JANVIER 1982

N° 8207

SOME THEORETICAL RESULTS FOR GENERALIZED RIDGE REGRESSION ESTIMATORS

C. FOURGEAUD[*], C. GOURIEROUX[**], J. PRADEL[***]

[*] PARIS I et CEPREMAP

[**] PARIS IX et CEPREMAP

[***] PARIS I

1. INTRODUCTION.

The ridge regression procedure has been introduced by HOERL-KENNARD (1970), initially to avoid the ill effects of quasi-collinearity on ordinary least squares estimators. This method consists in adding a symmetric positive matrix (biasing factor) to the design matrix.

In this paper, we examine some interpretations and theoretical properties of the ridge regression estimators. In particular we want to discuss the usual justifications of this method, such as :

- the improvement of the OLS estimator for some values of the parameters
- the possibility to include some a priori information on the parameters
- the property to avoid "wrong signs"
- and, finally, the advantage of its use to solve the difficulties introduced by quasi-collinearity of the explanatory variables.

In section 2, we recall the expressions of the ordinary and generalized ridge regression estimators (ORR and GRR) ; then we examine the connection with the principal component estimators and the effect of the biasing factor on the length and on the sign of the estimators.

The GRR estimator can be interpreted as an OLS estimator based on transformed explanatory variables and can be characterized in the class of such OLS estimators (section 3).

In section 4, we construct a confidence region centered at the GRR estimator and we compare it with the OLS'one.

Section 5 is devoted to the proof that, contrary to the usual belief, the OLS estimator gives most frequently the right signs of the parameters. The optimisation problem, which has to be solved, appears to be the dual problem of that associated with Gauss Markov theorem.

Finally, in section 6, using the interpretation of the GRR estimator as an OLS one, we derive, in the two parameters case, the necessary and sufficient condition for the transformed explanatory variables to be less collinear than the initial ones. This condition is compared with the condition for improving the OLS estimator.

2. RIDGE ESTIMATORS.

2.a The model.

Let us consider the classical linear model : $Y = Xb + u$, where Y is the $(n,1)$ vector of observations on the endogenous variable, X is the (n,G) matrix of observations on G exogenous variables assumed of full rank G ; b is the vector of unknown parameters and u is the error vector , such that $Eu = 0$ and $Vu = \sigma^2 I$.

The matrix X can be expressed using the spectral representation : $X = \sum_{i=1}^{G} \mu_i P_i Q_i'$, where the n-dimensional (resp. G.

dimensional) vectors P_i (resp. Q_i) are mutually orthogonal with unit length. Denoting by Q the (G,G) orthogonal matrix, whose column vectors are the Q_i $i = 1,\dots,G$ it follows :

$$X'X = \sum_{i=1}^{G} \mu_i^2 Q_i Q_i' = Q \ \mathrm{diag}(\mu^2)Q' = Q \ \mathrm{diag}(\lambda)Q'$$

where diag(λ) is the diagonal matrix, whose diagonal elements λ_i are the eigenvalues of $X'X$: $\lambda_i = \mu_i^2 > 0$, $i = 1,\dots,G$.

2.b The generalized ridge regression estimator (G.R.R.)

The GRR estimator, introduced by HOERL-KENNARD (1970a,b), is defined by :

$$(1) \qquad \tilde{b}(X,k) = [X'X + Q \ \mathrm{diag}(k)Q']^{-1} X'Y$$

where diag(k) is a diagonal matrix of biasing factors k_i . In the particular case of equal k_i : $k_i = K$, $i = 1,\dots,G$, the previous estimator is simply the ordinary ridge regression estimator (O.R.R.) :

$$(2) \qquad \tilde{b}(X,K) = (X'X + KI)^{-1} X'Y$$

The O.L.S. estimator $\hat{b} = (X'X)^{-1}X'Y$ is obtained for $K = 0$.
The G.R.R. estimator is generally biased :

$$(3) \qquad E \ \tilde{b}(X,k) = Q \ \mathrm{diag}(\frac{\lambda}{\lambda+k})Q'b$$

its covariance matrix is given by :

(4) $$V\,\tilde{b}(X,k) = \sigma^2\, Q\, \mathrm{diag}\left[\frac{\lambda}{(\lambda+k)^2}\right] Q'$$

Therefore the mean squared error matrix of $\tilde{b}(X,k)$ is :

(5) $$R[\tilde{b}(X,k)] = E[\tilde{b}(X,k) - b]\,[\tilde{b}(X,k) - b]'$$

$$= \sigma^2\, Q\, \mathrm{diag}\left[\frac{\lambda}{(\lambda+k)^2}\right] Q' + Q\, \mathrm{diag}\left(\frac{k}{\lambda+k}\right) Q'bb'Q\, \mathrm{diag}\left(\frac{k}{\lambda+k}\right) Q$$

2.c Linear transformation on the parameters.

If the initial model is reparametrized : $Y = Z\beta + u$ where $Z = XD$ and $\beta = D^{-1}b$, the OLS estimator of β is deduced from the OLS estimator of b by : $\hat{\beta} = D^{-1}\,\hat{b}$. This property is no longer valid for GRR estimator. $Z'Z = D'X'XD = D'Q\,\mathrm{diag}(\lambda)\,Q'D$ can be written as : $Z'Z = Q^*\,\mathrm{diag}(\lambda^*)\,Q^{*\prime}$,where Q^* is an orthogonal matrix and λ_i^* , $i = 1,\dots,G$, are the eigenvalues of $Z'Z$. Therefore the GRR estimator of β is :

$$\tilde{\beta}(Z,k) = [Z'Z + Q^*\,\mathrm{diag}(k)Q'^{*}]^{-1}\, Z'Y$$
$$= D^{-1}\,[X'X + D'^{-1}\, Q^*\,\mathrm{diag}(k)\, Q^{*\prime}\, D^{-1}]^{-1}\, X'Y$$

and is equal to $D^{-1}\,\tilde{b}(X,k)$ iff :

$$D'^{-1}\, Q^*\mathrm{diag}(k)Q^{*\prime}\, D^{-1} = Q\,\mathrm{diag}(k)\,Q'$$

These two matrices are in general distinct, for instance when the linear transformation corresponds to a change of units : $D = \mathrm{diag}(\alpha)$, where $\alpha_i > 0$. However, if D is an orthogonal matrix, $D'Q$ is also orthogonal, $Q^* = D'Q$, $\lambda^* = \lambda$ and $\tilde{\beta}(Z,k) = D^{-1}\,\tilde{b}(Z,k)$. In particular for $D = Q$, we have : $\beta = Q'b$, $Z'Z = \mathrm{diag}(\lambda)$; the explanatory variables Z_i are mutually orthogonal and the GRR estimator is obtained from the O.L.S. estimator by a diagonal matrix :

$$\tilde{\beta}_i(Z,k) = \frac{\lambda_i}{\lambda_i + k_i} \hat{\beta}_i \quad , \quad i = 1,\ldots,G \; .$$

If $k_i > 0$, the term $\frac{\lambda_i}{\lambda_i + k_i}$ may be interpreted as a shrinkage fraction (VINOD (1977)).

