*Electronic Journal of Applied Statistical Analysis*
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 2, 304 – 311
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n2p304


# THE EFFECTS OF PERTURBING EIGENVALUES IN THE PRESENCE OF MULTICOLLINEARITY

**Ethelbert Chinaka Nduka*, Maxwell Azubuike Ijomah**

*Department of of Mathematics & Statistics, University of port Harcourt, Nigeria*

Received 13 October 2010; Accepted 25 August 2011
Available online 14 October 2012

***Abstract***. *Multicollinearity is a linear dependency between two or more explanatory variables in the regression models which can seriously distort the least squares estimates. The variance of the Ordinary Least Squares Estimates would be inflated and the regression coefficients often indetermined in the presence of multicollinearity. Therefore, biased estimators are suggested as alternatives to the Ordinary Least Squares Estimator. In this study, a new method of solving multicollinearity problem through perturbation of eigenvalues is proposed. The performance of this estimator is evaluated by comparing it with some existing estimators (i.e. Ridge regression and the Principal Component regression) in terms of mean squared error. The comparisons show that these new method appears to be superior to other estimators in terms of a reduction in the mean squared error when certain conditions (loss of information in the number of variables) are satisfied. Hence, the accuracy of the parameter estimate increases.*

***Keywords***: *Multicollinearity, perturbation, eigenvalues, ordinary least squares, principal component, ridge regresssion.*

## 1. Introduction

Multiple regression models are widely used in applied statistical techniques to quantify the relationship between a response variable $Y$ and multiple predictor variables $X_i$, ($i$ = 1, 2, ...,$p$) where $p$ is the number of regressors, and we utilize the relationship to predict the value of the response variable from a known level of predictor variables, and to the statistical hypotheses.

* E-mail: etelnduka@yahoo.com

In estimating the parameters of the linear regression model using the Ordinary least squares method, one assumption made is that the independent variables $X_i$ are not linearly (perfectly) correlated. In other words, the design matrix **X** is of full rank ( e.g. rank (**X**) = $x$) . However, this is not usually so in practice – Hence, there is some degree of intercorrelation among the explanatory variables due largely to the interdependence of much economic magnitude over time. When this assumption is violated, and there is intercorrelation in the explanatory variables, we say there exist collinearly or multicollinearty between the explanatory variables [1]. The problem of multicollinearity is being able to separate the effects of two (or more) variables on an outcome variable. If two variables are significantly related, it becomes impossible to determine which of the variables accounts for variance in the dependent variable. As a rule of thumb, the problem occurs when explanatory variables are more highly correlated with each other than they are with the dependent variable. Mathematically, the problem is that the **X** matrix is not full rank. When this occurs, the **X** matrix (and hence the **X'X** matrix) has determinant zero and can not be inverted. High interpredictor correlations will lead to less stable estimate of regression weights.
When using multiple regression, occasionally we experience an apparent contradiction of intuition or theory when one or more of the regression coefficients seem to have the wrong sign. This makes a serious interpretative problem, as it is really difficult to explain a negative estimate (say) of a parameter to the model user when that user believes that the coefficient should be positive [6]. The relationship can be problematic if the parameter estimates obscure some functional relationship of interest to a researcher. It becomes very difficult to identify the separate effect of the variables involved precisely. Collinearity therefore makes it more difficult to achieve significance of the collinear parameters. In fact, since the regression coefficients are interpreted as the effect of change in their corresponding variables, all other things held constant, our ability to interpret the coefficients declines the more persistent and severe the collinearly. The lack of precision is manifested by the existence of potentially large sampling variance for estimators of the unknown parameters and high correlation between affected estimators.
In this paper, we propose a new method to handle the problem of multicollinearity via perturbation of eigenvalues.

## 2. Some Solutions to Multicollinearity

One suggestion that has been frequently made in trying to overcome the problem of multicollinearity is to collect new data [8]. Sometimes, the problem of multicollinearity occurs due to inadequate or erroneous data. Unfortunately, this is not always possible since some analysis must be based on the available data. Furthermore, this solution is not possible when the presence of multicollinearity is the result of internal constraints of the system being studied [7].
Dropping of variables that duplicate the role of other 'more' important variables has also been proposed as a natural solution to the multicollinearity problem. However, for models that strictly adhere to some theoretical framework, this is equivalent to massive loss of information.
Two other most popular methods are; the Ridge Regression and the Principal Component Analysis. Unfortunately, the ridge estimator is not generally accepted as a vital alternative to the Ordinary Least Squares (OLS) estimator because it depends on unknown parameters. In any specific application of ridge regression, there is no guarantee that the sample estimate is a

member of the class of more accurate estimates. The efficiency of ridge estimation therefore depends upon the estimate of prior information the researcher has regarding the population model.

