and useful to compare and select the suitable estimator.

(a) Standard error of the parameters
(b) 95% confidence interval (CI) for parameters
(c) Mean squared error of the model
(d) Co-efficient of determination of the model
(e) The scalar mean squared error of the parameter vector
(f) Mean squared error matrix of the parameters

Numerical Illustrations and Results

The Monte Carlo Simulation Data Set

To analyze collinearity problem five independent and one dependent variables are generated by using the method of Monte Carlo Simulation. One hundred observations are generated for each variable.

In this paper, McDonald, and Galarneau, (1975) method is used to generate Monte Carlo simulation random variables (data) and the procedure is given below:

$$X_{ij} = (1-\rho^2)^{1/2} Z_{ij} + \rho Z_{i6},\ Y_i = (1-\rho^2)^{1/2} Z_i + \rho Z_{i6}$$
$$, \text{i} = 1, \ldots, 100,\ \text{j} = 1, \ldots, 5, \quad (10)$$

whereZ_{ij}, is independent standard normal pseudo-random numbers and ρ^2 is correlation between any two explanatory variables and ρ = 0.99. These variables are then standardized so that **X'X** is in a correlation form.

The estimated Durbin-Watson value for this data set is **2.1788**. The critical lower (d_l) and upper (d_u) values at 1% significance level is $\boldsymbol{d_l}$ **= 1.44** and $\boldsymbol{d_u}$ **= 1.65** (for sample size 100 and regressors 5), respectively. The estimated value for this model lies between $\boldsymbol{d_u}$ and **4-**$\boldsymbol{d_u}$, hence, it can be conformed that at the **1%** level of significance there is no autocorrelation.

Collinearity results and analysis for Monte Carlo Simulation Data

(a) Correlation Analysis: The generated variables are standardized into unit length scaling system. A simple measure of collinearity is to inspect the off-diagonal elements of **X'X** matrix.

The off-diagonal elements of **X'X**values are greater than 0.9718 and all thesesvalues are closed to oneso it can be conformed that strong collinearity exists among the regressors.

(b) VIF Analysis: The diagonal elements of $(\mathbf{X'X})^{-1}$ is called VIF and it is used to measure the collinearity. The diagonal elements of $(\mathbf{X'X})^{-1}$are 28.7508, 38.2249, 34.2826, 36.1997, 35.1739.

The diagonal elements are greater than 5 and in respect of these values the collinearity is conformed among the regressors.

(c) Conditional Index Number Analysis: The condition index numbers K_i of the matrix $\mathbf{X'X}$ is given in table - 1.

Table 1: Eigenvalue and Condition Index Numbers for Simulation Data.

Eigen Value	$\lambda_1 =$ 4.903	$\lambda_2 =$ 0.031	$\lambda_3 =$ 0.026	$\lambda_4 =$ 0.024	$\lambda_5 =$ 0.017
K_i	$K_1 =$ 1.0	$K_2 =$ 158.3	$K_3 =$ 189.3	$K_4 =$ 207.2	$K_5 =$ 290.8

The above table the maximum condition index number is 290.8. This shows that the regressors are strongly correlated. The Eigen system analysis also one of the indication of collinearity among the regressors.

(d) Variance-Decomposition Analysis: Using Belsey, etc., (1980) method the condition indices for the data set is given in table -2.

Table 2: The variance-decomposition proportions

Number	Eigenvalue	Condition Indices	X_1	X_2	X_3	X_4	X_5
1	4.90259	2.21418	0.00141	0.00107	0.00119	0.00113	0.00116
2	0.03098	0.17601	**0.59545**	0.05966	0.02798	0.07116	0.26569
3	0.02590	0.16094	0.03873	0.01072	**0.67422**	0.37130	0.01504
4	0.02366	0.15383	0.33143	**0.46041**	0.00013	0.13048	0.29576
5	0.01687	0.12988	0.032998	0.46815	0.29649	0.42594	0.42237

If any of the variance-decomposition proportion greater than 0.5 are recommended guidelines. In variance-decomposition proportionπ_{21} and π_{33} are greater than 0.5. This means that second and third singular values are associated with the collinearity that is inflating the variances of $\hat{\beta}_1$ and $\hat{\beta}_3$.

Ordinary Least Squares Results for Monte Carlo Simulation Data

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	79.002	15.800	634.60	**0.000**
Residual Error	94	2.340	0.025		
Total	99	81.343			

S = 0.157792 R-Sq = **97.1%** R-Sq(adj) = **97.0%**

Predictor	Coef	SE Coef	T	P
Constant	-0.0044	0.0166	-0.26	**0.794**
X1	0.2334	0.0943	2.48	0.015
X2	0.2107	0.1067	1.98	**0.051**
X3	0.1145	0.0998	1.15	**0.254**
X4	0.1411	0.1057	1.34	**0.185**
X5	0.2961	0.1055	2.81	0.006

From the above ANOVA table the estimated F value is 634.6 (P=0.000) and R^2 is 97%. According to these results it cannot be concluded that this OLS model is significant. But, if we consider the individualparameter result, the constant and variables X_2, X_3 and X_4 are not significant to the model at 5% level of significance. This problem was occurred because of the collinearity among the regressors.