2.d G.R.R. estimator and principal component estimator.

Let us assume that the eigenvalues are indexed such that : $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_G$. The principal component estimator of order r $(r < G)$:

$$\tilde{b}_r = \sum_{i=1}^{r} \frac{1}{\mu_i} Q_i P_i' Y = \sum_{i=1}^{r} \frac{1}{\sqrt{\lambda_i}} Q_i P_i' Y$$

may be considered as a limit case of G.R.R. estimator. It can be obtained by choosing : $k_i = 0 \quad i = 1,\ldots,r$, $k_i = +\infty$, $i = r+1,\ldots,G$ (MARQUARDT (1970)).

In this section, we are looking for an ORR estimator giving the best approximation of $\tilde{b}_r$.

The O.R.R. estimator may be decomposed into :

$$\tilde{b} = \sum_{i=1}^{G} \frac{\sqrt{\lambda_i}}{\lambda_i + K} Q_i P_i' Y$$

The criterium to minimize is the norm $e(K)$ of the difference between the linear mappings transforming Y in $\tilde{b}_r$ and $\tilde{b}$ respectively. It is given by :

$$e^2(K) \quad = \quad \text{Tr } AA' \qquad \text{with} \quad A = \sum_{i=1}^{r} \frac{1}{\sqrt{\lambda_i}} Q_i P_i' - \sum_{i=1}^{G} \frac{\sqrt{\lambda_i}}{\lambda_i + K} Q_i P_i'$$

Replacing A by its expression :

$$e^2(K) \quad = \quad \sum_{i=1}^{r} \left(\frac{\sqrt{\lambda_i}}{\lambda_i + K} - \frac{1}{\sqrt{\lambda_i}} \right)^2 + \sum_{i=r+1}^{G} \left(\frac{\sqrt{\lambda_i}}{\lambda_i + K} \right)^2$$

$$= \sum_{i=1}^{r} \frac{K^2}{\lambda_i \, (\lambda_i + K)^2} + \sum_{i=r+1}^{G} \frac{\lambda_i}{(\lambda_i + K)^2}$$

The first order condition is :

$$\frac{de^2(K)}{dK} = \sum_{i=1}^{r} \frac{2K}{(\lambda_i + K)^3} - 2 \sum_{i=r+1}^{G} \frac{\lambda_i}{(\lambda_i + K)^3} = 0$$

Since $\dfrac{de^2(o)}{dK} = -2 \displaystyle\sum_{i=r+1}^{G} \frac{1}{\lambda_i^2} < 0$ and $\dfrac{de^2(K)}{dK} \sim \dfrac{2r}{K^2} \geq 0$

if K tends to infinity, it is easily seen that the previous equation always has at least a solution corresponding to a minimum of $e^2(K)$.

2.e Study of the "shrinkage" effect.

We have seen in subsection 2.c that the ridge procedure with $k_i \geq 0$ implies a shrinkage of the OLS estimator of the principal parameters $\beta = Q'b$. Consequently, we have : $||\tilde{b}(X,k)||^2 = ||\tilde{\beta}(X,k)||^2 \leq ||\hat{\beta}||^2 = ||\hat{b}||^2$. However this shrinkage effect is only a global one and it is interesting to examine this effect on each coordinate of the estimator. To simplify the study, we consider the case of two exogenous variables $G = 2$ such that : $X'X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

The ORR estimator is given by :

$$\tilde{b}(X,K) = (X'X + KI)^{-1} X'X \hat{b} = \frac{1}{(1+K)^2 - \rho^2} \begin{bmatrix} 1+K-\rho^2 & K\rho \\ K\rho & 1+K-\rho^2 \end{bmatrix} \hat{b}$$

and the first coordinate of $\tilde{b}(X,K)$ is :

$$\tilde{b}_1(X,K) = \frac{1}{(1+K)^2 - \rho^2} [(1+K-\rho^2) \hat{b}_1 + K\rho \hat{b}_2]$$

$\tilde{b}_1(X,K)$ is a function of $K \in [0,\infty[$, whose first derivative is :

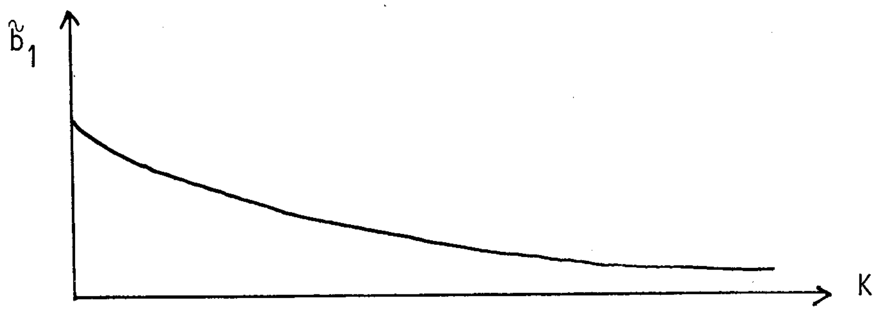
$$\frac{d \tilde{b}_1(X,K)}{dK} = \frac{1}{\{(1+K)^2 - \rho^2\}^2} \{K^2(-\hat{b}_1 - \rho \hat{b}_2) + 2 K(\rho^2-1) \hat{b}_1 + (\rho^2 - 1)(\hat{b}_1 - \rho \hat{b}_2)\}$$

This derivative has the same sign as a polynomial of degree two, whose discriminant is :