The principal component regression is also proposed, hoping that the linear combination of the *X*'s as a regressor will be able to keep all the variables into the model. Depending on the structure of the relationship among the regressors, component loadings may affect parameter estimates, e.g., loadings are similar, resulting in regression coefficients that are similar for all regressors. This will cause problem in interpretation because the relative importance of predictors is being masked by the way the linear combination is formed. Furthermore, the problem of choosing an optimum subset of predictors in a principal component analysis persists. A possible strategy is to keep only a few of the first components, but the components are chosen to explain the explanatory variables rather than the dependent variable, and so nothing guarantees that the principal components which 'explain' the regressors are relevant for the dependent variable. Also, the method uses less of information contained in the sample since the number of components retained is smaller than the number of the regressors. For this reason, the principal component regression is considered inappropriate as a solution to multicollinearity [5].

The literature on linear models with special focus on multicollinearity spans several decades already. However, no optimal solution has been achieved so far, so that there is still a continuing active interest on the problem.

## 3. Proposed Solution to Multicollinearity Problem

We will follow a line similar to [3] via the eigenvalue perturbation to identify points of influential eigenvalues to individual estimators of interest $\eta$, say. Since it has been found from literatures that multicolinearity exists whenever at least one of the eigenvalues of **X'X** is small [4]. To study how a perturbation of eigenvalue can have influence on model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{\varepsilon} \sim N(0, \delta^2\mathbf{I})$ we write the perturbed distribution of $\boldsymbol{\varepsilon}$ in the form of $\boldsymbol{\varepsilon} \sim N(0, \delta^2\mathbf{A}^{-1})$, where **A** is a diagonal matrix with diagonal elements $S_i = 1 + al_i$ (for $i$=1,2,…..,$n$) with $\mathbf{l} = (l_1, \ldots\ldots\ldots\ldots, l_n)$ and **a** are the direction and the magnitude of perturbation respectively.

Let $\mathbf{S} = (S_1, \ldots\ldots\ldots S_n)'$, the perturbation caused by the variation of **a** through $S = 1 + al$ spans a surface of $\eta(S(a))$ in an $(\eta + 1)$ dimensional space.

Now from the ridge regression estimation proposed by Hoerl and Kennard (1970). The ridge regression estimation is really a family of estimators given by:

$$\hat{\boldsymbol{\beta}}(\kappa) = (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})\mathbf{X}'\mathbf{Y} \qquad (2)$$

Where $\kappa > 0$ is a constant, often called the shrinkage or biasing parameter. In our case, we extend the above estimator by introducing the eigenvalues thereby having $\hat{\boldsymbol{\beta}}(\tau) = [\mathbf{p}^1(\boldsymbol{\lambda} + \tau\mathbf{I})^{-1}\mathbf{p}\ ]\mathbf{X}'\mathbf{Y}$ where $\lambda$= diagonal eigenvalues and **p**= corresponding eigenvectors.

Motivated by the relationship between the singular values ($\delta_i's$) of the design matrix **X** and eigenvalues of **X'X** in the above equation,, which the cross-product matrix **X'X** is a component of the variance-covariance matrix (i.e. $\delta^2(\mathbf{X}'\mathbf{X})^{-1}$ ) of $\boldsymbol{\beta}$, it is easy to see the role that eigenvalues play in collinearity.

