CONTRÔLE INFLACIONÁRIO NA DETERMINAÇÃO DOS COEFICIENTES DE REGRESSÃO EM PROBLEMAS ENVOLVENDO VARIÁVEIS NÃO ORTOGONAIS [1]

IVAN BARBOSA MACHADO SAMPAIO [2]

SINOPSE.- Por ocasião da inversão da matriz X'X no cômputo usual dos coeficientes de regressão, o condicionamento dos vetores não ortogonais na matriz X, bem como a possível alta correlação entre alguns dêstes, podem comprometer sèriamente a estimativa dos coeficientes de regressão correspondentes, superestimando-os em valor absoluto. Esta inflação é causada principalmente pela natureza da matriz invertida, acentuando-se à medida que esta se aproxima da singularidade. Algumas vêzes, a simples transformação de variáveis adequadas em tôrno da média é suficiente para reduzir uma alta correlação, entretanto, através do artifício matemático que consta da adição de k, $0 \leqslant k \leqslant 1$, aos elementos diagonais de X'X na forma de matriz de correlação, pode-se exercer razoável contrôle daquela inflação. Os coeficientes assim obtidos, embora "biased", se aproximam mais de seus valores reais.

INTRODUÇÃO

Variáveis independentes determinadas "a posteriori" são geralmente fontes de vetores não ortogonais, como o caso de dados meteorológicos, pêso e número de nódulos, quantidade de ração consumida e outros.

Em se relacionando variáveis dêsse tipo a um vetor sob estudo, digamos $\vec{Y}$, pode-se obter uma regressão múltipla num modêlo linear.

O têrmo linear aqui empregado implica tão sòmente na linearidade dos coeficientes de regressão, podendo as variáveis assumir valores algébricos de ordem n. Por exemplo, a regressão

$$Y = b_0X_0 + b_1X_1 + b_{11}X_1^2 + b_2X_2 + b_{12}X_1X_2$$

seria considerada um modêlo linear.

Para o caso de vetores ortogonais, a matriz de correlação obtida seria uma identidade, e a determinação dos coeficientes de regressão não sofreria interferência alguma proveniente de correlação entre variáveis, o que no caso não existiria. Entretanto, para o caso de vetores não ortogonais, a matriz de correlação conterá valores diferentes de zero nos elementos não diagonais. A observação dêsses elementos, que em valores absolutos podem variar de 0 a 1, indicará a ocorrência de inflação nos coeficientes. À medida que cada elemento se aproxima do valor 1, a matriz tende à singularidade e conseqüentemente há maior inflação nos valores dos coeficientes. Essa singularidade se comprova fàcilmente pela característica de simetria da matriz estudada. Por exemplo, digamos que a matriz de correlação A (n × n) se apresenta da seguinte maneira:

$$\begin{bmatrix} 1 & \cdots & a_{13} & \cdots & a_{1n} \\ \vdots & & & & \vdots \\ a_{31} & \cdots & 1 & \cdots & \vdots \\ \vdots & & & & \vdots \\ a_{n1} & \cdots & & \cdots & 1 \end{bmatrix}$$

onde o elemento a_{13}, correlação das variáveis 1 e 3, é próximo de 1. Devido à simetria da matriz, o elemento a_{31} terá o mesmo valor e se os demais elementos das linhas 1 e 3 forem baixos valores de correlação, exceto os diagonais, iguais à unidade, as citadas linhas serão aproximadamente iguais e o determinante da matriz será, portanto, fração próxima de zero.

Também conduzindo à singularidade, Riley (1955) cita o caso da matriz formada pelas equações normais, quando o modêlo envolve uma equação polinomial de grau elevado. Neste caso, a matriz simétrica assim definida terá a última coluna aproximadamente proporcional à anterior, recaindo no caso de um determinante quase nulo. O próprio Riley propôs na ocasião que, em se resolvendo $Y = \beta X$, onde X é matriz quase singular, fôsse tentado o sistema modificado.

$$Y = (X + kI)\gamma, \qquad (1)$$

onde k seria uma constante positiva inferior a 1. A nova matriz $C = X + kI$ é definitivamente melhor condicionada que X, uma vez que seus determinantes em função dos "eigenvalues" são:

$$\det X = \prod_{i=1}^{n} \lambda_i \text{, que é menor que}$$

$$\det C = \prod_{i=1}^{n} (\lambda i + k).$$

Como $X = C - kI$,

$$X^{-1} = C^{-1} + kC^{-2} + \ldots + k^m C^{m-1} + \ldots$$

e os valores dos coeficientes seriam

$$\begin{aligned} \beta &= X^{-1}Y \\ &= C^{-1}Y + kC^{-2}Y + \ldots + k^m C^{-m-1}Y + \ldots \\ &= \gamma + kC^{-1}\gamma + \ldots + (kC^{-1})^m\gamma + \ldots \end{aligned} \qquad (2)$$