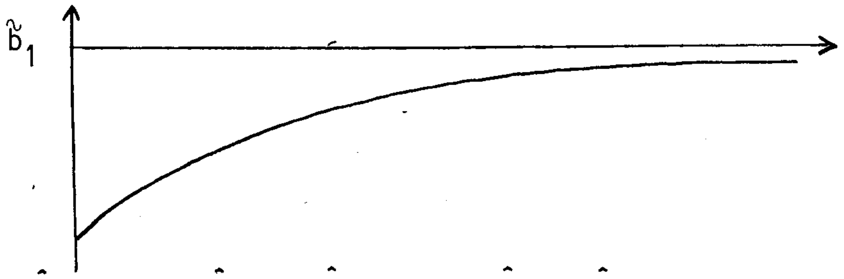
$$\Delta' = \rho^2(1-\rho^2) (\hat{b}_2^2 - \hat{b}_1^2)$$

Several cases have to be distinguished depending on the values of $\hat{b}_1$ and $\hat{b}_2$. The graph of $\tilde{b}_1(X,K)$ is given below in each of these cases :

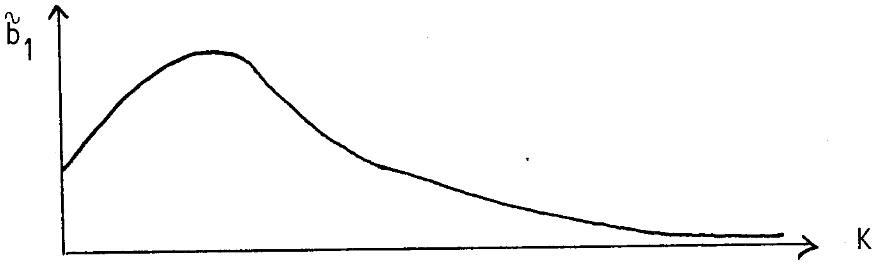
1) $\hat{b}_1 > 0$; $|\hat{b}_1| > |\hat{b}_2| \, |\rho|$



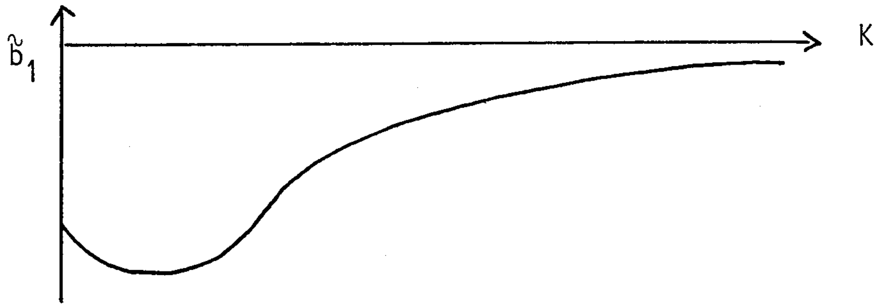
2) $\hat{b}_1 < 0$; $|\hat{b}_1| > |\hat{b}_2|\ |\rho|$



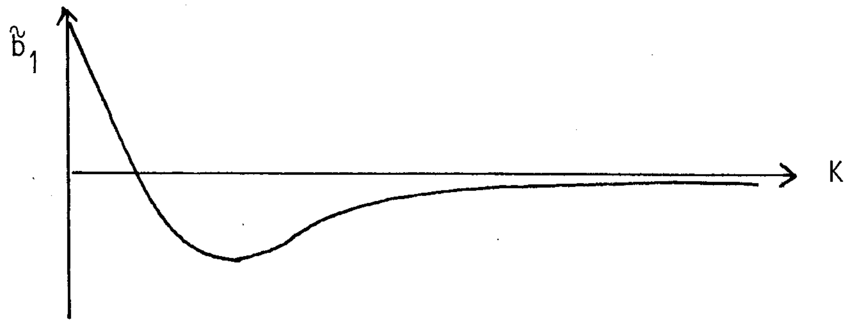
3) $\hat{b}_1 > 0$; $\hat{b}_1 + \rho\,\hat{b}_2 > 0$; $\hat{b}_1 - \rho\,\hat{b}_2 < 0$



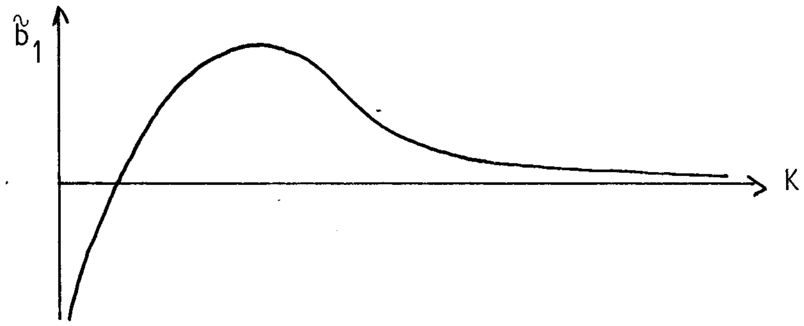
4) $\hat{b}_1 < 0$; $\hat{b}_1 + \rho\,\hat{b}_2 < 0$; $\hat{b}_1 - \rho\,\hat{b}_2 > 0$



5) $\hat{b}_1 > 0$; $\hat{b}_1 + \rho\,\hat{b}_2 < 0$; $\hat{b}_1 - \rho\,\hat{b}_2 > 0$



6) $\hat{b}_1 < 0 \quad ; \quad \hat{b}_1 + \rho\, \hat{b}_2 > 0 \quad ; \quad \hat{b}_1 - \rho\, \hat{b}_2 < 0$



We see from this example that :

- the absolute value of $\tilde{b}_1(X,K)$ may increase (see 3) and 4))
- there may be one and at most one change of sign (see 5) and 6)). This result can be also deduced from a property shown by MACDONALD (1980) .

Finally let us remark that, when there is a change of sign, the limit sign of $\tilde{b}_1(X,K)$ is not always equal to the sign of b_1 . The problem of "right" sign of the estimator of b_1 will be studied in greater detail in section 5.

3. THE G.R.R. ESTIMATOR AS AN OLS ESTIMATOR.

3.a An interpretation of the G.R.R. estimator.

The estimator $\tilde{b}(X,k)$ may be written under the form $\tilde{b}(X,k) = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$ and therefore may be interpreted as an O.L.S. estimator.

