A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 132, 1993

*Władysław Milo**

ON THE USEFULNESS OF REGULARIZATION IDEAS OF ESTIMATION: THE LINEAR MODEL CASE

Abstract. In the paper we present an analysis of negative effects of ill-conditioning for the performance of LSE. These results will be observed through the behaviour of LSE's variance, MSE, sample standard deviation, sample multiple correlation coefficient, F and t-statistics. We also include some results on ill-conditioning effects induced by data centering, weighting. To overcome these negative effects we propose new versions of regularization criteria for the linear model case. The resultant regularising estimators are consistent and asymptotically normal.

Key words: linear models, regularization, ill-conditioning, regularizing estimators.

1. INTRODUCTION

In econometrics and statistical literature one can find discussion about the existence and effects of

i) multicollinearity of the columns of matrix x,
ii) almost-multicollinearity of the columns of matrix x,
iii) bad-conditioning of matrix x,
iv) high correlation of explanatory variables $x_1, \ldots, x_k$.

These four concepts (i) - (iv) are intertwining in the scopes of their meaning. These readers who trace advancement in numerical and sta-

* Lecturer at the Institute of Econometrics and Statistics of the University of Łódź.

tistical data analysis on the one side and in applied econometrics and statistics on the other, feel strongly the necessity to fix the strict meaning of (i) - (iv) and to carry on the analysis of their relationships. One of the practical reasons of this necessity is the need to create good diagnostic methods and programming diagnostic packages for identification of the existence of (i) - (iv), its sources, and consequences.

It is known that such diagnostic tools can not be created without diagnostic acts concerning such phenomena as outliers, missing values, autocorrelation, influential observations, stability of the parameters and models.

In this paper we concentrate our analysis on bad conditioning of the data matrix with some points of reference to multicollinearity, almost-multicollinearity and strong correlation.

Negative effects of (iii) will be discussed through the analysis of the effects of centering, weighting, standardization.

In § 2 we try to define these concepts and give some notes on their complexity.

In § 3 we discuss relationship between almost-multicollinearity and bad-conditioning.

In § 4 there is a discussion of negative and positive effects of centering, weighting, standardization of x.

As a device of overcoming negative effects of bad conditioning we propose regularising estimators.

In § 5 we propose some regularising ideas and estimators.

In § 6 a short discussion on asymptotic properties of regularising estimators is given.

2. ON INTERWINING CONCEPTS OF RELATIONSHIP BETWEEN EXPLANATORY VARIABLES

The term "multicollinearity" [see: K e n d a l l, B u c k l a n d (1971)] denotes linear dependence between the predictor (sometimes called: independent) variables. In algebraic terms it denotes the linear dependence of columns in the data matrix x of dimension $n \times k$, i.e. the matrix with rank $(x) = k_o < k$. In the case of the linear model describing the random $n \times 1$ vector Y.

$$Y = x\beta + W, \quad k_o < k, \quad P_Y = N_Y(x\beta, \sigma^2 I) \tag{1.1}$$

where the relation $k_o < k$ or, equivalently, the relation $x\delta = 0$, $\delta \neq 0$ characterizes "multicollinearity" of the columns of x, and "$P_Y = N_Y(x\beta, \sigma^2 I)$ "denotes" the normal distribution of the random vector Y with mean $x\beta$ and dispersion $\sigma^2 I$". By (1.1) OLSE (Ordinary Least Squares Estimator) does not exist. Other estimates are no-unique and their sample standard deviations go to infinity. These are the most notorious negative effects of assumptions (i) and (1.1) on the OLSE.

Other known negative effects are: biasedness of estimators, predictors, residuals, singularity but still normality of distributions of generalized inverse estimators, predictors and residuals, as well as, that the sum of squares of predictors and residuals, based, on Moore-Penrose inverse, is X^2-distributed.

There are other interpretations of multicollinearity. One can find them in the works of [J o h n s t o n (1962), S i l v e y (1969), G u n s t (1983), M a s o n, G u n s t, W e b s t e r (1975), F a r r a r, G l a u b e r (1967), H a r v e y (1977), C h a t t e r j e e, P r i c e (1977)]. For example J. Johnston A. Harvey say that (i) occurs if two or more explanatory variables are highly correlated, i.e. if (iv) occurs. They do not distinguish the model

$$Y = x\beta + W, \quad k_o = k, \quad \text{corr } (x) \text{ high}, \quad P_Y = N_Y(x\beta, \sigma^2 I) \qquad (1.2)$$

where x:n x k is a real matrix and corr (x) has purely non-probabilistic descriptive meaning from the model

$$Y = x\beta + W, \quad k_o = k, \quad \text{corr } (x) \text{ high}, \quad P_Y = N_Y(x\beta, \sigma^2 I) \qquad (1.2a)$$

where the matrix x is a random n x k matrix, and corr (x) has normal meaning. It is known that corr $(x) = D_x^{-1/2} x' C x D_x^{-1/2}$ is a descriptive correlation matrix since it is defined, through $C = I - n^{-1}11'$ and $D_x = \frac{1}{n-1}$ diag $(x'_{.1}Cx_{.1}, \ldots, x'_{.k}Cx_{.k})$.

However in the case of (1.2a) corr $(x) = \Delta_x^{-1/2} X \Delta_x^{-1/2}$, where $\Delta_x = \text{diag}\,(\sigma^2_{x_1}, \ldots, \sigma^2_{x_k})$, $\sigma^2_{x_i} = \text{var } x_i$, $i = 1, \ldots, k$, $X \equiv x'C_\mu x$.

