SINET: Ethiop. J. Sci., 23(1):67–72, 2000
 ISSN: 0379–2897

TOTAL SHRINKAGE VERSUS PARTIAL SHRINKAGE IN MULTIPLE LINEAR REGRESSION

Eshetu Wencheko

Department of Statistics, Faculty of Science
Addis Ababa University, PO Box 1176, Addis Ababa, Ethiopia

ABSTRACT: The paper discusses the merits of partial shrinkage of the ordinary least square estimator of the coefficients of the multiple regression model of full rank. Theoretical comparisons of scalar and matrix-valued risks of the partially shrunken and totally shrunken estimators are given. The strategy of partial shrinkage is applied to two data sets.

Key words/phrases: Partial and total shrinkage, variance inflation factors

INTRODUCTION

We consider the multiple linear regression model $M\{\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I}_n\}$ where $\mathbf{y}$ is the observed n-vector of response variable, $\mathbf{X}$ is a $n \times p$ non-stochastic regression matrix of full column rank, β is a p-vector of unknown but fixed vector of regression coefficients while $\sigma^2 > 0$ is the unknown constant variance of the error terms.

It is known that, although unbiased, the ordinary least squares estimator (LSE) of the regression coefficients $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ has some deficiencies that, at times, make its usefulness questionable. This happens when two or more exogenous variables of the regression matrix are strongly linear dependent. Under such circumstances a family of biased linear homogenous estimators commonly known as generalised ridge estimators (Obenchain, 1975; 1978) can be used instead of $\mathbf{b}$. Members of this family outperform the LSE on sub-domains of the parameter space of the regression coefficients.

Theoretical mean square error (MSE) comparisons between $\mathbf{b}$ and any member of the generalised ridge estimators can be found, among others, in Trenkler (1986) and Trenkler (1981). It is also possible to make theoretical comparisons of MSE's of members within the same family of biased estimators. The following section of this paper is devoted to the MSE comparison of the shrunken estimator $c\mathbf{b}$, $c \in (0, 1)$, (Mayer and Willke, 1973) and the so-called generalised shrunken estimator.

THE GENERALISED SHRUNKEN ESTIMATOR

A member of this class that we consider in this paper is a generalisation of the shrunken estimator (Mayer and Willke, 1973), which will henceforth be referred to as the generalised deterministic shrunken estimator (Eshetu Wencheko, 1998). This estimator has the form $\mathbf{b}(C) = C\mathbf{b}$, $C = \text{diag}(c_j)$, $c_j \in (0, 1]$, $j = 1, \ldots, p$. Note that if all diagonal elements of C are equal we have the estimator of Mayer and Willke (1973).

It is known that the deterministic generalised shrunken estimator has bias vector and covariance matrix $B(C\mathbf{b}) = (C - I_p)\beta$ and $\text{Cov}(C\mathbf{b}) = \sigma^2 C(X'X)^{-1}C$, respectively. The total variance of $C\mathbf{b}$, that is the sum of diagonal elements of $\text{Cov}(C\mathbf{b})$, is $V(C\mathbf{b}) = \sigma^2 \text{ tr } C(X'X)^{-1}C'$. The scalar risk, that is the sum of the squared norm of the bias vector and total variance, is $G(C\mathbf{b}) = \beta'(C-I_p)^2\beta + V(C\mathbf{b})$. Similarly, $B(c\mathbf{b}) = (c-1)\beta$ and $\text{Cov}(c\mathbf{b}) = \sigma^2 c^2 (X'X)^{-1}$ and $G(c\mathbf{b}) = (c-1)^2 \beta'\beta + V(c\mathbf{b})$. Having said the above about the two shrunken estimators we now give the following results with regard to their risks. In the first result MSE stands for the mean square error matrix.

Theorem 1:

$$\begin{aligned} \text{MSE}(C\mathbf{b}) < \text{MSE}(c\mathbf{b}) \Leftrightarrow\ & C(X'X)^{-1}C - c^2 (X'X)^{-1} \\ < & \ [(c-1)^2 \beta\beta' - \text{diag}\,(c_j - 1)^2 \beta\beta']/\sigma^2 \\ = & \ [\text{diag}(c-1)^2 - \text{diag}\,(c_j - 1)^2]\beta\beta'/\sigma^2, \end{aligned}$$

where "$<$" stands for the Löwner order of matrices.