PROPERTY 1 :

$$\tilde{b}(X,k) = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$$

Where $\tilde{X} = \sum_{i=1}^{G} \frac{\lambda_i + k_i}{\sqrt{\lambda_i}} P_i Q_i' = X + \sum_{i=1}^{G} \frac{k_i}{\sqrt{\lambda_i}} P_i Q_i'$

Proof :

$$(\tilde{X}'\tilde{X})^{-1} = Q \operatorname{diag}\left(\frac{\sqrt{\lambda}}{\lambda+k}\right)^2 Q' = Q \operatorname{diag}\left(\frac{\lambda}{(\lambda+k)^2}\right) Q'$$

and $$(\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y = \left(\sum_{i=1}^{K} Q_i Q_i' \frac{\lambda_i}{(\lambda_i+k_i)^2}\right)\left(\sum_{j=1}^{K} \frac{\lambda_j+k_j}{\sqrt{\lambda_j}} Q_j P_j'\right) Y$$

$$= \sum_{i=1}^{K} \frac{\sqrt{\lambda_i}}{\lambda_i+k_i} Q_i P_i' Y$$

Therefore $(\tilde{X}'\tilde{X})^{-1} \tilde{X}'y = \tilde{b}(X,k)$

Q.E.D.

PROPERTY 2 : The column vectors of $\tilde{X}$ are linear combinations of the column vectors of X :

$\tilde{X} = XC$ with $C = Q \operatorname{diag}\left(\frac{\lambda+k}{\lambda}\right) Q'$

Proof :

$$XQ\operatorname{diag}\left(\frac{\lambda+k}{\lambda}\right) Q' = \left[\sum_{i=1}^{G} \sqrt{\lambda_i} P_i Q_i'\right]\left[\sum_{j=1}^{G} \frac{\lambda_j+k_j}{\lambda_j} Q_j Q_j'\right]$$

$$= \sum_{i=1}^{G} \frac{\lambda_i+k_i}{\sqrt{\lambda_i}} P_i Q_i' = \tilde{X}$$

Q.E.D.

Therefore to estimate b by ridge regression is equivalent to estimating $b^* = C^{-1} b$ by O.L.S. and to considering that the resulting estimator is an estimator of b (and not of b^*) . $\tilde{b}$ can be viewed as an error on variables estimator, the "right" explanatory variables X being replaced by the erroneous ones $\tilde{X}$.

3.b The risk function of the G.R.R. and of the OLS estimators.

First let us remark that the comparison of the risk matrices $R[\tilde{b}(X,k)]$ and $R(\hat{b})$ for the usual order relation $(\ll)$ on symmetric matrices is equivalent to the comparison of $R[\tilde{\beta}(X,k)]$ and $R(\hat{\beta})$.

The condition $R[\tilde{\beta}(X,k)] \ll R(\hat{\beta})$ can be written as :

$$\sigma^2 \operatorname{diag}\left(\frac{\lambda}{(\lambda+k)^2}\right) + \operatorname{diag}\left(\frac{k}{\lambda+k}\right) \beta\beta' \operatorname{diag}\left(\frac{k}{\lambda+k}\right) \ll \sigma^2 \operatorname{diag}\left(\frac{1}{\lambda}\right)$$

(6) $$\frac{\beta\beta'}{\sigma^2} \ll \operatorname{diag}\left(\frac{2}{k} + \frac{1}{\lambda}\right)$$

The latter condition may be interpreted in several ways :

a) For the ridge regression estimation procedure to be better than the OLS one, the k_i must be choosen according to condition (6) . This condition depends on the unknown parameter $\frac{\beta}{\sigma}$. It is in particular satisfied, if : $\frac{\|\beta\|^2}{\sigma^2} \leq \underset{i}{\operatorname{Min}} \left(\frac{2}{k_i} + \frac{1}{\lambda_i}\right)$.

In the case of ORR procedure, this sufficient condition becomes :

$\frac{\|\beta\|^2}{\sigma^2} \leq \frac{2}{K} + \frac{1}{\lambda_1}$ (λ_1 is the largest eigenvalue of $X'X$). Therefore

$\frac{2}{K} \geq \frac{\|\beta\|^2}{\sigma^2} - \frac{1}{\lambda_1}$.

If K is contrained to be positive, this condition is implied by the THEOBALD's one (THEOBALD (1974)) : $K \leq \frac{2\sigma^2}{\|\beta\|^2} = \frac{2\sigma^2}{\|b\|^2}$.

b) Another approach consists in giving a priori values for the constants k_i and in determining the values of the parameters, for which the estimation procedure is improved by G.R.R..

PROPERTY 3 : Inequality (6) will be satisfied for some values of $\frac{\beta}{\sigma}$, if and only if

$k_i > 0$ or $k_i < - 2 \lambda_i$ $\forall$ i = 1,...,G

Proof :

In effect the G.R.R. estimator is preferable to the OLS one for some values of $\frac{\beta}{\sigma}$ if and only if the matrix $\operatorname{diag}\left(\frac{2}{k} + \frac{1}{\lambda}\right)$ is non negative. This condition is equivalent to $k_i(2\lambda_i+k_i) \geq 0$ $\forall$ i = 1,...,G , which gives the result.

Q.E.D.

In spite of the usual pratice, the OLS procedure can be improved for some values of the parameters by choosing the constants k_i sufficiently negative $(k_i < - 2 \lambda_i)$. When the k_i's are fixed according to the conditions of property 3, the domain of parameters where the GRR estimator is better than the OLS one is given by :

$$\sum_{i=1}^{G} \frac{\left(\frac{\beta_i}{\sigma}\right)^2}{\frac{2}{k_i} + \frac{1}{\lambda_i}} \leq 1$$

The domain is simply an ellipsoïd. This ellipsoïd has to be transformed by an orthogonal mapping to be expressed in the parameter of interest b .

$$b'Q \operatorname{diag}\left(\frac{k\lambda}{2\lambda+k}\right) Q'b \leq \sigma^2$$

3.c Least squares on modified exogenous variables.

From property 2, the GRR estimator is an OLS one computed from the explanatory variables : $\tilde{X} = XQ \operatorname{diag}\left(\frac{\lambda+k}{\lambda}\right) Q'$.

Therefore it seems natural to introduce the set of OLS estimators defined by : $\hat{b}^* = (X^{*\prime}X^*)^{-1} X^{*\prime}Y$, where $X^* = XC^*$ and C^* is regular.