Macroeconomic Impact of Foreign Direct Investment (MIFDI) Data

Based on Sun (1998, 2001) theory a Macroeconomic Impact of Foreign Direct Investment Data were collected in Sri Lanka from 1978 to 2004 to analyze the collinearity problems. This data set consists one dependent variable (Total Domestic Investment) and four independent variables (Foreign Direct Investment, Gross Domestic Product Per Capita, Exchange Rate and Interest Rate). This data set is time series, therefore the five variables should be analyzed to find whether all variables are following common trend with same order. For this purpose the *unit root test* is tested. At 1% level of significance all five variables are cointegrated to same order of integration coefficients 1.

The next test is to be tested that, the linear model of selected variables is the homoskedasticity (i.e., constant error variance). For this analysis the Durbin-Watson test is carried out. The estimated Durbin-Watson value for this data set is **2.0131**. The critical lower (d_l) and upper (d_u) values at 1% significance level is $\boldsymbol{d_l}$ **= 0.878** and $\boldsymbol{d_u}$ **= 1.515** (for sample size 27 and regressors 4), respectively. The estimated value for this model lies between $\boldsymbol{d_u}$ and **4-**$\boldsymbol{d_u}$, hence, it can be conformed that at the **1%** level of significance there is no autocorrelation.

Collinearity Results and Analysis for MIFDI Data

(a) **Correlation Analysis**: Correlation between the variables are 0.59016 -0.66621, -0.58280, -0.93693, -0.96745 and 0.90446. From these values it can be said that strong positive and negative correlation exists among regressors. Hence, it is confirmed that collinearity exists among the regressors.

(b) **VIF Analysis**: The diagonal elements of $(\mathbf{X'X})^{-1}$ is used to measure the VIF and the values are 1.83902, **23.7106, 9.78299, 15.7075** and it is given below.

From the above result three diagonal values are greater than 5. This is an indication of collinearity present among the regressors.

(c) **Conditional Index Analysis**: The eigenvalues and condition index numbers of **X'X** are given in table - 3.

Table 3: Eigenvalues and conditional index numbers for MIFDI Data

Eigen Value	$\lambda_1 = 5623.84$	$\lambda_2 = 21.39$	$\lambda_3 = 5.72$	$\lambda_4 = 0.18$
Condition Index	$K_1 = 1$	$K_2 =$ **262.92**	$K_3 =$ **983.19**	$K_4 =$ **31243.56**

From the above table it can be confirmed that strong collinearity exists among regressors.

From the above overall collinearity analysis for MIFDI data it can be said that strong collinearity among the independent variables.

Ordinary Least Squares Results for MIFDI Data

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	3.3779	0.8445	91.42	**0.000**
Residual Error	22	0.2032	0.0092		
Total	26	3.5811			

S = 0.0961092 R-Sq = **94.3%** R-Sq(adj) = **93.3%**

Predictor	Coef	SE Coef	T	P
Constant	4.179	4.3670	0.96	**0.349**
FDIR	0.0864	0.0278	3.11	0.005
GDPPCR	0.8886	0.3031	2.93	0.008
EXR	-0.6558	0.3000	-2.19	**0.040**
IRR	0.1501	0.1221	1.23	**0.232**

From the above ANOVA table the F value is 91.42 (P=0.000) and R^2 is 94.3%. Based on these results it cannot be confirmed that this OLS model is significant. However, let's consider the individual parameter results, the constant and variables EXR and IRR are not significant to the model at 5% level of significance. This occurred because of the collinearity among the regressors.

It is confirmed that the collinearity exists in the above simulation and MIFDI data sets. Hence, the alternative way to reduce or avoid the collinearity is to use biased estimator instead of using unbiased (OLSE) estimator. In this research paper the biased ORRE is suggested to use to fit the model and compare the stochastic properties of OLSE with ORRE. The stochastic properties are given in table – 4.

Table-4: OLSE and RRE Stochastic Properties in Simulation Data

Stochastic Property	**OLSE**	**ORRE**
R^2	0.971	0.971
σ^2	0.0249	0.00031
scalar mse(estimator)	0.0528	0.0099

Stochastic Properties Analysis for Simulation Data

As mentioned in section 4.3, the ridge regression linear model is used to fit regression model for the simulation data. The biasing parameter is estimated for this data set and k = 0.023. The stochastic properties such as coefficient of determination, mean squared error of the model and scalar mean squared error of the estimator of OLSE and ORRE for the simulation data are given below table - 4.

Considering the above stochastic properties R^2 is same for both estimators whereas σ^2 and scalar mse(estimator) for OLSE is higher than that of ORRE. Hence, the ORRE is better than the OLSE.

The 95% confidence interval and confidence width of OLSE and ORRE are given below table - 5.