It is obvious that the statistical meanings of (1.2) and (1.2a) are different.

With the expression "multicollinearity" we associate the ex-

pression "near or almost multicollinearity of columns of x". It is [see: G u n s t (1983)] characterized by

$$x\delta = d, \quad \|d\| \leq \eta \|\delta\|, \quad d, \quad \delta \neq 0 \iff k_o = k \tag{1.3}$$

The linear model describing (y, x) with the almost-multicollinear columns of x has the form

$$Y = x\beta + W, \; k_o = k, \; x\delta = d, \quad \|d\| \leq \eta \|\sigma\|, \quad P_Y = N_Y(x\beta, \sigma^2 I) \tag{1.3a}$$

In defining bad-conditioning of x and a corresponding model for bad-conditioned data we use popular index of bad-conditioning.

$$\nu = \nu_x = \lambda_k^{1/2}\lambda_1^{-1/2} \tag{1.4}$$

where $\lambda_k^{1/2} = \lambda_{max}^{1/2}(\chi)$, $\lambda_1^{1/2} = \lambda_{min}^{1/2}(\chi)$ are singular values of the matrix $\chi = x'x$, appropriately, the largest and the smallest and λ_k, λ_1 are the largest and the smallest eigenvalues of x'x.

It will be said that the matrix x is bad-conditioned if

$$\nu^2 > \nu_*^2 \tag{1.5}$$

where ν_*^2 is the threshold value of ν^2 distinguishing bad- and good-conditioned matrices x.

Therefore our model describing bad-conditioned results of observations is of the form

$$Y = x\beta + W, \quad k_o = k, \quad \nu^2 > \nu_*^2, \quad P_Y = N_Y(x\beta, \sigma^2 I) \tag{1.6}$$

By (1.6) it is seen that the statistical meaning of bad-conditioning and correlation in the context of (1.6) and (1.2) - (1.2a) is different. It can happen that strong correlation in x can coincide with strong bad-conditioning of x. It can also happen that strong correlation coincides with the small bad-conditioning. It can also happen that for a given matrix x all phenomena (ii)-(iv) occur simultaneously. Up to now there has not been clear cut separation devices for them. Some qualitative interrelationships will be shown in § 3.

3. ON RELATIONSHIPS BETWEEN EXPLANATORY VARIABLES

Remember that there are distinguished four kinds (see (i) (iv) in § 2) of relationships between explanatory variables. Let us start with (i). It is characterized by

$$x\delta = \sum_{j=1}^{k} x_{.j}\delta_j = 0, \quad \exists_j: \delta_j \neq 0, \quad j = 1, \ldots, k \tag{3.1}$$

or equivalently as $k_o < k$.

In (3.1) the vector δ is not too informative. In order to change its qualitative definitional role we replace [see: G u n s t (1983)] the vector δ with the eigen vector $v_{.1}$ corresponding to the eigenvalue $\lambda_1 = 0$. Hence, (3.1) takes more diagnostic form

$$xV_{.1} = 0, \quad V_{.1} \neq 0 \tag{3.1a}$$

The relation (3.1a) enables us to detect the structure of collinear relationship between the columns of data matrix x. Note that if the parameter vector β coincides with $V_{.1}$, then

$$Y = xV_{.1} + W = W, \quad \lambda_1 = 0 \tag{3.2}$$

The point $\beta = V_{.1}$ is, therefore, a pathological point of parameter space. It annihilates the signal $x\beta$. Such atrophy of signal evidently spoils specification efforts and should be taken into considerations in constructing nested and non-nested types of testing procedures. It is the break-down point for any sensible linear specification. So, for $k_o < k$ and $\beta = V_{.1}$, it is useless to estimate β and to do testing. Less hopeless but still very serious situation we have in the close neighbourhood of this break-down point $\beta = V_{.1}$. More details can be found in [M i l o (1989)].

For $\beta \neq V_{.1}$ and β and $V_{.1}$, sufficiently far away, we can sensibly estimate β and test it in spite of (3.1a). The relation between multicollinearity and almost-multicollinearity can be seen from the discussion of (1.3a). If $d = 0$ and $k_o < k$, then (1.3a) becomes (1.1). Hence, the last two modified forms of assumptions are linkage forms between (i) and (ii). In this respect some more comments illuminate the mater: - in (1.3a) the quantity $\bar{\eta} = \frac{\|d\|}{\|\delta\|}$

can measure the degree of almost multicollinearity. Another option is the quantity $\bar{\eta} = \frac{\|x\delta - \hat{d}\|}{\|x\delta\|} \leq 1$, where $\hat{d}$ approximates d. They measure closeness to dependence or distance from independence, $k_o < k$ (or (3.1)) is not equivalent to non-orthogonality since even for $k_o = k$ (independent system x of vectors) the system x does not need to be orthogonal, i.e. there is always

$$x_{.1} \perp x_{.j}, \quad i \neq j = 1, \ldots, k \rightarrow x\delta = 0, \quad \delta \neq 0 \tag{3.3}$$

but

$$(x\delta = 0, \quad \delta \neq 0) \Longrightarrow x_{.i} \perp x_{.j}, \quad i \neq j = 1, \ldots, k \tag{3.4}$$

