AN OPTIMAL PROPERTY OF THE (r, k)-CLASS ESTIMATORS IN THE REGRESSION MODEL WITH CONCOMITANT VARIABLES

José Raúl Martínez

Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Av. Haya de la Torre y Medina Allende
Ciudad Universitaria, 5000 – Córdoba – ARGENTINA
e-mail: jmartine@mate.uncor.edu

ABSTRACT

In the present paper we display an optimal property, in terms of the Generalized Mean Square Error Criterion, of the class estimators in the regression model with concomitant variables.

Key words and phrases: Concomitant variables; generalized mean square error, (r, k)-class estimators.

I INTRODUCTION

The uses of concomitant variables in many fields of biological, physical, industrial, and economical disciplines are well known and proved to be important and useful. In the last time, the use of concomitant information in the evaluation of mortality data or other age – dependent all – or – none phenomena (e. g.; onset of chronic disease) has attracted the attention of many researches. It is also known that researchers are faced every day with the problems of multicollinearity. The statistical consequences of this in a linear regression model have been studied in great detail. It is well known that in situations of multicollinearity, it becomes difficult to obtain precise estimates of separate effects of the variables involved in the regression model. The method of least squares produces large sampling variance of the estimated regression coefficients, which in turn gives rise to the possibility that otherwise significant coefficients may be dropped from the analysis improperly.

In order to circumvent these problems, many alternative estimator have been suggested.

Among them the (r, k)-class estimators. This class was introduced by Baye and Parker (1984), combining the ordinary ridge regression estimator (**ORR**) (Hoerl and Kennard (1970)) and the principal components regression estimator (**PCR**) (Marquardt (1970)) as alternative estimators to the ordinary least square estimator (**OLS**) in the linear regression model under the multicollinearity explanatory variables.

Martínez (1990) displays an optimal property, in terms of the Generalized Mean Square Error criterion, of this class of estimators. (See Appendix for further details).

In the present paper we display conditions, similar to those in Martínez (1990), under which there exist values r_0 and k_0 such that the (r_0, k_0)-estimator exhibits,

simultaneously, Generalized Mean Square Error (GMSE) less than GMSE of ORR, PCR and OLS in the linear regression model with concomitant variables.

II. THE MODEL AND NOTATION

Let us consider the model (Rao, (1973)),

$$Y^* = X\beta + \mu, \qquad \mu \equiv C\gamma + \varepsilon, \qquad (2.1)$$

where Y^* is an n dimensional column vector of response variables, β, γ are p, q dimensional column vectors of unknown regression coefficient parameters, respectively, and X, C, ε, satisfy the following assumptions:

i) X is an nxp matrix of explanatory vectors standardized in such way X'X is a correlation matrix and C an nxq matrix of concomitant vectors, which are observable and satisfy $\text{rank}(X, C) = p + q\ (\leq n)$ and $X'C \neq 0$.

ii) ε is an n dimensional column vector of errors with the following mean vector and variance matrix:

$$E(\varepsilon) = 0 \qquad \text{Var}(\varepsilon) = \sigma^2 I,$$

where I is the n dimensional identity matrix and σ^2 is an unknown variance of error.

Let $S = [S_1, S_2, ..., S_p]$ be an orthogonal matrix such that $S'X'XS = D$, where $D = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$ are eingenvalues of X'X.

Let $\alpha = S'\beta$ and $Z = XS$, then (2. 1) can be written as

$$Y^* = XSS' + \mu$$

$$Y^* = Z\alpha + \mu$$

In this way we can express the estimators of α in terms of the eingenvalues $\lambda_1, \lambda_2, ..., \lambda_p$.

Note that if β is any particular estimator of β then $\alpha = S'\beta$ is the corresponding estimator of α and $\alpha'\alpha = \beta'SS'\beta = \beta'\beta$.

In this model, we consider the (r, k)-class estimators for regression coefficients without using concomitant variables explicitly:

$$\beta^*(r, k) = Sr(Dr + kIr)^{-1}Sr'X\,Y^*, \qquad (2.2)$$

where $0 \leq r \leq p$, $k \geq 0$, $Sr = [S_1, S_2, ..., S_r]$ and $Dr = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$.