Then : $\hat{b}^* = C^{*-1}(X'X)^{-1}X'Y = C^{*-1}\hat{b}$ and the risk function of $\hat{b}^*$ is :

$$R[\hat{b}^*] = (C^{*-1} - I)\, bb'(C^{*-1} - I)' + \sigma^2\, C^{*-1}(X'X)^{-1}\, C^{*-1\prime}$$

This estimator $\hat{b}^*$ is better than the OLS one in a neighbourhood of $\frac{b}{\sigma} = 0$, iff :

$$(X'X)^{-1} - C^{*-1}(X'X)^{-1}\, C^{*-1\prime} \text{ is non negative } . \qquad (7)$$

There exist some estimators of the previous form which improve the OLS procedure and which are not G.R.R. estimators.

If we only consider the admissible estimators, we know from RAO(1978) that they can be written as BAYES estimators :

$\check{b} = W[I + (X'X)W]^{-1}\, X'Y$, where W is the a priori expectation of $\frac{bb'}{\sigma^2}$.

Assuming W to be positive definite, $\hat{b}^*$ is a BAYES estimator if :

$$\hat{b}^* = \check{b} \iff \forall\, Y \quad C^{*-1}(X'X)^{-1}\, X'Y = [X'X + W^{-1}]^{-1}\, X'Y$$

$$\iff C^{*-1}(X'X)^{-1}\, X' = [X'X + W^{-1}]^{-1}\, X'$$

We obtain : $C^{*-1} = [X'X + W^{-1}]^{-1} X'X$

(8) $$C^{*} = I + (X'X)^{-1} W^{-1}$$

The associated estimator satisfies (7) ,

$$(X'X)^{-1} - [X'X + W^{-1}]^{-1} X'X[X'X + W^{-1}]^{-1} \gg 0$$

or, equivalently :

(9) $$(X'X)^{1/2} [X'X + W^{-1}]^{-1} X'X [X'X + W^{-1}]^{-1} (X'X)^{1/2} \ll I$$

If we denote by $\hat{d} = (X'X)^{1/2} \hat{b}$ the OLS estimator of $d = (X'X)^{1/2} b$ and by $\hat{d}^{*} = (X'X)^{1/2} \hat{b}^{*} = (X'X)^{1/2} [X'X + W^{-1}]^{-1} (X'X)^{1/2} \hat{d}$ the estimator corresponding to $\hat{b}^{*}$, condition (9) can be written as :

(10) $$\hat{d}^{*\prime}\hat{d}^{*} \leq \hat{d}'\hat{d}$$

Therefore the choice of such a BAYES estimator implies a shrinkage on $\hat{d}$.

Finally we have the following property :

PROPERTY 4 : In the class of estimators $\hat{b}^{*} = C^{*-1} \hat{b}$ which are admissible $(C^{*} = I + (X'X)^{-1} W^{-1})$, the GRR estimators are characterized by the symmetry of C^{*} .

Proof :

C^* is a symmetric matrix if and only if $X'X\,W = W\,X'X$, (since $X'X$ and W are both symmetric). Therefore W and $X'X$ have the same eigenvectors and W can be written :

$W = Q\,\mathrm{diag}(\omega)'\,Q'$, where the ω_i's are the eigenwalues of W. We obtain the standard form of GRR estimator by setting $k_i = \frac{1}{\omega_i}$ $i = 1,\ldots,G$.

Q.E.D.

Let us remark that in this case the k_i's are positive. In particular the G.R.R. estimators obtained with some values of k_i negative are not admissible ; however it can be better than the OLS estimator on a neighbourhood of zero.

4. CONFIDENCE REGION BASED ON G.R.R. ESTIMATOR.

In the context of ridge regression which leads to biased estimators $\tilde{b}$, we shall construct biased confidence regions centered at $\tilde{b}$.

We consider confidence regions $\tilde{W}(\alpha,B)$ defined by :

$$\tilde{W}(\alpha,B) = \{b/(\tilde{b}-b)'\Omega(\tilde{b}-b) \leq h\}$$

and $P_b\,[\tilde{W} \ni b] = P_b\,[(\tilde{b}-b)'\Omega(\tilde{b}-b) \leq h] \geq 1 - \alpha \qquad \forall\ b \in B$

where $1 - \alpha$ is the confidence level and B is a given set of values of b . Ω is a definite positive matrix, which will be chosen later.

Under the normality assumption of the disturbances, we have : $\tilde{b}(X,k) \sim N[Q'\mathrm{diag}\left(\frac{\lambda}{\lambda+k}\right)Qb \; ; \; \sigma^2 \, Q' \, \mathrm{diag}\left[\frac{\lambda}{(\lambda+k)^2}\right]Q]$

Therefore : $\frac{1}{\sigma^2}(\tilde{b} - b)' \, Q' \, \mathrm{diag}\left[\frac{(\lambda+k)^2}{\lambda}\right]Q \, (\tilde{b} - b)$ has a chi-square distribution $\chi^2(G,\delta)$ with G degrees of freedom and a non centrality parameter $\delta = \frac{1}{\sigma^2} \, b'Q'\mathrm{diag}\left(\frac{k^2}{\lambda}\right)Q \, b$.

Choosing $\Omega = \frac{1}{\sigma^2} \, Q' \, \mathrm{diag}\left[\frac{(\lambda+k)^2}{\lambda}\right] \, Q$, we see that :

$$(\tilde{b} - b)' \, \Omega(\tilde{b} - b) \sim \chi^2(G,\delta)$$

Finally a confidence region available for any $b \in B_a$

where : $$B_a = \left\{ b \; / \; \frac{b' \, Q \, \mathrm{diag}\left(\frac{k^2}{\lambda}\right)Q'b}{\sigma^2} < a \right\}$$

is given by :

$$\text{(11)} \qquad \tilde{W}[\alpha, B_a] = \{ b \; / \; \frac{1}{\sigma^2} \, (\tilde{b}-b)'Q'\mathrm{diag}\left(\frac{(\lambda+k)^2}{\lambda}\right)Q(\tilde{b}-b) \leq h \}$$

where h is the $1 - \alpha$ quantile of $\chi^2(G,a)$: $h = \chi^2_{1-\alpha}(G,a)$

This region can only be used if σ^2 is known. If σ^2 is unknown, this parameter can be estimated as usual by s^2 , the sum of squares of OLS residuals divided by $n - G$. This quantity is independent from $\hat{b}$ and also from $\tilde{b} = C^{-1}\hat{b}$. Therefore the confidence region is deduced from the previous one by replacing σ^2 by s^2 and $\chi^2_{1-\alpha}(G,a)$ by $\mathcal{F}_{1-\alpha}(G,n-G,a)$.