- from (1.1) and (1.3a) it can be seen that (i) and (ii) are codefining linear models and influence properties of statistics defined on the elements of these models. Thanks to (1.6) one can see that (iii) codefines the linear model (1.6) describing bad-conditioned results of observations. Therefore (iii) has partly numerical and partly statistical role and meaning. In order to depict relationship between (ii), (iii) and (i) we will use the idea of singular value of decomposition SVD [see: M a g n u s-N e u d e c k e r (1988)]. Due to SVD we can write $x = \sum_{i=1}^{k} \lambda_i^{1/2} U_{.i} V_{.i}$, where $\{U_{.i}, V_{.i}\}$ are eigenvectors corresponding to the eigenvalues $\{\lambda_i\}$ of xx' and $x'x$. We rewrite $x\delta = 0$ as

$$\sum_{i=1}^{k_0} \lambda_i^{1/2} U_{.i} V'_{.i} \delta = 0 \tag{3.5}$$

or after dividing (3.5) by $\lambda_1^{1/2}$

$$u_{.1} \frac{V_1' \delta}{\|\delta\|} + \ldots + \sqrt{\nu}\, u_{.k} \frac{V_{.k}\delta}{\|\delta\|} = 0 \tag{3.5a}$$

This equation relates bad-conditioning (iii) and multicollinearity (i). To connect near-multicollinearity (ii) with bad-conditioning (iii) we use SVD and (1.3) to obtain

$$u_{.1} \frac{V'_{.i}\delta}{\|\delta\|} + \ldots + \sqrt{\nu}\, u_{.k} \frac{V'_{.i}\delta}{\|\delta\|} = \frac{1}{\sqrt{\lambda_1}\,\|\delta\|} d \tag{3.6}$$

Squaring the last equation gives more diagnostic equation

$$\nu^2 \frac{(V'_{.k}\delta)^2}{\|\delta\|} + \sum_{j=1}^{k-1} \frac{(V'_{.j}\delta)^2}{\|\delta\|^2} = \lambda_1^{-1}\eta^{-2} \qquad (3.6a)$$

Special cases of (3.6a) are

a) if $\delta = v_{.k}$ then $\nu^2 = \lambda_1^{-1}\eta^{-2}$ or $\eta^{-2} = \lambda_k$

b) if $\delta = v_{.j}$, $j = 1, \ldots, k-1$, then $\|d\|^2 = \lambda_1$

c) if $\delta = \beta$, then

$$u_{.1} \frac{V'_1\beta}{\|\beta\|} + \ldots + \nu u_{.k} \frac{V'_{.k}\beta}{\|\beta\|} = 0$$

$$u_{.1} \frac{V'_{.1}\beta}{\|\beta\|} + \ldots + \nu u_{.k} \frac{V'_{.k}\beta}{\|\beta\|} = \frac{1}{\sqrt{\lambda_1}\|\beta\|} d$$

$$\nu^2 \frac{(V'_{.k}\beta)^2}{\|\beta\|^2} + \sum_{j=1}^{k-1} \frac{(V'_{.j}\beta)^2}{\|\beta\|^2} = \frac{\|d\|^2}{\|\beta\|^2} \frac{1}{\lambda_1}$$

d) If $\delta = \beta = V_{.k}$, then $\nu = \lambda_1^{-1}\eta^{-2}$ and $\eta^{-2} = \lambda_k$.

In order to relate strong correlation (iv) with bad-conditioning (iii) one needs to consider two cases: (.) x is non-random real matrix without regard to the situation when x is a sample value of random vector $X = (X_1, \ldots, X_k)$, (..) X is a random matrix with n rows and k columns.

Suppose that in the case of (.) we transform x into $x_* = CxD_x^{-1/2}$.

For standardized matrix x_* we obtain counterparts of

$$x_*\delta_* = 0 \quad \text{multicollinearity} \qquad (3.7)$$

$$x_*\delta_* = d_*, \; \|d_*\| \leq \eta_* \|\delta_*\|, \; d_*, \; \delta_* \neq 0 \quad \text{almost multicollinearity} \qquad (3.7a)$$

$$\nu_{x_*} > \nu_*, \quad \nu_{x_*} = \frac{\lambda_k(x_*)}{\lambda_1(x_*)} = \frac{\lambda_{*k}}{\lambda_{*1}} \quad \text{bad-conditioning of } x_* \qquad (3.7b)$$

Due to SVD we can rewrite (3.7), (3.7a) as

$$u_{*1} = \frac{V'_{*1}\delta_*}{\|\delta_*\|} + \ldots + \nu_{x_*} u_{*k} \frac{V'_{*k}\delta_*}{\|\delta_*\|} = 0 \qquad (3.8)$$

or

$$u_{*1}\frac{v'_{*1}\delta_*}{\|\delta_*\|}+\ldots+\nu_{x_*}u_{*k}\frac{v'_{*k}\delta_*}{\|\delta_*\|}=\frac{1}{\sqrt{\lambda_{*1}}\|\delta_*\|}d_* \qquad (3.8a)$$

the above relations in obvious way relate bad-conditioning of the matrix x_* with multicollinearity and almost-multicollinearity of the columns of x_*. Since $x_* = \text{corr } x$, therefore we connected in (3.8) both multicollinearity, bad-conditioning and, if the elements of corr (x) are greater than ±0,7, strong correlations.