Note that $\beta^*(r, k)$ is a linear transformation of OLS estimator β^*. In effect,

$$\beta^*(r, k) = H\beta^*$$

where $H = Sr(Dr + kIr)^{-1}Sr'$.

The (r, k)-class is a general estimator which includes the ORR, PCR and OLS estimators as special cases:

$\beta^*(p, k) = \beta^*(k)$, the ORR estimator

$\beta^*(r, 0) = \beta^*r$, the PCR estimator

$\beta^*(p, 0) = \beta^*$, the O LS estimator.

The Generalized Mean Square Error (GMSE) of $\beta^*(r, k)$ is given by

$$GMSE\beta^*(r, k) = E[\beta^*(r, k) - \beta]'W[\beta^*(r, k) - \beta],$$

where $W = I$ or $W = X'X$.

When $W = I$, the $GMSE\beta^*(r, k)$ represents the MSE of the estimator of the regression coefficients. When $W = X'X$, the $GMSE\beta^*(r, k)$ represents the MSE of the predictor.

It is not difficult to see that

$$GMSE\beta^*(r, k) = GMSE\beta(r, k) + 2\sum_{i=1}^{r} \alpha_i \eta_i(\mu_i/\lambda i)\delta_i[\delta_i - 1] + \sum_{i=1}^{r} \eta_i^2\mu_i\delta_i^2/\lambda_i^2$$

where $\beta(r, k)$ is the (r, k) –estimator in the model without concomitant variables,

$$\mu_i = \begin{cases} 1 & \text{for all } i = 1, 2, ..., p \quad \text{when } W = I \\ \lambda_i & \text{for all } i = 1, 2, ..., p \quad \text{when } W = X'X \end{cases}$$

$\delta_i = \lambda i/(\lambda i + k)$ for all $i = 1, 2, ..., p$ and η_i is the i-esima component of the vector $\eta \equiv S'X'C\gamma$.

In this section we display an optimal property of the $\beta^*(r, k)$ estimators.

III. THE MAIN RESULT

Theorem: If $\alpha'\alpha$ is bounded and r_0 is such that $\sigma^2/2\alpha'\alpha > \lambda r_{0+1}$ and $\alpha_i\eta_i \geq 0$ for $i = r_{0+1}, ..., p$, then for all $0 < k < \min\{ \sigma^2/4 \max_{i=1,...,p} \alpha_i^2 ; \min_{i=r_0+1,...,p} = \lambda i / |\alpha_i|(\sigma\lambda i^{-1/2} - |\alpha_i|)\}$

we have:

i) $GMSE\beta^*(r_0, k) < GMSE\beta^*(k)$
ii) $GMSE\beta^*(r_0, k) < GMSE\beta^*(r_0)$

Proof of i): From (2.3) and recalling that for each integer r $(1 \leq r \leq p)$ and $k \geq 0$ (See Nomura and Okhubo, 1985)

$$GMSE\beta(r, k) = \sum_{i=1}^{r} \mu_i[\alpha_i^2(\delta_i - 1)^2 + \sigma^2 \delta_i^2/\lambda_i] + \sum_{i=r}^{p} \mu_i \alpha_i^2$$

we obtain:

$$GMSE\beta^*(r_0, k) - GMSE\beta^*(k) = \sum_{i=ro+1}^{p} \mu_i(\alpha_i^2 - \sigma^2 \delta_i^2/\lambda_i] - \sum_{i=ro+1}^{p} \mu_i/(\lambda_i + k)^2 (\alpha_i k - \eta_i)^2$$

$$\equiv T_1 + T_2$$

Since $\mu_i > 0$ for all i = 1, 2 , ..., p, results $T_2 > 0$. On the other hand, we have

$0 < k < \min\{ \sigma^2/4 \max_{i=1,...,p} \alpha_i^2 ; \min_{i=ro+1,...,p} = \lambda i / |\alpha_i|(\sigma \lambda i^{-1/2} - |\alpha_i|)\}$, then it is not difficult to see

that $|\alpha_i| < \sigma \lambda i^{-1/2} \delta_i$ and since $\mu_i > 0$ for all i, it follows that $T_1 < 0$. Hence i) is proved.