From the least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ of $\boldsymbol{\beta}$, and introducing the eigenvalue decomposition of **X'X** we have that:

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{p}\boldsymbol{\lambda}^{-1}\mathbf{p}' = \sum \boldsymbol{\lambda}_i^{-1}\mathbf{p}_i\mathbf{p}_i' \tag{3}$$

and so with perturbation of the eigenvalues, we can write:

$$\hat{\boldsymbol{\beta}}(\tau) = \sum (\lambda_i + \tau\mathbf{I})^{-1}\mathbf{p}'\mathbf{p}\mathbf{X}'\mathbf{Y} = \sum (\lambda_i + \tau\mathbf{I})^{-1}\mathbf{c}_i\mathbf{p}_i' \tag{4}$$

where $\mathbf{c}_i = \mathbf{p}_i\mathbf{X}'\mathbf{Y}$. $\mathbf{p}_i$ is an eigenvector. If we now consider the variance–covariance matrix of $\boldsymbol{\beta}$, we have:

$$V(\hat{\boldsymbol{\beta}}) = \delta^2\mathbf{p}(\lambda + \tau\mathbf{I})^{-1}\mathbf{p}' = \delta^2 \sum (\lambda_i + \tau\mathbf{I})^{-1}\mathbf{p}_i\mathbf{p}_i' \tag{5}$$

So we can rewrite the variance as:

$$V(\hat{\boldsymbol{\beta}}) = \delta^2[p_1, p_2, \ldots\ldots\ldots, p_k]^{-1} \begin{vmatrix} \lambda + \tau I & 0 & 0 \\ \ldots\ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & \Omega & \lambda_k + \tau I \end{vmatrix} \begin{bmatrix} p_1^{\,1} \\ M \\ p_k^{\,1} \end{bmatrix}$$

$$V(\hat{\boldsymbol{\beta}}) = \delta^2[p_1, p_2, \ldots\ldots\ldots, p_k] \begin{bmatrix} \frac{p^1}{\lambda_1 + \tau I} \\ \Omega \\ \frac{P_K}{\lambda_K + \tau I} \end{bmatrix}$$

$$Var(\hat{\boldsymbol{\beta}}) = \delta^2 \left| \frac{p_1 p_1^{\,1}}{\lambda_1 + \tau I} + \ldots\ldots\ldots\ldots + \frac{P_k p_k^{\,1}}{\lambda_k + \tau I} \right|$$

Thus perturbing the eigenvalues of the original predictor variable cross-products matrix to "push" them away from zero, thereby reducing the amount of collinearity of the variables.

In order to generate multivariate normal vectors to be used, the following procedure was adopted:

1. For fixed sample size *n* (*n* = 30, 100 and 200) we generate 5 replicates, each of size *n*, from the regression: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_5 x_5 + \varepsilon$.
2. Initial (Guess) values were assigned to $\beta$'s. These are $\beta_0 = 1$, $\beta_1 = 2, \beta_2 = 3, \beta_3 = 4, \beta_4 = 5$ and $\beta_5 = 6$.
3. We then calculated the *p* x *p* correlation matrix for this simulated data and extract the p eigenvalues and order them from largest (position 1) to smallest (position p).
4. The number of variables in the multi-normal distributions to be generated is to be predetermined. The SAS V9 package will be used for this analysis.

We shall use these steps in performing the Monte Carlo experiment using statistical software package (SAS) to access the performance of the eigenvalue in the presence of multicollinearity at different sample sizes.

## 4. Simulated Data and Result

In order to evaluate the performance of eigenvalues in identifying the true model, when we have multi-collinear predictors, the following simulation study was conducted : we have sample sizes *n*= 30, 100 and 200, with fixed model size (*m*= number of explanatory variables in the regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ to five variables). The (true) coefficient vectors $(\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4, \beta_4 = 5, \beta_5 = 6)$ and the disturbance vector $\varepsilon$, is normally distributed with $\mu = 0$ and $\sigma(\varepsilon) = 1$.

Multicollinear **X** (30, 100, 200 : 5) matrices yielding **X'X** with different condition index have been generated by the procedure laid down as stated above. We shall test with multicollinear correlation coefficients of 0.999, 0.55 and 0.19. This will enable us to have an in-depth knowledge of the study. The author wrote his program (in SAS) for computation needed in this work.

Tables 1 to 3 present the average of the estimates at different sample sizes *n*= 30, 100 and 200. In order to capture the true behavior of the eigenvalues in the presence of multicollinearity, the variables were tested at different levels of collinearity using multicollinear correlation coefficient (m.c.e). As stated above, we classify as follows: perfect collinearity= 0.999, medium= 0.55 and low= 0.19.

Table 1 summarizes the parameter estimates from different methods as well as the root mean squared error (RMSE) assessing the predictive ability of the model. For principal component regression, only two components were used in the model since it already accounts for 80% of the total variation in the five independent variables.