In the case (..) we have

$$X_* = C_{\xi X} X D_X^{-1/2} \quad \text{where} \quad C_{\xi X} = I - \frac{1}{n}1(\xi X),$$

$$D_X^{1/2} = \text{diag}\,(\sigma_{X_1}, \ldots, \sigma_{X_k}), \quad X_* = X_*'X_*.$$

Therefore, our characterizations of multicollinearity, almost-multicollinearity, bad-conditioning are as follows

$$X_*\delta_* = 0 \qquad (3.9)$$

$$X_*\delta_* = d_*, \quad \|d_*\| \le \eta_*\|\delta_*\|, \quad d_*, \ \delta_* \neq 0 \qquad (3.9a)$$

$$\nu_{X_*} > \nu_* \qquad (3.9b)$$

$$\text{corr } X = X_*'X_* \text{ with elements greater than } \pm 0{,}7 \qquad (3.9c)$$

Using SVD, as above, it is easy to formulate forms of (3.9), (3.9a), (3.9b) that clearly combine four above types of relationships between explanatory variables.

4. ON CONSEQUENCES OF BAD-CONDITIONING

In §§ 1-3 we occasionally mentioned effects of existence of bad-conditioning for the characterization of multicollinearity, almost-multicollinearity, strong-correlation as well as properties of statistics defined for x in the context of linear models. Now we will use formulas derived in Appendix A. As it is easily seen from A1-A21:

- the values of Y, $Y'Y$, $\bar{Y}$, Y^{-2}, $Y'CY$ generally depend on the values of bad-conditioning index ν also called "condition number ν of matrix x" (see: A10-A14),
- the values of Y, $Y'Y$, $\bar{Y}$, $Y'CY$ do not depend on ν if, for each $i = 1, \ldots, k$, $v_{\cdot i} \perp \beta$, $\beta \neq 0$ (see: A18).

- residuals $E = MY$, sum of squares $E'E = Y'MY$ of residuals, the sample variance $\hat{\sigma}^2 = (n-k)^{-1}E'E$ do not depend on the level ν of x, (see: A2, A16, A16a),
- by dependence of $Y'CY$ (see: A11-A14) on the level of ν the sample correlation coefficient $R^2 = R^2(\nu) = \text{corr}\,(Y, B'X) = 1 - (Y'CY)^{-1}Y'MY$, $X = (X_1, \ldots, X_k)'$ also depends on ν.

It is easy to find that under the conditions:

$\lambda_1 \to 0$, $\lambda_i > 0$ relatively large, $c_{1i} = V'_{.i}\beta(V'_{.i}\beta - nU_i^{-2}) > 0$ $c_{2i} = V'_{.i}\beta(U'_{.i}W - nU_i^{-2}) > 0$, $i = 2, \ldots, k$, $c = W'MW > 0$, we have $\lim_{\lambda_1 \to 0} (Y'CY)^{-1}E'E \approx 0$, and $\lim_{\lambda_1 \to 0} R^2(\nu) \approx 1$. Otherwise, under the high values of ν there would be a tendency to overevaluate indications given by R^2. Other effects of bad-conditioning are as follows:

- both t-statistic as well as $F = (1 - R^2)^{-1} R^2(n - k)$ statistic used in testing significance of model parameters depend on ν,
- the Durbin-Watson statistic, by its definition, does not depend on the condition number ν of x,
- Durbin-Watson, Dent as well as Theil-Nagar estimators of autocorrelation coefficient do not depend on ν,
- the internally and externally studentized residuals [see: C o o k-W e i s b e r g (1982)] do not depend on the condition number of x,
- recursive residuals do depend on $(n - k)$ condition numbers calculated for matrices $x'_{(i)}x_{(j)}$, $i = k, \ldots, n$, where $x_{(i)}$ is a submatrix $((i) \times k)$ of matrix x,
- values of the empirical influence curve $EIC_i = nX^{-1}x_{i.}E_i$ depend on ν,
- the sample influence curve SIC_i defined, in Cook-Weisberg book, as $SIC_i = (n - 1)(1 - M_{ii})^{-1}\chi^{-1}x_{1.}E_i$ depends on the condition number,
- the CUSUM test and the fluctuation test statistics for their definition [see: K r ä m e r, S o n n e r b e r g e r (1986)] do depend on the condition numbers of some submatrices of x'x,

- an instrumental variable estimator (for definition and interesting numerical and statistical properties see discussions given by [F a r e b r o t h e r (1988), K i v i e t (1987), P o l l o c k (1979)], does depend on the condition number x.

The above mentioned negative and positive effects of bad-conditioning existence for the properties of statistics defined for the elements of linear model in a clear way show us when and why regularizing estimation methods are useful in the sense of increasing stability of estimators with respect to bad-conditioning. Such estimators are derived from certain regularizing functionals. Some of them we will show in the next paragraph. The negative effects that were listed above are not the only ones. Many of them would extend to at least simple simultaneous linear models, special types of single equations models. We did not touch problems of the size and sensitivity of negative effects on the level of bad--conditioning. They can be tackled among others, by the use of matrix differential calculus tools (for their exposition see, for instance, [M a g n u s, N e u d e c k e r (1988)].

It has to be remembered that in the above notes the bad-conditioning was only confirmed as existing in x and said to be harmful. There were no arguments why it exists. In general, it is very difficult to give them on the grounds of numerical or statistical analysis.

There are, however, situations when we can explicitly state why bad-conditioning arises. They take place, under certain conditions as the result of centering, weighting, standardizing the data matrix x.

