A Modified Least-Squares Approach to Mitigate the Effect of Collinearity in Two-Variable Regression Models

Martin L. William *and* L.Maria Alphonse Ligori
Department of Statistics, Loyola College, Chennai, India

Abstract

This paper presents a modification to ordinary least squares (OLS) method with a view to overcoming the ill-effects of collinearity on the OLS estimates of the regression parameters in a linear model with two explanatory variables. This modified approach leads to estimates that are, to a large extent, better than OLS estimates under the mean square error criterion and also overcome the overestimation problem that plagues the OLS estimates. Although a few attempts to get improved estimates have been made by some authors, the method developed here takes a route that has not been hitherto ventured in the context of addressing collinearity issues.

Keywords and Phrases: *Collinearity, Ordinary Least Squares, Relative Efficiency*

1. INTRODUCTION

'Collinearity' among the regressors of a linear model is among the most endemic concerns raised not only by theoreticians but also by practitioners involved in modeling real-life data. The severity of the problem of multicollinearity in linear models may be gauged by the fact that more than 200 articles discussing this problem have appeared during the past few decades. Farrar and Glauber (1967), Stewart (1987), Mason and Perreault Jr.(1991) and Fox and Monette (1992), to mention a few, discussed issues related to multicollinearity problem. One simple 'solution' to multicollinearity problem is removal of the variable(s) affected by collinearity after carrying out a preliminary diagnosis. However, removing a variable is not always a prudent action as it results in completely losing information on the effects of that regressor on the response variable. A number of diagnostic methods are followed for identifying the problematic variable. Reference is made to Marquardt (1970), Willan and Watts (1978), Belsley *et al* (1980) for multicollinearity diagnostics. Further, there has been a plethora of recent articles [by Greene (1999), Hansen (1999), Dutta and Ahmed (1999) and others] that have dealt with economic phenomena and have discussed various ways of dealing with the collinearity problem. A review of these works reveals that the issue of collinearity is widely prevalent and there is a lack of consensus on ways of handling it.

As is well known, in the presence of multicollinearity, OLS is likely to yield 'poor' estimates of the regression parameters. The estimates are of incorrect or counter-intuitive signs and/or are of implausible magnitudes. Improving the estimates in the presence of collinearity among the regressors without losing out any regressor is a problem worth addressing because available procedures to overcome the weakness in OLS estimates have not found wide acceptance in practice as is found from the articles mentioned above. Alternative solutions that have been suggested in the past include using Bayesian estimation (Zellener (1971), Leamer (1973), Leamer (1978)) for some parameters and Ridge Regression (Hoerl and Kennard, 1970a, Hoerl and Kennard, 1970b, Hoerl and Kennard, 1976)). The drawback in all these procedures is that they are based on many assumptions that are not always practically viable in real-time analysis.

There have been attempts in the past to get improved estimates for regression parameters under mean square error criterion, by relaxing the requirement that the estimates be unbiased. Vizcarrondo and Wallace (1968) investigated this issue in the context of restrictions among the regression parameters which arise in testing linear hypothesis. According to these authors,

multicollinearity is a 'situation in which there exists at least one linear restriction that would yield estimates better than OLS estimates according to the mean square error criterion'. Sclove (1968) also addressed the problem of improving the estimates in linear regression, again under the mean square error criterion, by incorporating preliminary tests of significance of the regression coefficients. In short, both the afore-mentioned works involve imposing restrictions on the parameters, testing the restrictions and then getting 'improved' estimates for the remaining parameters.

The more interesting route would be to get improved estimates initially by retaining all parameters (equivalently, all regressors) that one originally starts with and then perform tests on the parameters. But, this route does not seem to have been pursued so far. The present paper is an attempt in this direction.

This paper is organized as follows: In Section 2, we propose a modification to the OLS method and derive the 'modified' estimates for the regression parameters. Section 3 presents the comparison of the modified estimates with OLS estimates under the 'mean square error' criterion. Section 4 also discusses another comparison of the two estimates vis-à-vis the problem of overestimation. Section 5 contains concluding remarks and indicates some of the ongoing investigations that are currently being pursued.

2. THE PROPOSED MODIFICATION TO OLS METHOD AND THE NEW ESTIMATE

Consider a linear model with two regressors given by

$$Y_i = a_1 X_{1i} + a_2 X_{2i} + e_i,\ i = 1, 2, \ldots, n \qquad e_i \sim N(0, \sigma^2) \qquad (2.1)$$

Without loss of generality we assume that X_1 and X_2 are normalized so that $\Sigma X_{1i}^2 = \Sigma X_{2i}^2 = 1$ and hence there is no intercept in (2.1).