**Table 1. Performance evaluation of estimators using average of parameter estimates for n = 30.**

| | | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | RMSE |
|---|---|---|---|---|---|---|---|
| Simulated Parameters | High | 2 | 3 | 4 | 5 | 6 | |
| | Medium | 2 | 3 | 4 | 5 | 6 | |
| | Low | 2 | 3 | 4 | 5 | 6 | |
| OLS | Cor coeff =0.99 | -3.79 | -4.31 | 4.06 | 1.08 | 3.35 | 0.97 |
| | Cor.coeff.= 0.55 | -0.07 | -0.48 | 0.45 | 0.12 | 0.37 | 0.97 |
| | Cor.coeff.= 0.19 | 0.16 | -0.24 | 0.23 | 0.06 | 0.19 | 0.97 |
| Ridge Estimator | High | -0.06 | -0.67 | 0.82 | -0.17 | 0.50 | 1.15 |
| | Medium | -0.01 | -0.41 | 0.40 | 0.08 | 0.33 | 0.98 |
| | Low | 0.16 | -0.23 | 0.22 | 0.06 | 0.18 | 0.97 |
| Perturbed Eigenvalue Estimator | High | -0.09 | -0.79 | 0.94 | -0.20 | 0.56 | 1.14 |
| | Medium | -0.01 | -0.42 | 0.41 | 0.08 | 0.34 | 0.98 |
| | Low | 0.16 | -0.23 | 0.22 | 0.06 | 0.18 | 0.97 |
| Principal Component Regression | High | 0.17 | 2.97 | 3.55 | 3.45 | 3.82 | 0.97 |
| | Medium | 0.19 | -0.01 | 0.99 | -0.66 | 0.47 | 0.97 |
| | Low | 0.20 | -0.01 | 0.69 | -0.37 | 0.14 | 0.97 |

**Table 2. Performance evaluation of estimators using average of parameter estimates for n = 100.**

| | | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | RMSE |
|---|---|---|---|---|---|---|---|
| Simulated Parameters | High | 2 | 3 | 4 | 5 | 6 | |
| | Medium | 2 | 3 | 4 | 5 | 6 | |
| | Low | 2 | 3 | 4 | 5 | 6 | |
| OLS | Cor coeff =0.99 | -2.65 | -0.49 | 2.19 | 0.46 | 0.62 | 1.02 |
| | Cor.coeff.= 0.55 | -0.19 | -0.05 | 0.24 | 0.05 | 0.07 | 1.02 |
| | Cor.coeff.= 0.19 | -0.03 | -0.03 | 0.12 | 0.03 | 0.03 | 1.02 |
| Ridge Estimator | High | -0.07 | -0.13 | 0.28 | 0.01 | 0.05 | 1.04 |
| | Medium | -0.12 | -0.05 | 0.21 | 0.04 | 0.06 | 1.02 |
| | Low | -0.03 | -0.03 | 0.12 | 0.03 | 0.03 | 1.02 |
| Perturbed Eigenvalue Estimator | High | -0.09 | -0.15 | 0.32 | 0.01 | 0.05 | 1.03 |
| | Medium | -0.12 | -0.06 | 0.21 | 0.04 | 0.06 | 1.02 |
| | Low | -0.03 | -0.03 | 0.12 | 0.03 | 0.03 | 1.02 |
| Principal Component Regression | High | 0.05 | 0.29 | 1.83 | -0.43 | 2.69 | 1.02 |
| | Medium | 0.07 | 0.07 | -0.32 | -0.17 | 0.39 | 1.02 |
| | Low | 0.12 | 0.07 | -0.11 | -0.17 | 0.11 | 1.02 |