In the case of centering, it is easy to check that if $\nu_x = 1$, then $\nu_{x,Cx} > 1$ if $\nu_x > \bar{U}_1^{-2} U_k^{-2}$, where U_i^{-2} is the square of average of the elements of the eigenvector $U_{.i}$. If the last inequality holds, then centering operation introduces bad-conditioning. Otherwise, it does not.

In the case of weighting on the RHS of x_k (called variables scaling) with the weight matrix $W = \mathrm{diag}\{w_{jj}\}_1$, $0 < w_{jj} \leq 1$, due to Fan Ky theorem [see: M a g n u s, N e u d e c k e r (1988)], we have: if $\nu_x = 1$ then $\nu_{WXW} > 1$ if $(w^2_{kk}-1)(w^2_{11}-1)^{-1} > \chi_{11}\chi_{kk}^{-1}$, where w_{kk}, w_{11} are the largest and the smallest weights

and χ_{11} and χ_{kk} are the first and last element of the matrix x'x. It means that if the before scaling matrix x had the ideal $\nu_x = 1$ than the obtained matrix XW will have an increased condition number $\nu_{WXW} > 1$.

In the case of standardizing operation we transform a matrix x into $x_* = CxD_x^{-1/2}$. Its cross product's form $\chi_* = x_*'x_* = D_x^{-1/2}x'CxD_x^{-1/2}$ is called descriptive dispersion matrix if x is a sample value of random vector $X = (X_1, X_2, \ldots, X_k)'$. It can be found that if

$$x'_{.j}Cx_{.j} > \frac{n-1}{n - (\frac{2n-1}{n-1}) - \frac{n^2}{\lambda_1}\bar{U}_k}, \quad \text{then} \quad \nu_{x_*} > 1,$$

$$\nu_x = 1, \quad x'_{.j}Cx_{.j} = \min_j (x_{.j}Cx_{.j})$$

or

$$\nu_{x_*} > 1 \text{ if } \nu_x = 1, \; x'_{.j}Cx_{.j} > \frac{n-1}{(n - \frac{1}{n-1}) - n^2\lambda_1^{-1}\bar{U}_k}.$$

Equivalent conditions of negative effects are

$$\left.\begin{array}{l} \bar{U}_k > 0, \; n > 1+x'_{.j}Cx_{.j} \\ \bar{U}_k < 0, \; n < 1+x'_{.j}Cx_{.j} \end{array}\right\}, \quad \lambda_1 \geq \frac{4(n-1)^2\bar{U}_k(n-1-x'_{.j}Cx_{.j})}{(n-2)^2 x'_{.j}Cx_{.j}}$$

5. REGULARIZING ESTIMATORS

As it is well known [see the works of H o e r l, K e n n a r d (1970), V i n o d (1978), F a r e b r o t h e r (1978), T r e n k l e r (1985), V i n o d, U l l a h (1981)] bad-conditioning produces instability of l.s. estimators in the case of linear model. There are many ways to reduce this instability with respect to small changes in the elements of x'x or x'y. One popular option is to derive regularizing ridge type estimators by minimizing

$$\phi_{HK}(B) = \phi_o(B) + \gamma\beta'\beta$$

The obtained, parametric in γ, family of estimators has the form

$$B_{HK}(\gamma) = (x'x + \gamma I)^{-1}x'Y$$

and its empirical counterpart is of the form

$$B_{HK}(\hat{\gamma}) = (x'x + \hat{\gamma}I)^{-1}x'Y, \quad \hat{\gamma} = (B'B)^{-1}E'E\,(n-k)^{-1}.$$

Note that B_{HK} belongs to a set of families of regularizing estimators. In this paper we introduce new families of regularizing estimators. Their detailed discussion is given in [M i l o (1988, 1989)]. Now we present new criteria functions. They are as follows

$$\phi_{R1}(\gamma_1, \nu) = \phi_o(\beta) + \nu_{\underset{\sim}{x}}(\gamma_1, \nu), \quad \gamma_1 = \frac{\beta'\beta}{\sigma^2} \tag{5.1}$$

where $\nu_{\underset{\sim}{x}}(\gamma_1, \nu) = \dfrac{\sigma^2+\lambda_k\beta'\beta}{\sigma^2+\lambda_1\beta'\beta} = \dfrac{1 + \lambda_k\gamma_1}{1 + \lambda_1\gamma_1}$.

$$\phi_{R2}(\gamma_2, \nu) = \phi_o(\beta) + \nu_{\underset{\sim}{x}}(\gamma_2, \nu), \quad \gamma_2 = \frac{\beta'\chi\beta}{\sigma^2} \tag{5.2}$$

$$\nu_x(\gamma_2, \nu) = \frac{\sigma^2 + \lambda_k\beta'\chi\beta}{\sigma^2 + \lambda_1\beta'\chi\beta} = \frac{1 + \nu\lambda_1\gamma_2}{1 + \lambda_1\gamma_2}$$

$$\phi_{R3}(\nu) = (1 - \nu^{-1})\phi_o(\beta) + \nu^{-1}\,\|V_{.1} - \beta\|^2 \tag{5.3}$$

where $V_{.1}$ is the eigenvector corresponding to the least eigenvalue of $x'x$.

These new criteria functions can be motivated by the following reasoning.