The result below follows from Theorem 1 above:

Theorem 2:

$$G(\mathbf{Cb}) < G(c\mathbf{b}) \Leftrightarrow \operatorname{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}^2 - c^2(\mathbf{X}'\mathbf{X})^{-1}) < [(c-1)^2\boldsymbol{\beta}'\boldsymbol{\beta} - \boldsymbol{\beta}'\operatorname{diag}(c_j - 1)^2\boldsymbol{\beta}]/\sigma^2 = [\operatorname{diag}(c-1)^2 - \operatorname{diag}(c_j - 1)^2]\boldsymbol{\beta}'\boldsymbol{\beta}/\sigma^2.$$

RATIONALE FOR PARTIAL SHRINKAGE

The estimator of Mayer and Willke (1973) provides a biased estimator that is a simple scalar multiple of **b**. The scalar shrinkage factor which is less than unity makes the length of the LSE shorter. The same is true about the generalised shrunken estimator. In the presence of strongly collinear regressors some or all variance inflation factors (VIF) of the covariance matrix of LSE could be unacceptably large. A shrinkage of the LSE by a scalar reduces the magnitude of all of the VIF's irrespective of size. If some of the regressors exhibit mild collinearity and, therefore the VIF's of the LSE estimator lie between 1 and 10, such biased estimators are regarded as quite acceptable. The reasoning behind such a conclusion is based on the fact that whenever the maximum VIF does not exceed 10 or is not smaller than unity the bias is taken as tolerable. Note that in the ideal orthogonal design all diagonal elements of the covariance of the LSE (with standardised regression matrix) are equal to unity. Marquardt (1970), Marquardt and Snee (1975), Montgomery and Askin (1981), Snee (1983) and Trenkler (1981) had advocated the foregoing rationale. This basis is used to motivate the introduction of partial shrinkage strategy.

We notice that a single shrinkage constant does not take into consideration the magnitude of individual VIF's; it simply shrinks all components indiscriminately and, thus the naming total shrinkage. The strategy employed in what we will call partial shrinkage would leave those VIF's with acceptable size unaltered – no need to shrink these. On the other hand, a selective shrinkage will be performed with regard to the remaining ones that do not fall within what is sometimes referred to as the Marquardt interval, that is simply the interval [1, 10]. The shrinkage matrix **C** should also be chosen such that the coefficient of determination R^2 will be acceptable.

EXAMPLES - DEMONSTRATION OF PARTIAL SHRINKAGE

In this section we demonstrate how partial shrinkage can be applied to two data sets that have been considered in the regression literature as typically collinear. The first of these is a result from a chemical engineering experiment (Hald 1952, p. 645ff), and the severity of multicollinearity of the data is expressed by the condition number 37.1063. The second data set from Montgomery and Askin (1981) deals with household-level electricity consumption. The degree of collinearity is given by the condition number 36.6544. In this paper we call the quotient of the largest spectral value to the smallest (of X'X) as the condition number of the matrix X.

The LSE for each data set and the corresponding VIF's are given in Table 1 and Table 2, respectively. In the case of the first data set all VIF's are greater than 10, while in the second two components (third and fifth) of the LSE have acceptable VIF's.

Table 1. Results for the chemical engineering data.

b_j	31.6071	27.5003	2.2612	-8.3531
v_j	38.4962	254.4232	46.8684	282.5129
c_j^2	0.0260	0.0039	0.0213	0.0035
c_j	0.1612	0.0627	0.1459	0.0595
$c_j b_j$	5.0951	1.7241	0.3300	-0.4970
$\sqrt{5}\, c_j b_j$	11.3930	3.8852	0.7379	-1.1113
$\sqrt{10}\, c_j b_j$	16.1121	5.4521	1.0436	-1.5717

Table 2. Results for household-level electricity consumption data.