**Table 3. Performance evaluation of estimators using average of parameter estimates for n = 200.**

| | | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | RMSE |
|---|---|---|---|---|---|---|---|
| Simulated Parameters | High | 2 | 3 | 4 | 5 | 6 | |
| | Medium | 2 | 3 | 4 | 5 | 6 | |
| | Low | 2 | 3 | 4 | 5 | 6 | |
| OLS | Cor coeff =0.99 | -2.33 | -0.28 | 1.36 | 0.75 | 0.60 | 1.05 |
| | Cor.coeff.= 0.55 | -0.18 | -0.03 | 0.15 | 0.08 | 0.07 | 1.05 |
| | Cor.coeff.= 0.19 | -0.04 | -0.02 | 0.08 | 0.04 | 0.03 | 1.05 |
| Ridge Estimator | High | -0.05 | -0.07 | 0.14 | 0.06 | 0.03 | 1.07 |
| | Medium | -0.11 | -0.03 | 0.12 | 0.07 | 0.05 | 1.05 |
| | Low | -0.04 | -0.02 | 0.07 | 0.04 | 0.03 | 1.05 |
| Perturbed Eigenvalue Estimator | High | -0.07 | -0.09 | 0.16 | 0.06 | 0.04 | 1.06 |
| | Medium | -0.12 | -0.03 | 0.13 | 0.07 | 0.05 | 1.04 |
| | Low | -0.04 | -0.02 | 0.07 | 0.04 | 0.03 | 1.05 |
| Principal Component Regression | High | 0.04 | 0.06 | 0.36 | 1.12 | 2.56 | 1.06 |
| | Medium | 0.06 | 0.02 | -0.50 | 0.22 | 0.38 | 1.05 |
| | Low | 0.09 | 0.01 | -0.03 | 0.13 | 0.11 | 1.05 |

A critical look at the above tables reveal that the new method performed better than the ridge regression method using the RMSE which is specifically what we set out to minimize; so we present the results demonstrating that the RMSE is reduced through the use of this new method especially under severe collinearity. Among the three methods studied, the new method most consistently performed best for all sample sizes (i.e. 30, 100 and 200) especially under severe collinearity.

## 5. Conclusion

Perturbation model can produce superior prediction or at least comparable predictive ability to the usual solution to the problem of multicollinearity like Principal Component Regression and Ridge Estimation

This new model estimated through perturbation of eigenvalues can accurately estimate the more important variables estimated early in the iterative process. The comparisons show that these new method appears to be superior to other estimators in terms of a reduction in the mean squared error when certain conditions are satisfied. Hence, the accuracy of the parameter estimate increases. Indirectly, the impact of multicollinearity is reduced. Therefore, this new method via perturbation of eigenvalues can be considered in the linear regression model in order to obtain a better regression equation. An important strength of this new method also, is that it indicates clearly in which directions solutions should be sought, unlike the Ridge Regression which does not provide insight about the subspace of that explains the response well, as pointed out by Frank and Fredman, (1993). Furthermore, the new method also exhibited some advantages over the principal component regression which uses less of information contained in the sample since the number of components retained is smaller than the number of the regressors and as such is considered inappropriate as a solution to multicollinearity.
(Koutsoyannis, 1977).

Regardless of model fit, the capability of the perturbation model in estimating the contributions of the more important variables holds.

## References

[1]. Drapper N.R. and Smith H. (1981). *Applied Regression Analysis*. New York: John Wiley and Sons, Inc.
[2]. Frank, I.E. and Friedman, J.H. (1993). A Statistical view of some chemometrics regression tools. *Technometrics*, 35, 109 – 148.
[3]. Hoerl, A.E. and Kennard, R.W.(1970), Ridge Regression: Biased Estimation for Non-Orthogonal Problems. *Technometrics*, 2, 109 – 120.
[4]. Judge, G.G, Hill, R.C., Griffiths, W.E., Lee, T.C. (1980). *Introduction to the Econometrics*. New York: John Wiley.
[5]. Kuotsoyiannis, A. (1977). *Theory of Econometrics, Second Edition*, London: Macmillan.
[6]. Mullet, G. M. (1976). Why regression coefficients have the wrong sign? *J. Qual. Technol.*, 8, 121-126.
[7]. Rawlings, J. O., Pantula, S.G, Dickey, D. A. (1998). *Applied Regression Analysis: A research tool. 2nd ed*. New York: Springer-Verlag.
[8]. Ryan T.P., (1997). *Modern Regression Methods*. New York: John Wiley and Sons.
[9]. Freund, R.J. and R.C. Littell (1991). *SAS System for Regression, 2nd ed.* SAS Institute, Cary, NC.
[10]. Wetherill, G.B., Duncombe, .P., Kenward M., Paul S.R., Vowden B.J., (1986). *Regression Analysis with Application*. New York: Chapman and Hall Ltd.