Suppose that the least squares system of normal equations

$$\chi B = x'Y$$

is unstable bad-conditioned with respect to the small changes of the elements of $x'x$ or $x'Y$. In order to increase the stability of this system we propose to regularize it, i.e.

$$(x'x + \frac{\sigma^2}{\beta'\beta} I)\beta = x'Y$$

or putting it into another useful notation

$$(x'x + \gamma_1^{-1}I)\beta = x'Y.$$

By linear algebra facts the condition number $\nu_{\underset{\sim}{x}}$ for the regularized matrix $\underset{\sim}{\chi} \equiv (x'x + \gamma_1^{-1}I)$ equals

$$\nu_{\underset{\sim}{x}} = \frac{\underset{\sim}{\lambda}_k}{\underset{\sim}{\lambda}_1}, \quad \text{where} \quad \underset{\sim}{\lambda}_k = \lambda_k + \gamma_1^{-1}, \quad \underset{\sim}{\lambda}_1 = \lambda_1 + \gamma_1^{-1}$$

We postulate that this condition number be small as possible. It is, by definition (through γ_1 a function of β, and σ^2). The regularizing part of estimation quality functional is here equal $\nu_x(\gamma_1, \nu) = \nu_{\underset{\sim}{x}}$. By minimizing ϕ_{R1} with respect to β we also minimize $\nu_{\underset{\sim}{x}}$ which was postulated. By the rules of differential calculus we obtain parametric family estimators

$$B_{R1} = (x'x + \frac{\nu - 1}{\sigma^2\lambda_1^{-1}(1 + \lambda_1\gamma_1)^2} I)^{-1}x'Y \tag{5.4}$$

Replacing σ^2 and γ_1 with $\hat{\sigma}^2 = (n - k)^{-1}Y'MY$, $\hat{\gamma}_1 = B'B\hat{\sigma}^{-2}$, where $B = X + Y$, $x = (x'x)^{-1}x'$, we obtain

$$\hat{B}_{R1} = (x'x + \frac{\nu - 1}{\hat{\sigma}^2\lambda_1^{-1}(1 + \lambda_1\hat{\gamma}_1)^2} I)^{-1}x'Y \tag{5.4a}$$

For given x, Y the formula (5.4a) gives us one member of the (5.4) type estimators.

Repeating the above argumentation for the second type of regularization, i.e.

$$(x'x + \frac{\sigma^2}{\beta'\chi\beta} I)\beta = x'Y$$

or in alternative form

$$(x'x + \gamma_2^{-1}I)\beta = x'Y,$$

we arrive at the parametric family of estimators

$$B_{R2} = \frac{m^2}{m^2 + \sigma^2\lambda_1(\nu - 1)} B, \quad m^2 = (\delta^2 + \lambda_1\beta'\chi\beta)^2 \tag{5.5}$$

and the empirical regularizing estimator has the form

$$\hat{B}_{R2} = \frac{\hat{m}^2}{\hat{m}^2 + \hat{\sigma}^2\lambda_1(\nu - 1)} B, \quad \hat{m}^2 = (\hat{\delta}^2 + \lambda_1 B'\chi B)^2 \tag{5.5a}$$

In the case of (5.3) by the minimization of convex regularizing combination ϕ_{R3} with respect to β we obtain

$$B_{R3} = (x'x + \frac{1}{\nu - 1} I)^{-1} (x'Y + \frac{1}{\nu - 1} V_{.1})$$

It is known that our estimators [see: M i l o (1988), (1989)] or comments in the included Appendix) that the following statements are true.

Theorem 1. If Y is normally distributed with the mean $x\beta$ and dispersion $\sigma^2 I$ and $A = (x'x + \gamma_*^{-1} I)^{-1} x'$,

$$\gamma_*^{-1} = \frac{\gamma - 1}{\sigma^2 \lambda_1^{-1} (1 + \lambda_1 \gamma_1)^2}, \quad \text{then} \quad B_{R1} \sim N(A \times \beta, \ \sigma^2 AA').$$

Theorem 2. Under the assumptions of Th. 1 and $\lim_{n\to\infty} \lambda_1 = \infty$, $\lim_{n\to\infty} \lambda_1^{-1}\lambda_i = \bar{\nu}_i \in R$, $\beta'\beta = k$ the family B_{R1} is a consistent family of estimators.

Theorem 3. Under the assumptions of Th. 2 we have $\operatorname{plim}_{n\to\infty} B_{R2} = \beta$ i.e. the family B_{R2} is a consistent family of estimators.

Theorem 4. If the assumptions of Th. 2, except $\beta'\beta = k$ replaced with $\beta'\beta = k(\beta' v_{.1})^2$ hold, then B_{R3} is consistent and normally distributed estimator.

From the definitions of B_{R2}, B_{R3} and estimators bias it follows.

Theorem 5. Under the assumptions of Th. 4, and

$$\operatorname{cov}(B, \frac{Y'MY}{Y'N_1Y}) = 0, \quad \operatorname{cov}(Y'MY, (Y'N_1Y)^2 = 0, \quad ac \neq 0,$$

$$a = (n - k)^{-1}(\lambda_k - \lambda_1), \quad c = a\sigma^2(n - k)d^{-1}, \quad d = \xi Y'N_1Y,$$