It remains to compare the confidence region (11) with that

$\hat{W}$ corresponding to the OLS estimator :

$$\hat{W}(\alpha) = \{b \,/\, \frac{1}{\sigma^2} (\hat{b}-b)'Q'\mathrm{diag}(\lambda)Q(\hat{b}-b) \leq \chi^2_{1-\alpha}(G)\}$$

The region $\tilde{W}$ is smaller than $\hat{W}$ if :

$$\frac{Q'\ \mathrm{diag}\left(\frac{(\lambda+k)^2}{\lambda}\right)Q}{\sigma^2\ \chi^2_{1-\alpha}(G,a)} \gg \frac{Q'\ \mathrm{diag}(\lambda)Q}{\sigma^2\ \chi^2_{1-\alpha}(G)}$$

$$\Longleftrightarrow \quad \frac{\mathrm{diag}\left(\frac{(\lambda+k)^2}{\lambda}\right)}{\chi^2_{1-\alpha}(G,a)} \gg \frac{\mathrm{diag}(\lambda)}{\chi^2_{1-\alpha}(G)}$$

$$\Longleftrightarrow \quad \chi^2_{1-\alpha}(G,a) \leq \chi^2_{1-\alpha}(G)\ \frac{(\lambda_i+k_i)^2}{\lambda_i^2} \qquad 1 = 1,\ldots,G$$

$$\Longleftrightarrow \quad \chi^2_{1-\alpha}(G,a) \leq \chi^2_{1-\alpha}(G)\ \underset{i}{\mathrm{Min}}\ \frac{(\lambda_i+k_i)^2}{\lambda_i^2}$$

There exists $a > 0$ satisfying this condition, if $\frac{(\lambda_i+k_i)^2}{\lambda_i^2} > 1$ $\forall\ i = 1,\ldots,G$. These inequalities are the same as the inequalities obtained in the comparison of the risk functions(property 3). When they are satisfied, the maximal neighbourhood for which $\tilde{W}$ is more precise than $\hat{W}$ is : $\{b \,/\, \frac{1}{\sigma^2}\ b'\ Q\ \mathrm{diag}\left(\frac{k^2}{\lambda}\right)Q'b \leq a_0\}$, where a_0 is such that :

$$\chi^2_{1-\alpha}(G,a_0) = \chi^2_{1-\alpha}(G)\ \underset{i}{\mathrm{Min}}\ \frac{(\lambda_i+k_i)^2}{\lambda_i^2}$$

5. ESTIMATION OF THE SIGNS OF THE COEFFICIENT.

It has often been suggested (VINOD (1978), HOERL-KENNARD (1970)) that ridge regression procedure was "a new hope for avoiding wrong signs".

We have already seen in subsection 2.e , that $\tilde{b}_1(X,K)$ could have the same sign as $\hat{b}_1$ for any K . Therefore, in such a case, there is no reason to use ridge regression for obtaining a better sign.

In some other cases (2.e : 5), 6)), $\tilde{b}_1$ could have the opposite sign for some K . As noted by MAC DONALD (1980), "a sign change in $\tilde{b}_1$ occurs if $\hat{b}_1$ differs in sign from the correlation coefficient between Y and X_1" ; however the sign of this correlation coefficient is not in general the same as the sign of the coefficient b_1 .

In this section, a property is established, about the linear estimators giving "most frequently" the right sign. For instance let us consider the parameter b_1 associated with the first explanatory variable X_1 . The other coefficients and the corresponding variables are denoted by b^* and X^* . A linear estimator of b_1 is of the form : $\bar{b}_1 = \alpha' Y$ with $\alpha \in IR^n$ and we are looking for α such that $\bar{b}_1$ is a solution of the problem :

$$(P) \quad \begin{cases} \underset{\alpha}{\text{Max}} \ \underset{b^*}{\text{Min}} \ P_{b_1, b^*}[\bar{b}_1 > 0] \quad , \quad \text{if } b_1 > 0 \\ \\ \underset{\alpha}{\text{Max}} \ \underset{b^*}{\text{Min}} \ P_{b_1, b^*}[\bar{b}_1 < 0] \quad , \quad \text{if } b_1 < 0 \end{cases}$$

PROPERTY 5 : Under the normality of the disturbances, the solutions of (P) are :

$\bar{b}_1 = \nu_1 \hat{b}_1$ where $\nu_1 > 0$

and $\hat{b}_1$ is the OLS estimator of b_1

Proof :

Let us consider the case $b_1 > 0$:

$$P_{b_1,b^*}[\bar{b}_1 > 0] = P_{b_1,b^*}[\alpha'(X_1 b_1 + X^* b^* + u) > 0] = \Phi\left(\frac{\alpha' X_1 b_1 + \alpha' X^* b^*}{\sigma \|\alpha\|}\right)$$

where Φ is the cumulative function of the standard normal distribution.

The minimal value of $P_{b_1,b^*}[\bar{b}_1 > 0]$ with respect to b^* is :

$$\begin{cases} \Phi[-\infty] = 0 & \text{if} \quad \alpha' X^* \neq 0 \\ \Phi\left(\dfrac{\alpha' X_1 b_1}{\sigma \|\alpha\|}\right) & \text{if} \quad \alpha' X^* = 0 \end{cases}$$

The maximization with respect to α leads to the optimization problem

$$(P_1) \quad \begin{cases} \max\limits_{\alpha} \Phi\left(\dfrac{\alpha' X_1 b_1}{\sigma \|\alpha\|}\right) \\ \text{s.t.} \quad \alpha' X^* = 0 \end{cases}$$

Since $b_1 > 0$, the solutions α of (P_1) can be written : $\alpha = \nu_1 a$,

where $\nu_1 > 0$ and a is a solution of the problem $(\tilde{P}_1)$:

$$(\tilde{P}_1) \quad \begin{cases} \max\limits_{a} \quad a' X_1 \\ \text{s.t.} \quad \begin{aligned} a' X^* &= 0 \\ a' a &= 1 \end{aligned} \end{cases}$$

or of the problem $(\tilde{\tilde{P}}_1)$:

$$(\tilde{\tilde{P}}_1) \quad \begin{cases} \max\limits_{a} \quad a' X_1 \\ \text{s.t.} \quad \begin{aligned} a' X^* &= 0 \\ a' a &\leq 1 \end{aligned} \end{cases}$$