Denoting $\mathbf{Y} = (Y_1, .., Y_n)^T$ and $\mathbf{X} = \begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ \vdots & \vdots \\ X_{1n} & X_{2n} \end{bmatrix}$,

the OLS estimates for the regression parameters are given by

$$\hat{a}_{OLS} = (\hat{a}_{1OLS}, \hat{a}_{2OLS})^T = (\mathbf{X^T X})^{-1}(\mathbf{X^T Y})$$

These estimates are unbiased for a_1 and a_2 but their variances are large when the correlation between X_1 and X_2 is high. In the sequel, we suggest a modification to OLS that leads to improved estimates. The procedure is as follows:

Step 1: Obtain the initial OLS estimates of a_1 and of a_2 by regressing Y on X_1 and on X_2 separately. The estimates are

$$\hat{a}_1(0) = \Sigma Y X_1 / \Sigma X_1^2 \quad \text{and} \quad \hat{a}_2(0) = \Sigma Y X_2 / \Sigma X_2^2.$$

Step 2: Considering the model $Y_i - \hat{a}_2(0) X_{2i} = a_1 X_{1i} + e_i$, $i = 1, 2, \ldots, n$, obtain a revised OLS estimate of a_1. Similarly, obtain the revised estimate for a_2 by regressing $Y - \hat{a}_1(0) X_1$ on X_2. These estimates are given by

$$\hat{a}_1(1) = \Sigma(Y - \hat{a}_2(0) X_2) X_1 / \Sigma X_1^2 = \hat{a}_{1OLS}(1 - w) \qquad (2.2)$$

$$\hat{a}_2(1) = \Sigma(Y - \hat{a}_1(0) X_1) X_2 / \Sigma X_2^2 = \hat{a}_{2OLS}(1 - w) \qquad (2.3)$$

where $w = r_{12}^2$, r_{12} being the observed coefficient of correlation between X_1 and X_2.

We shall refer to these estimates as ***Modified OLS*** estimates and note that these are biased estimates since $E(\hat{a}_1(1)) = a_1(1 - w)$ and $E(\hat{a}_2(1)) = a_2(1 - w)$. We also note that in the case of perfect orthogonality or complete absence of collinearity (i.e. $r_{12} = 0$), the Modified OLS estimates are identical to the OLS estimates $\hat{a}_{1OLS}$ and $\hat{a}_{2OLS}$ and hence unbiased.

Here, we recall a suggestion of Tukey (1960) who proposed multiplying the usual OLS estimate by a constant between zero and unity to get a variance smaller than that of the OLS estimate. Tukey made this proposal specifically for the quadratic term in a polynomial regression model with linear and quadratic terms. The modified OLS estimates obtained above formalizes the proposal of Tukey and at the same time its applicability is not restricted to polynomial regression models.

3. COMPARISON OF THE MODEIFIED ESTMATES WITH OLS ESTIMATES

As the modified estimates are biased, the mean square error criterion is employed to study their performances and compare with the OLS estimates. With some computations, it is found that

$MSE(\hat{a}_1(1)) = \sigma^2 (1 - w) + a_1^2 w^2$ and $MSE(\hat{a}_2(1)) = \sigma^2 (1 - w) + a_2^2 w^2$.

We note that when $r_{12} = 0$, the mean square errors of the modified estimates are equal to that of the usual OLS estimates. Henceforth, we restrict ourselves to situations where $r_{12} \neq 0$.

We now carry out a comparison of $\hat{a}_1(1)$ with $\hat{a}_{1OLS}$

Consider D = MSE ($\hat{a}_{1OLS}$) – MSE ($\hat{a}_1(1)$)

$$= \frac{w}{1-w}[w^2 a_1^2 - w(\sigma^2 + a_1^2) + 2\sigma^2]$$

$$> \frac{w}{1-w}[w^2 a_1^2 - w(2\sigma^2 + a_1^2) + 2\sigma^2]$$

$$= \frac{w}{1-w}[(w a_1^2 - 2\sigma^2)(w-1)]$$

Clearly, D > 0 if $w < 2\sigma^2 / a_1^2$

Case (i): $2\sigma^2 / a_1^2 > 1$

In this case, D > 0 for all $w \leq 1$, which means that, the modified estimate $\hat{a}_1(1)$ is preferable to $\hat{a}_{1OLS}$ irrespective of the value of *w*.