$N_1 = x(x'x)^{-1}x'$ we have

$$\operatorname{bias}^2 \hat{B}_{R2} < \operatorname{bias}^2 B_{R3} \Leftrightarrow \sum_1^k \frac{\beta_i^2}{\lambda_i + \gamma} > \frac{\gamma^2 2\gamma V'_{.1} \beta - 2\gamma V_1' \beta(\lambda_1 + \gamma)}{(\lambda_1 + \gamma)^2},$$

$$\operatorname{bias}^2 \hat{B}_{R2} > \operatorname{bias}^2 B_{R3} \Leftrightarrow \sum_1^k \frac{\beta_i^2}{\lambda_i + \gamma} < \frac{\gamma^2 2\gamma V'_{.1} \beta - 2\gamma V_1' \beta(\lambda_1 + \gamma)}{(\lambda_1 + \gamma)^2},$$

$$\gamma = \frac{1}{\nu - 1}.$$

Since $MSE(B) = \sigma^2 \sum_1^k \lambda_i^{-1}$, and $MSE(B_{R3}) = \beta'\beta + \beta'\chi\chi_\nu^{-2}\chi\beta + \sigma^2 tr\chi\chi_\nu^{-2} +$ $+ \beta'\chi\chi_\nu^{-1}$ therefore it is easy to find conditions under which $MSE(B_{R3}) < MSE(B)$. Similar reasoning will lead us to fix superiority conditions for other regularizing estimators.

Summarizing, we can say that regularizing estimators are useful because they provide more stable solutions for the system or normal equations. Condition numbers for regularized matrices are under given conditions smaller then for non-regularized ones. Similarly new regularizing estimators can be more precise in terms of smaller values of MSE. These reasons as well as those given in §§ 2-5 speak for the usefulness of using regularizing estimators.

APPENDIX

A. Effects of bad-conditioning. In § 4 we used the following results

$$B = x^+Y, \quad x^+ = \chi^{-1}x', \quad \chi = x'x \tag{A1}$$

$$Y = xB, \quad E = Y - \hat{Y} = MY, \quad M = I - xx^+ \tag{A2}$$

$$Y'Y = Y'(I - M)Y, \quad E'E = Y'MY \tag{A3}$$

$$R^2 = 1 - \frac{E'E}{Y'CY} = 1 - \frac{Y'MY}{Y'CY}, \quad C = I - n^{-1}11' \tag{A4}$$

Due to SVD we have

$$M - 1 = I - \sum_{i=1}^k U_{.i}U'_{.i}, \chi = \sum_1^k \lambda_i V_{.i}V'_{.i} \tag{A5}$$

$$\chi^{-1} = \sum_{i=1}^k \lambda_i^{-1}V_{.i}V'_{.i} \quad \text{for} \quad k_o = k, \quad I - M = \sum_{i=1}^k U_{.i}U_{.i} \tag{A6}$$

$$x^+ = \sum_{i=1}^k \lambda_i^{-1/2}V_{.i}U'_{.i}, \quad xx^+ = \sum_{i=1}^k U_{.i}U'_{.i}, \quad x^+x = I. \tag{A7}$$

$$x'W = \sum_{i=1}^k \lambda_i^{1/2}V_{.i}U'_{.i}W, \quad x^+W = \sum_{i=1}^k \lambda_i^{-1/2}V_{.i}U'_{.i}W \tag{A8}$$

$$B = \beta + x^+W, \quad B - \beta = \sum_{i=1}^k \lambda_i^{-1/2}V_{.i}U'_{.i}W \tag{A9}$$

$$Y = \Sigma\, \lambda_i^{1/2}U_{.i}V'_{.i}\beta + W \tag{A10}$$

$$Y'Y = \sum_{i=1}^{k} \lambda_i (V'_{.i}\beta)^2 + W'W + 2\sum_{i=1}^{k} \lambda_i^{1/2} V'_{.i}\beta U'_{.i}W \tag{A11}$$

$$n\bar{Y}^2 = n\sum_{i=1}^{k} \bar{\lambda}_i^2 (V_i\beta)^2 + n\bar{W}^2 + 2n \sum_{1}^{k} \lambda_i \bar{U}_i \bar{W} V_{.i}\beta \tag{A12}$$

$$\bar{Y} = n^{-1}Y'1, \quad \bar{U}_i = n^{-1}U'_{.i}1, \quad \bar{W} = n^{-1}W'1 \tag{A13}$$

$$Y'CY = Y'Y - n\bar{Y}^2 \tag{A14}$$

$$\hat{Y} = \sum_{1}^{k} \lambda_i^{1/2} U_{.i}V'_{.i}\beta + \sum_{1}^{k} U_{.i}U'_{.i}W \tag{A15}$$

$$E = MW = W - \sum_{1}^{k} U_{.i}U'_{.i}W \tag{A16}$$

$$E'E = W'W - \sum_{1}^{k} (U'_{.i}W) \tag{A16a}$$

$$MSE\ B = \frac{\sigma^2}{\lambda_k}(1 + \ldots + \nu) \tag{A17}$$

$$Y'CY = W'W - n\bar{W}^2 \quad \text{if} \quad V_{.i} \perp \beta \quad \text{for each } i \tag{A18}$$

$$Y'CY = W'W - n\bar{W}^2 + \sum_{1}^{k} \lambda_i (1 - n\bar{U}_i^2) + 2\Sigma\, \lambda_i^{1/2}(W'U_{.i} - n\bar{U}_i\bar{W}) \tag{A19}$$

if for each $V_{.i} = \beta$

$$Y'CY \approx W'W - n\bar{W}^2 \quad \text{if} \quad n\bar{U}_i^2 \approx 1, \quad W'U_{.i} \approx n\bar{U}_i\bar{W} \tag{A20}$$

$Y'CY$ strongly depends on the values of expressions

$$\lambda_k(1 - n\bar{U}^2)(V'_{.k}\beta)^2, \quad U'_{.k}W - n\,\bar{U}_k W V'_{.k}\beta \tag{A21}$$

B. Properties of regularizing estimators.

Due to definition of A, γ from § 5 and Chebyshev inequality $\lim_{n\to\infty} \xi||B_{R1}-\beta|| = 0$. Hence plim $B_{R1} = \beta$. Normality follows from known theorems given in, f.ex. [S r i v a s t a v a, K h a t r i (1979)].