The dual problem $(\approx{D}_1)$ of $(\approx{P}_1)$ is :

$$(\overset{\approx}{D}_1) \quad \begin{cases} \underset{a}{\text{Min}} \ a'a \\ \text{s.t.} \ \begin{array}{l} a'X^* = 0 \\ a'X_1 \geq 1 \end{array} \end{cases}$$

and is equivalent to :

$$(\tilde{D}_1) \quad \begin{cases} \underset{a}{\text{Min}} \ a'a \\ \text{s.t.} \ \begin{array}{l} a'X^* = 0 \\ a'X_1 = 1 \end{array} \end{cases}$$

This problem is exactly that associated with the determination of the BLUE of b_1 and therefore has the unique solution

$$a' = \left[X_1'(I-X^*(X'^*X^*)^{-1}X'^*)X_1\right]^{-1} X_1'(I-X^*(X'^*X^*)^{-1}X'^*)$$

Finally the solution of (P_1) are $\bar{b}_1 = \nu_1\hat{b}_1$. These solutions do not depend on b_1 . It is easily seen that the resolution of the program

$$(P) \qquad \underset{\alpha}{\text{Max}} \ \underset{b^*}{\text{Min}} \ P_{b_1,b^*}(\bar{b}_1 < 0) \qquad \text{if} \quad b_1 < 0$$

leads to the same solution.

Q.E.D.

Remark 1 : The previous approach can be applied to the other coefficients $b_2,\ldots,b_G$. The linear estimators of b giving most frequently the right signs of $b_1,b_2,\ldots,b_G$ are $\bar{b} = \text{diag}(\nu)\,\hat{b}$, with $\nu_i > 0$, $i = 1,\ldots,G$. The GRR estimator is not in general of this form.

Remark 2 : However, if we consider the principal parameters $\beta_1,\ldots,\beta_G$, we have $\tilde{\beta}(X,k) = \text{diag}\left(\frac{\lambda}{\lambda+k}\right)\hat{\beta}$. Therefore, if $\lambda_i+k_i > 0 \quad i = 1,\ldots,G$, $\tilde{\beta}(X,k)$ gives most frequently the right signs of $\beta_1,\ldots,\beta_G$, but not more frequently than $\hat{\beta}$.

Remark 3 : Finally if we impose to the GRR estimator to simultaneously give the right signs of the β_i's and to improve the precision, we must have :

$\lambda_i+k_i > 0$ and $\{k_i > 0$ or $k_i < -2\lambda_i\}$ $\qquad i = 1,\ldots,G$

This implies the usual restrictions $k_i > 0 \quad i = 1,\ldots,G$

6. RIDGE REGRESSION AND QUASI-COLLINEARITY.

The sufficient condition, given by THEOBALD (1974) : $0 < K < \frac{2\sigma^2}{\|b\|^2}$, for the GRR estimator to be better than the OLS one, does not depend on the explanatory variables and then does not depend on the degree of multicollinearity.

However the relation with the collinearity problems appears, when we consider the exact conditions. For instance, if $G = 2$, the region of improvement of OLS, derived in 3.b :

$$\{\beta_1,\beta_2,\sigma \;/\; \frac{(\frac{\beta_1}{\sigma})^2}{\frac{1}{k_1}+\frac{2}{\lambda_1}} + \frac{(\frac{\beta_2}{\sigma})^2}{\frac{2}{k_2}+\frac{1}{\lambda_2}} < 1\}$$

increases, when λ_1 or λ_2 decreases.

The problem of collinearity can be examined precisely in the case $G = 2$. In this case a measure of the degree of collinearity is the angle between X_1 and X_2 .
It can be interesting to see if, as suggested by SWINDEL (1974), the variables $\tilde{X}_1$ and $\tilde{X}_2$ associated with the GRR estimator are less collinear than X_1 and X_2 are.

The orthogonal matrix Q can be written as :

$$Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

and $X'X$ is given by :

$$X'X = Q\,\mathrm{diag}(\lambda)\,Q' = \begin{bmatrix} \lambda_1\cos^2\theta + \lambda_2\sin^2\theta & (\lambda_1-\lambda_2)\sin\theta\cos\theta \\ (\lambda_1-\lambda_2)\sin\theta\cos\theta & \lambda_1\sin^2\theta + \lambda_2\cos^2\theta \end{bmatrix}$$

where $\lambda_1 \geq \lambda_2$.

The transformed variables $\tilde{X}_1$, $\tilde{X}_2$ are such that :

$$\tilde{X}'\tilde{X} = Q\,\mathrm{diag}(\alpha)\,Q' \qquad \text{with} \qquad \alpha_i = \frac{(\lambda_i+k_i)^2}{\lambda_i}$$

The acute angle corresponding to $\tilde{X}_1$, $\tilde{X}_2$ is greater than the one corresponding to X_1 , X_2 iff :

$$\frac{[(\lambda_1-\lambda_2)\sin\theta\cos\theta]^2}{(\lambda_1\cos^2\theta+\lambda_2\sin^2\theta)(\lambda_1\sin^2\theta+\lambda_2\cos^2\theta)} \geq \frac{[(\alpha_1-\alpha_2)\sin\theta\cos\theta]^2}{(\alpha_1\cos^2\theta+\alpha_2\sin^2\theta)(\alpha_1\sin^2\theta+\alpha_2\cos^2\theta)}$$

This condition can be simplified in :

$$(\lambda_1^2+\lambda_2^2)\ \alpha_1\ \alpha_2 \ \geq\ (\alpha_1^2+\alpha_2^2)\ \lambda_1\ \lambda_2$$

$$\Longleftrightarrow \quad (12) \quad 0 \ \geq\ (\lambda_2\ \alpha_1 - \lambda_1\ \alpha_2)\ (\lambda_1\ \alpha_1 - \lambda_2\ \alpha_2)$$

This inequality is satisfied iff :

$$\frac{\lambda_2}{\lambda_1} \leq \frac{\alpha_1}{\alpha_2} \leq \frac{\lambda_1}{\lambda_2}$$

or, equivalently :

$$\left(\frac{\lambda_1+k_1}{\lambda_1}\right)^2 \leq \left(\frac{\lambda_2+k_2}{\lambda_2}\right)^2 \quad \text{and} \quad (\lambda_1+k_1)^2 \geq (\lambda_2+k_2)^2$$

If λ_1+k_2 and λ_2+k_2 are positive, these conditions become :

$$\frac{\lambda_1+k_1}{\lambda_1} \leq \frac{\lambda_2+k_2}{\lambda_2} \quad \text{and} \quad \lambda_1+k_1 \geq \lambda_2+k_2 \ .$$

The first condition is expressed in term of the shrinkage coefficients $\frac{\lambda_i}{\lambda_i+k_i}$; it means that,smaller is the eigenvalue,greater is the shrinkage. This is the condition of "declining deltas" given in VINOD (1978).