Case (ii): $2\sigma^2 / a_1^2 < 1$

In this case, D > 0 for $w \leq 2\sigma^2 / a_1^2$. Also with some computations, we have

$$D > a_1^2\left[\frac{\sigma^2}{a_1^2}\frac{1}{(1-w)} - 1\right]$$

From this we get D > 0, if $w \geq 1 - \left[\frac{\sigma^2}{a_1^2}\right]$. Thus the modified estimate $\hat{a}_1(1)$ is preferable to $\hat{a}_{1OLS}$ for all 'w' outside the interval $\left(\frac{2\sigma^2}{a_1^2}, 1 - \frac{\sigma^2}{a_1^2}\right)$.

If $\frac{\sigma^2}{a_1^2} \geq \frac{1}{3}$, the interval mentioned above does not exist and hence the modified estimate outperforms the OLS estimate for every value of w.

If $\frac{\sigma^2}{a_1^2} < \frac{1}{3}$, then the comparison of the performances of $\hat{a}_1(1)$ and $\hat{a}_{1OLS}$ is required for $w \in \left(\frac{2\sigma^2}{a_1^2}, 1 - \frac{\sigma^2}{a_1^2}\right)$.

The table below presents the efficiency of $\hat{a}_1(1)$ relative to $\hat{a}_{1OLS}$ for various choices of $\frac{\sigma^2}{a_1^2}$ corresponding to different choices of 'w' over $\left(\frac{2\sigma^2}{a_1^2}, 1 - \frac{\sigma^2}{a_1^2}\right)$. The relative efficiency (RE) is computed as $RE = \frac{MSE(\hat{a}_{1OLS})}{MSE(\hat{a}_1(1))}$.

Table 1 [Efficiency of $\hat{a}_1(1)$ relative to $\hat{a}_{1OLS}$ for various choices of σ^2/a_1^2 and w]

w \ σ^2/a_1^2	0.05	0.1	0.15	0.2	0.25	0.3
0.1	1.010101	1.111111	1.149425	1.169591	1.182033	1.190476
0.15	0.904977	1.094391	1.176471	1.222307	1.251564	1.27186
0.2	0.78125	1.041667	1.171875	1.25	1.302083	1.339286
0.25	0.666667	0.969697	1.142857	1.254902	1.333333	1.391304
0.3	0.571429	0.892857	1.098901	1.242236	1.347709	1.428571
0.35	0.496278	0.820513	1.048951	1.218583	1.349528	1.453664
0.4	0.438596	0.757576	1	1.190476	1.344086	1.470588
0.45	0.395257	0.70609	0.956938	1.163636	1.336898	1.48423
0.5	0.363636	0.666667	0.923077	1.142857	1.333333	1.5
0.55	0.34188	0.639488	0.900901	1.132343	1.338688	1.52381
0.6	0.328947	0.625	0.892857	1.136364	1.358696	1.5625
0.65	0.324675	0.624512	0.902256	1.160261	1.40056	1.624915
0.7	0.330033	0.641026	0.934579	1.212121	1.474926	1.724138
0.75	0.347826	0.680851	1	1.306122	1.6	1.882353
0.8	0.384615	0.757576	1.119403	1.470588	1.811594	2.142857
0.85	0.456621	0.903955	1.342282	1.771872	2.192982	2.605863
0.9	0.613497	1.219512	1.818182	2.409639	2.994012	3.571429

From the table, it is found that 'by and large', the modified estimate $\hat{a}_1(1)$ is preferable to $\hat{a}_{1OLS}$ for $\frac{\sigma^2}{a_1^2} \geq 0.20$. A special highlight is the 'higer relative efficiency' of the modified estimates for the highly troublesome values of w (near 0.9 in which case 'collinearity' becomes a serious issue). Even for the ratio close to 0.1, we observe that for the highly 'troublesome' value of w, the modified estimate overtakes the ordinary OLS estimate. For still smaller values of σ^2 (relative to a_1^2) , the above computations reveal that OLS itself can be preferred over the modified approach proposed in Section2. Similar comparison can be made for $\hat{a}_2(1)$.

4. AN INTERESTING PROPERTY OF THE MODIFIED ESTIMATE

We know that the angle between two vectors x and y is given by

$$\cos\theta = < x, y > = \frac{x^T y}{\sqrt{x^T x}\sqrt{y^T y}}$$

Consider the angle between

$$< E(\hat{a}(1), a > = \frac{E(\hat{a}(1)^T a}{\sqrt{[E(\hat{a}(1))]^T E(\hat{a}(1))}\sqrt{a^T a}}$$

Since $E(\hat{a}(1)) = a(1-w)$, we get

$$\cos\theta = \frac{(1-w)a^T a}{(1-w)\sqrt{a^T a}\sqrt{a^T a}} = 1$$

$$\Rightarrow \theta = 0$$

Thus the 'revised' estimates, on an average, lies in the direction of the regression arameter(s).This property is also satisfied by the OLS estimate(s).