Proof of ii): It follows that from (2.3) that

$$GMSE\beta^*(r_0, k) - GMSE\beta^*(r_0) = [GMSE\beta(r_0, k) - GMSE\beta(r_0)]$$
$$+ 2\sum_{i=1}^{r_0} \alpha_i \eta_i(\mu_i/\lambda i)\delta_i[\delta_i - 1]$$
$$+ \sum_{i=1}^{r_0} \eta_i^2 \mu_i/\lambda_i^2(\delta_i^2 - 1)^2$$
$$\equiv T_1 + T_2 + T_3.$$

From Martínez (1990) (see Appendix) it follows that $T_1 < 0$. Now, since $\mu_i > 0$, $\delta_i < 1$ and
$\alpha_i \eta_i \geq 0$ for all i = 1, 2,..., p we obtain $T_2 < 0$ and $T_3 < 0$. Thus ii) is proved.

Collorary: For r_o and k we have

a) $GMSE\beta^*(r_0, k) < GMSE\beta^*(r_0) < GMSE\beta^*$
b) $GMSE\beta^*(r_0, k) < GMSE\beta^*(k) < GMSE\beta^*$

Proof: We need only to prove the second inequality of a) and b).

$$GMSE\beta^*(r_0) - GMSE\beta^* = [GMSE\beta(r_0) - GMSE\beta] - \sum_{i=ro+1}^{p} \eta_i^2 \mu_i/\lambda_i^2$$

$$\equiv T_1 + T_2.$$

From Martínez (1990) (see Appendix) it follows that $T_1 < 0$ and since $\mu_i > 0$ for all i results
$T_2 < 0$ and thus a) is proved.
Finally,
$GMSE\beta^*(k) - GMSE\beta^* =$

$$[GMSE\beta(k) - GMSE\beta]$$

$$+ 2\sum_{i=1}^{p} \alpha_i \eta_i(\mu_i/\lambda i)\delta_i[\delta_i - 1] + \sum_{i=1}^{ro} \eta_i^2\mu_i/\lambda_i^2(\delta_i^2 - 1)^2$$

$$\equiv T_1 + T_2 + T_3.$$

Since $\mu_i > 0$, $\delta_i < 1$ and $\alpha_i \eta_i \geq 0$ for all i = 1, 2,..., p we obtain $T_2 < 0$ and $T_3 < 0$. From Martínez (1990) it is easy to see that $T_1 < 0$. This proves the corollary.

IV APPENDIX

Theorem: If $\alpha'\alpha$ is bounded and r_o is such that λ_{ro+1} then for all $0 < k < \sigma^2/4\max_{i=1,...,p} \alpha_i^2$

we have:

i) $GMSE\beta(r_0, k) < GMSE\beta(r_0)$
ii) $GMSE\beta(r_0, k) < GMSE\beta(k)$,

Corollary: For r_0 and k, we have

i) $GMSE\beta(r_0, k)\ GMSE\beta(r_0, k) < GMSE\beta$
ii) $GMSE\beta(r_0, k) < GMSE\beta(k) < GMSE\beta$

$\beta(r_0, k)$ is the (r_0, k)- estimator of the parameter β in the linear regression model.

Acknowledgement

The author is indebted to the referees for their helpful comments and suggestions which led to an improved presentation of the results. The research for this paper was supported partially by CNPq – Brasil and CONICOR –Córdoba –Argentina.

REFERENCES

Baye, M. R. and Parker, D. F. (1984). Combining Ridge and Principal Component.
Regression: A Money Demand Illustration. *Communications in Statistics 13 (2), 197 – 205.*

Hoerl, A.E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation of Nonorthogonal Problems. *Technometrics 12, 55 – 67.*

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

Martínez, J. R. (1990). An optimal property of the (r, k)-class estimators. *Communications in Statistics 19(4), 1281 – 1289.*

Nomura, M. and Ohkubo, T.(1985). A note on Combining Ridge and Principal Component Regression. *Communications in Statistics 14(10), 2489 – 2493.*

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. *John Wiley.*