Table - 5: 95% CI for OLSE and RRE in Simulation Data

OLSE			**ORRE**		
Parameter	95% CI	CI Width	Parameter	95% CI	CI Width
-0.00435	[-0.03695, 0.02825]	0.0652	-0.00436	[-0.00798, -0.00075]	0.00723
0.23343	[0.04862, 0.41824]	0.3696	0.23382	[0.21347, 0.25417]	0.04070
0.21074	[0.00169, 0.41979]	0.4810	0.20791	[0.18490, 0.23092]	0.04602
0.11446	[-0.08109, 0.31002]	0.3911	0.11639	[0.09482, 0.13797]	0.04315
0.14415	[-0.06605, 0.34834]	0.4144	0.14229	[0.12013, 0.16569]	0.04556
0.28610	[0.08922, 0.50297]	0.4238	0.29474	[0.27188, 0.31761]	0.04573

Considering the 95% CI and confidence width the OLSE has wider confidence interval and large width respectively than that of ORRE. Therefore, ORRE is better estimator compare with OLSE for this data set.

The next stochastic property analysis is the mean squared error matrix of the OLSE and ORRE. The difference between the matrix mean squared error of OLSE and ORRE was estimated and the result shows that non-negative definite matrix. This mean the estimator ORRE is better than OLSE for this collinearity data set.

Stochastic Properties Analysis for MIFDI Data

The stochastic properties R^2, σ^2 and scalar mse(estimator) of OLSE and ORRE for the MIFDI data are given below table - 6.

Table - 6: OLSE and RRE Stochastic Properties in MIFDI Data

Stochastic Property	OLSE	ORRE
R^2	0.943	0.941
σ^2	0.00924	0.0096
scalar mse(estimator)	19.2754	15.9734

In the above table the stochastic properties R^2 and σ^2 are approximately same for both estimators, whereas the property scalar mse(estimator) for OLSE is higher than that of ORRE in MIFDI data, this means ORRE better than the OLSE to fit the model for this collinearity data set.

The 95% CI and confidence width were also studied for OLSE and ORRE, the results show that the OLSE has wider confidence interval and confidence width than the ORRE. Hence, it can be said that ORRE is better estimator compare with OLSE to fit model for MIFDI data set.

Finally, matrix mean squared error of OLSE and ORRE are analyzed for MIFDI data set. The difference between the matrix mean squared error of OLSE and ORRE was obtained and the output shows that non-negative definite matrix.

From the above overall stochastic properties analysis for OLSE and ORRE in both collinerity data sets it is confirmed that ORRE is better than that of OLSE when collinearity present among the regressors. Therefore, when collinearity presents among the regressors the most suitable estimator to fit the model is biased estimators.

Comments

Two independent collinearity data sets were analyzed for the purpose of studying the problems of collinearity among the regressors. The diagnosing and dealing methods for collinearity problems discussed in this paper are very useful to detect collinearity in the real data analysis.

Although ordinary least squares estimator is best linear unbiased estimator (BLUE) in the class of best linear unbiased estimators it is not a suitable method to fit regression model when collinearity presents among the regressors. An appropriate and most suitable method to fit regression model for collinearity data is biased estimation. The popular and useful biased estimation is ridge type estimators.

References

Afifi, A.A. and Clark, V. (1996).*Computer aided multivariate analysis***.**3rd edition.Chapman and Hall.

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction.*Technometrics,* 58, 139-148.

Belsey, D.A., Kuh, E. and Welsch, R.E. (1980).*Regression Diagnostics: Identifying Influence data and Source of Collinearity.*Wiley New York.

Draper, N.R. and Smith, H. (1998).*Applied regression analysis.*3rd edition, John Wiley and Sons, Inc.

Hoerl, A.E., (1964). Ridge Analysis.*Chemical Engineering Process Symposium* Series60, 67-77.

Hoerl, A.E. and Kennard, R.W. (1970a). Ridge Regression: Biased Estimation for Nonorthogonal Problems.Technometrics, 12, 55-67.

Hoerl, A.E. and Kennard, R.W. (1970b). Ridge Regression: Application to Nonorthogonal Problems.Technometrics, 12, 69-82.

Marquardt, D.W. (1970). Generalized Inverse, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation. *Technometrics,* 12, 591-612.

McDonald, G.C. and Galarneau, D.I. (1975).A Monte Carlo evaluation of some ridge-type estimators. *The American Statistical Association,* 70(**350**), 407-416.

Montgomery, D.C. and Peck, E.A. (1992).*Introduction to linear regression analysis.*John wiley and Sons, Inc.

Quirino, P. (2001). Multicollinearity and maximum entropy estimators.*Economics Bulletin,* 3(**11**), pp. 1-9.

Sun, H. (1988).*Macroeconomic Impact of Direct Foreign Investment in China: 1979-1996. UK*: Blackwell Publishers Ltd.

Sun, H. (2001).Foreign Direct Investment and Export Performance in China.*Journal of Regional Science*, 41(**2**), 317-336.