It is well known that in the presence of collinearity, the OLS tends to overestimate the regression parameters. In the sequel, we establish that the revised estimate(s) does not suffer from overestimation problem to the extent of the OLS estimate(s).

Consider the expected length of the OLS estimate (Montgomery *et al.,* 2003) namely

$$E(\hat{a}_{ols}^T \hat{a}_{ols}) = a^T a + \frac{2\sigma^2}{(1-w)}$$

We observe that, on the average, the length of the OLS estimate becomes extremely large for large values of *w*. Thus, when there is severe collinearity, the OLS estimates tends to 'exceedingly' overestimate the regression parameters.

In contrast, the expected length of the 'revised' estimate is given by

$$E(\hat{a}(1)^T \hat{a}(1)) = a^T a + 2\sigma^2(1-w)^2 - \sigma^2 w(2-w)\left(\frac{a_1^2}{\sigma^2} + \frac{a_2^2}{\sigma^2}\right)$$

$$< a^T a + 2\sigma^2(1-w)^2 < a^T a + \frac{2\sigma^2}{(1-w)} = E(\hat{a}_{ols}^T \hat{a}_{ols})$$

Thus the expected length of the revised estimate(s) is less than the expected length of the OLS estimate(s) and the severity of overestimation is reduced.

5. CONCLUDING REMARKS

The 'modified' approach discussed in this paper, overcomes the drawbacks of OLS estimates in the presence of collinearity to a large extent. However, the modified estimates do not perform uniformly well over OLS estimates as found in Section 3. The comparison of the two approaches depends on the ratios σ^2/ a_1^2 and σ^2/ a_2^2. Hence, it is pertinent to suggest a way of initially estimating these ratios and choose the approach based on these initial estimates. Investigations in this regard are presently in a preliminary stage and the issue will be addressed in a future communication. The generalization of the modified approach for multiple regression models involving more than two regressors will also be addressed in a future communication.

REFERENCES

1] Belsley,D.A., Kuh,E. and Welsh,R.E. 1980. Regression Diagnostics – Identifying Influential Data and Sources of Collinearity. John Wiley and Sons: New York.
2] Dutta,D and Ahmed,N. 1999. An aggregate import demand function for Bangladesh: a cointegration approach. *Applied Economics* **31**(4): 875-884.
3] Farrar,D.E. and Glauber,R.R. 1967. Multicollinearity in regression analysis: the problem revisited. *Review of Economics and Statistics* **49**: 92-107.
4] Fox,J. and Monette,G. 1992. Generalized collinearity diagnosis. *Journal of the American Statistical Association*. **1993**(87): 178-183.
5] Greene,C.A. 1999. On the impossibility of a stable and low GDP elasticity of money demand: the arithmetic of aggregation, replication and income growth. *Applied Economics* **31**(9): 1119-1127.
6] Hansen,E. 1999. A price-to-market model with unobserved variables: explaining New Zealand's import prices. *Applied Economics*. **31**(1): 3-8.
7] Hoerl,A.E. and Kennard,R.W. 1970a. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**: 55-67.
8] Hoerl,A.E. and Kennard,R.W. 1970b. Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**: 69-82
9] Hoerl,A.E. and Kennard,R.W. 1976. Ridge regression: Iterative estimation of the biasing parameter. *Commun.Stat.* **4**: 105-123.
10] Leamer,E.E. 1973. Multicollinearity: A Bayesian interpretation. *Rev.Econ.Stat*. **55**: 371-380.
11] Leamer,E.E. 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley: New York.
12] Marquardt,D.W. 1970. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* **12**: 591-612.
13] Mason,C.H. and Perreault Jr.,W.D. 1991. Collinearity, power and interpretation of multiple regression analysis. *Journal of Marketing Research*. **28**(3):268-280.
14] Montgomery C.Douglas, Elizabeth A.Peck and Geoffrey Vining,G. 2003. Introduction to Linear Regression Analysis. John Wiley and Sons, INC: New York.
15] Sclove,S.L. 1968. Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association* **63**: 596-606.
16] Stewart,G.W. 1987. Collinearity and least squares regression. *Statistical Science* **32**(1): 68-100.
17] Tukey, John W. (1960). Where do we go from here? *Journal of the American Statistical Association* **55**: 80-93.
18] Vizcarrondo-Carlos Toro and Wallace,T.D. 1968. *Journal of the American Statistical Association* **63**: 558-572.
19] Willian,A.R. and Watts,D.G. 1978). Meaningful multicollinearity measures. *Technometrics* **20**: 407-412.
20] Zellener,A (1971). An Introduction to Bayesian Inference in Econometrics, Wiley, New York.