b_j	-8.0726	9.7066	5.0385	3.4468	0.3226
v_j	191.2968	180.6651	3.9587	11.4890	1.0556
c_j^2	0.0052	0.0055	*	0.0870	*
c_j	0.0723	0.0744	*	0.2950	*
$c_j b_j$	-0.5837	0.7217	*	1.0169	*
$\sqrt{5}\, c_j b_j$	-1.3051	1.6137	*	2.2738	*
$\sqrt{10}\, c_j b_j$	-1.8458	2.2822	*	3.2157	*

*** left unchanged.**

For the purpose of demonstration we have chosen the reciprocals of those inflated VIF's and two multiples of the same to obtain a factor, say $c_j^2 = v_j^{-1}$, (the reciprocal of the *jth* VIF) component to reduce the VIF of the corresponding component of the partially shrunken estimator $c_j b_j$ to 1. Similarly, using $c_j^2 = 5/v_j$ and $c_j^2 = 10/v_j$ will reduce the associated inflation of b_j to 5 and 10, respectively. These resultant shrinkage factors are simply scalar multiples of the c_j's, namely $\sqrt{5}\ c_j$ and $\sqrt{10}\ c_j$ .The Tables provide the vectors of the three partially shrunken estimators of the coefficients of regression for the two models. In general, the partially shrunken estimators for the two data sets can be given as **Cb** with

$$\mathbf{C} = m\ diag\ (5.0951,\ 1.7241,\ 0.3300,\ -0.4970)$$

and

$$\mathbf{C} = m\ diag\ (-0.5837,\ 0.7217,\ 1,\ 1.0169,\ 10),$$

where $m \in [1, \sqrt{10}]$ is stochastic. In this paper only three possible m-values that gave rise to three shrunken estimators have been considered. Nonetheless, since m is random a simulation study can be conducted to generate as many partially shrunken estimators as needed.

Note that in Table 2 the two columns marked with asterisk are those where multiplication of the VIF's is not necessary. This means the third and fifth components remain unchanged.

Evidently we also observed that both total shrinkage and partial shrinkage do not bring about any change on the arithmetic signs of the components of the shrunken estimator. Nonetheless, partial shrinkage is a strategy that aims at reducing the inflated VIF's. This is achieved by appropriate choices of multipliers that will have distinct effects on the magnitude of the corresponding estimator component.

SUMMARY

The paper justifies the need for shrinkage of only those components of the standard estimator of the regression coefficients with large VIF's in the sense

discussed in the sections above. Obviously the amount by which we shrink a component introduces bias that is tolerable as long as the resulting VIF is within the permitted interval. Those component estimators that are left unaltered have acceptable variance and they are still unbiased. The idea of partial shrinkage is similar to that of estimation of a sub-vector.

REFERENCES

1. Eshetu Wencheko (1998). A generalised deterministically shrunken estimator. *SINET: Ethiop. J. Sci.* **21(2)**:273–277.
2. Hald, A. (1952). *Statistical Theory with Engineering Applications*. John Wiley, New York.
3. Marquardt, D.W. (1970). Generalised inverses, ridge regression, biased linear estimation and non-linear estimation. *Technometrics* **12**:591–612.
4. Marquardt, D.W. and Snee, R.D. (1975). Ridge regression in practice. *The American Statistician* **29**:3–20.
5. Mayer, L.S. and Willke, T.A. (1973). On biased linear estimation in linear models. *Technometrics* **15**:497–508.
6. Montgomery, D.C. and Askin, R.G. (1981). Problems of non-normality and multicollinearity for forecasting methods based on least squares. *American Institute of Industrial Engineers Transactions* **13**:102–114.
7. Obenchain, R.L. (1975). Ridge analysis following a preliminary test of the shrunken hypothesis. *Technometrics* **17**:431–445.
8. Obenchain, R.L. (1978). Good and optimal ridge estimators. *Annals of Statistics* **6**:1111–1121.
9. Snee, R.D. (1983). Discussion in "Developments in linear regression methodology: 1959–1982". *Technometrics* **25**:230–237.
10. Trenkler, D. (1986). *Verallgemeinerte Ridge Regression*. Mathematical Systems in Economics 104. Anton Hain Verlag, Meisenheim, Frankfurt/Main.
11. Trenkler, G. (1981). *Biased Estimators in the Linear Regression Model*. Mathematical Systems in Economics 58. Oelgeschlager, Gunn and Hain Verlag, Cambridge, Massachusetts.