In the case of $\hat{B}_{R1}$ consistency follows from the fact that $\lim_{n\to\infty} \xi\lambda_1^{-1}\hat{\gamma} = 0$, $\lim_{n\to\infty} \xi\lambda_i^{-1}\gamma^{-2} = a_i$, $\gamma = c[Y'(M + \lambda_1 N_1)Y]^{-2}Y'MY$, $c = (n - k)(\nu - k)\lambda_1$.

Consistency of B_{R2} follows from the fact that $\mathrm{cov}(Y'MY,(Y'MY,(Y'N_1Y)^2) = 0$ and under the assumption of th. 3

$$\lim \gamma = 1, \quad \lim \gamma^2 = 1, \quad \gamma = \frac{m^2}{m^2 + \sigma^2(\lambda_k - \lambda_1)}.$$

Normality of B_{R2} follows from normality of B and boundedness of γ.

Consistency of B_{R3} come from $\lim \frac{\lambda_1 \gamma}{\lambda_i + \gamma} = 0$, $\lim \frac{\gamma}{\lambda_i + \gamma} = c$.

Normality follows from theorems given in the book of Srivastava, Knatri. In the proof of Th. 5 it must be remembered that matrices M, N_1 are projection matrices and we additionally use assumptions about zero covariance between Y'MY and $(Y'N_1Y)^2$ and B.

REFERENCES

Chatterjee S., Price B. (1977): *Regression analysis by example*, N. Y., Wiley.

Cook R., Weisberg S. (1982): *Residuals and influence in regression*, London, Chapman and Hall.

Farebrother R. W. (1988): *Linear least squares computations*, N. Y., Marcell Dekker.

Farrar D., Glauber R. (1967): *Multicollinearity in regression analysis*, Rev. of Econ. and Statist., Vol. 49, p. 92-107.

Gunst R. F. (1983): *Regression analysis with multicollinear predictor variables*, Comm. in Statist. Theory and Methods, Vol. 19, p. 2217-2260.

Harvey A. (1981): *The econometric analysis of time series*, Oxford, Philip Allan.

Hoerl A., Kennard R. (1970): *Ridge regression*, "Technometrics", p. 55-67.

Johnston J. (1963): *Econometric methods*, N. Y., McGraw Hill.

Kendall M. G., Buckland W. (1971): *A dictionary of statistical terms*, Edinburgh Oliver and Boyd for the Intern. Statist. Inst.

Kiviet J. (1987): *Testing linear econometric Models*, Amsterdam, Ilpendam.

Kramar W., Sonnerberger H. (1986): *The linear regression model under test*, Heidelberg Physica Verlag.

Magnus J., Neudecker H. (1988): *Matrix differential calculus*, N. Y., Willey.

Mason R. et al. (1975): *Regression analysis and problems of multicollinearity*, Commun. in Statist., Vol. 4, p. 277-292.

Milo W. (1988): *Properties of regularizing estimators*, 18-th Europ. Meet. of Statist., August 22-26/1988, Berlin.

Milo W. (1989): *Comparative analysis of biased regularizing estimator ESEM89*, Münich, Sept., p. 4-8.

Pollock D. S. G. (1979): *The algebra of econometrics*, N. Y., Willey.

Silvey S. (1969): *Multicollinearity and imprecise estimation*, R. Roy. Statist. Soc. B., p. 539-552.

Srivastava M., Knatri C. (1979): *An introduction to multivariate statistics*, N. Y., North Holland.

Trenkler G. et al. (1985): *Updating the ridge estimator*, Comput. Statist. Quart., Vol. 2, p. 135-141.

Vinod H. (1978): *Simulation and extension of a MSE estimator in comparison with Stein and Technometrics*, No. 3, p. 491-496.

Vinod H., Ullah A. (1981): *Recent advances in regression methods*, N. Y., Marcel Dekker.

Władysław Milo

O UŻYTECZNOŚCI IDEI REGULARYZACJI PRZY ESTYMACJI MODELI LINIOWYCH

Celem artykułu jest pokazanie czytelnikowi użyteczności idei regularyzacji w zmniejszaniu lub dużej redukcji negatywnych skutków występowania złego uwarunkowania danych. Skutki te obserwowano w samym estymatorze metody najmniejszych kwadratów jak i jego statystycznych i numerycznych charakterystykach. Podstawowe analizowane charakterystyki tego estymatora to: MSE, wariancja, próbkowe odchylenie standardowe, próbkowy współczynnik korelacji wielokrotnej (inaczej: współczynnik determinacji), statystyki testu t-Studenta oraz testu F. Zbadano też skutki estymacyjne przeprowadzania takich operacji jak centrowanie, ważenie danych. W celu zmniejszenia negatywnych skutków złego uwarunkowania proponuje się stosowanie estymatorów regularyzujących. W omawianym modelu są one zgodne i asymptotycznie normalne.