The second condition means that the principal factors remain in the same order after transformation.

These conditions are in particular satisfied for the ORR estimator, with $K > 0$.

Remark 4 : If is clear that the previous conditions are not related to the conditions to improve the precision of OLS estimator $k_i > 0$ or $k_i < -2\lambda_i$.

7. CONCLUDING REMARKS.

Several properties lead to consider ridge regression as an alternative to least squares estimation in the multiple linear regression model.
The justifications usually given for the use of this procedure are the possibilities :

a) to improve the OLS estimator for some values of the parameters ;
b) to include some a priori information on the parameters ;
c) to solve the difficulties introduced by the quasi-collinearity.

The previous study allows us to discuss these justifications :

a) For some values of the parameters, the ridge estimator is better than the OLS one for the risk functions (THEOBALD (1974)) and also for the confidence regions. However the improvement may also be obtained with negative biasing factors k_i .
The ridge estimator can be interpreted as an OLS one based on transformed explanatory variables. Using other variables than the initial ones may lead to a better estimator.
b) If all the parameters are positive, the GRR estimators are interpretable as BAYES estimators, associated with a particular a priori

distribution of the parameters. This distribution is zero mean (LINDLEY - SMITH (1972)). This property shows that ridge regression should be used if the parameters are closed to zero ; this is an important limitation, which is not in general compatible with a priori ideas on the parameters.

An interpretation of the same kind has been presented by various authors (HOERL-KENNARD (1970a), KMENTA-LIN (1980), CRAIG-VAN NOSTRAND (1980), GALARNEAU-GIBBONS (1981)...). According to this interpretation, the GRR estimator is a constrained OLS estimator. It is solution of the problem :

$$\min_b \|y - Xb\|^2 \qquad \|b\|^2 \leq r^2$$

However this interpretation seems to be not correct. The usual proof consists in the introduction of a LAGRANGE multiplier λ and in the minimization of :

$$\|y - Xb\|^2 - \lambda(\|b\|^2 - r^2)$$

The first order condition gives : $b^* = (X'X + \lambda I)^{-1}X'Y$

where the multiplier satisfies :

$$Y'X(X'X + \lambda I)^{-2}X'Y = r^2$$

Therefore the biasing factor λ <u>depends on the observations</u> Y and is a random variable. The statistical properties of this estimator are different from the GRR estimator's properties.

c) The opinion that ridge regression leads to estimators giving the right sign more frequently than the OLS estimator is not correct. We proved that the OLS estimator is optimal for this problem. The ridge estimators only appear optimal for the determination of the sign of the principal parameters. Unfortunatly these parameters are often difficult to interprete from an economic point of view.

d) The utilisation of ridge regression for solving the quasi-collinearity problems has abondantly been discussed in the literature (see for instance SMITH-CAMPBELL (1980)). We have seen, in the two parameters case, that choosing k , such that the transformed variables are less collinear than the initial ones are, has no connection with the improvement of OLS.

If we impose conditions simultaneously for improving the OLS (for some values of b) for questing the right sign and for reducing the quasi-collinearity, we find the usual constraints : $k_i > 0$ (right sign, improvement), $\frac{\lambda_i+k_i}{\lambda_i}$ decreasing, λ_i+k_i increasing (reduction of quasi-collinearity).

These remarks show the difficulties of using ridge regression even in the simple case of multiple linear regression. Then it seems too early to propose direct generalizations of this procedure to more complex situations.

Finally let us remark that the previous study does not apply to the cases where the biasing parameters depend on the observations on the endogenous variable. In such a case the statistical properties have to be reexamined (see for instance JAMES-STEIN (1961)).

REFERENCES

CRAIG and R, VAN NOSTRAND (1980) : "Comment". Journal of the American Statistical Association, 369,92-94.

GALARNEAU and D, GIBBONS (1981) : "A simulation study of some ridge estimators", Journal of the American Statistical Association, 373, 131-139.

JAMES, W and C, STEIN (1961) : "Estimation with quadratic loss". Proceedings of the fourth Berkeley Symposium, Vol 1, 361-379.

HOERL, A and W, KENNARD (1970a) : "Ridge regression : biased estimation for non orthogonal problems", Technometrics 12, 55-68.

HOERL, A and W. KENNARD (1970b) : "Ridge regression : Application to non orthogonal problems, Technometrics 12, 69- .

KMENTA, J and K, LIN (1980) : "Some new results on ridge regression estimation", Discussion paper, University of Michigan.

LINDLEY, D and A, SMITH (1972) : "Bayes estimates for the linear model", Journal of the Royal Statistical Society, B, 30, 1-41.

MAC DONALD, G (1980) : "Some algebraic properties of ridge coefficients", Journal of the Royal Statistical Society, B, 42, 31-32.

MARQUARDT, D (1970) : "Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation", Technometrics 12, 591.

RAO, R (1976) : "Estimation of parameters in a linear model", Annals of Statistics, 4, 1023-1038.

SMITH, G and F., CAMPBELL (1980) : "A critique of some ridge regression methods", Journal of the American Statistical Association, 369, 74.

SWINDEL, B (1974) : "Instability of regression coefficient illustrated", The American Statistician 28, 63-65.

THEOBALD, C (1974) : "Generalization of mean square error applied to ridge regression", Journal of the Royal Statistical Society B, 36, 103-106.

VINOD, D (1978) : "A survey of ridge regression and related techniques for improvements over ordinary least squares", The Review of Economics and Statistics 60, 121-131.