Escolhido o valor de k, γ pode ser calculado por (1) e a seguir todos os têrmos da expansão (2). Note-se que cada têrmo desta contribui um pouco para o valor de β. Na verdade, tais têrmos adicionais contribuem muito pouco, afetando apenas casas decimais distantes da vírgula, mas caso as adições tenham valores relativamente substanciais e decresçam vagarosa-

[1] Recebido 4 jan. 1971, aceito 26 mar. 1971.

[2] Eng.º Agrônomo do Setor de Estatística Experimental e Análise Econômica do EPE, Ministério de Agricultura, 9.º andar, Brasília, DF., bolsista do Conselho Nacional de Pesquisas (CNPq. 14450/70).

mente, significa que o sistema está altamente mal condicionado e dificilmente se obterá uma solução razoável para o mesmo.

Hoerl (1962) também notou uma grande variação nos valores dos mesmos coeficientes para diferentes grupos de dados similares, devido à correlação e distribuição das variáveis. Verificou que utilizando o artifício matemático proposto por Riley (1955), variando gradualmente o valor de k, havia uma região onde os coeficientes variavam dràsticamente buscando uma posterior estabilização, ao passo que o valor da soma de quadrados do resíduo experimentava um aumento relativamente pequeno. Por êste processo, que Hoerl decidiu chamar solução "ridge", chega-se a coeficientes cujos valores não se deixam afetar por dados mal condicionados e mal distribuídos, prestando-se à predição de distintos grupos de dados similares.

Baseados nos trabalhos originais de Riley (1955) e Hoerl (1962), Hoerl e Kennard (1970a) introduziram o método "Ridge Trace" para resolver problemas não ortogonais. Considerando que o método dos quadrados mínimos pode conduzir a valores inflacionados dos coeficientes de regressão, Hoerl e Kennard propuseram controlar tal inflação através do uso da matriz de correlação X'X + kI, k variando de 0 a 1.

O novo sistema a resolver passa a ser, então,

$$\hat{\beta}^* = (X'X + kI)^{-1} X'Y.$$

Devido à alteração na matriz X'X, o valor da soma de quadrados do resíduo sofrerá acréscimos sucessivos à proporção que o valor de k aumenta, e sua magnitude é dada pela fórmula

$$SQR = Y'Y - (\hat{\beta}^*)' XY - k(\hat{\beta}^*)' (\hat{\beta}^*). \qquad (3)$$

Quando $k = 0$, a solução é a determinada pelo método dos quadrados mínimos, e o valor da SQR se reduz aos dois primeiros têrmos da equação (3).

Os coeficientes $\hat{\beta}^*$ assim estimados para cada valor de k são representados gràficamente em relação a k. À proporção que k aumenta, os coeficientes maiores, principalmente, diminuem seus valores absolutos até tornarem-se relativamente estáveis. Por outro lado, o valor da soma de quadrados do resíduo crescerá gradativamente. Examinando-se as linhas correspondentes aos coeficientes e à SQR, seleciona-se um valor de k que forneça coeficientes os mais estabilizados possíveis, mas que ainda corresponda a um acréscimo reduzido na soma de quadrado residual. Os gráficos assim obtidos oferecem fácil identificação para o melhor valor de k, como pode ser visto na Fig. 2.

Desde que k seja diferente de zero, os valores dos coeficientes obtidos conterão sempre uma fração de "bias"; entretanto, apresentarão valores mais próximos aos reais, uma vez que a inflação ocorrida no processo dos quadrados mínimos se deve exclusivamente ao efeito de correlação entre variáveis e a problemas computacionais originados pela quase singularidade da matriz X'X.

Material e Métodos

A fim de ilustrar as causas, conseqüências e contrôle da inflação nos valores dos coeficientes de regressão em problemas não ortogonais, foram coletados dados climatológicos considerados de importância para a cultura do milho no Estado de Iowa, E.U.A., no período de 1930 a 1962. As variáveis selecionadas para a equação de regressão foram:

X_1 = avanço tecnológico anual (efeito linear);
X_2 = avanço tecnológico anual (efeito quadrático);
X_3 = precipitação pluviométrica de maio;
X_4 = temperatura média de maio;
X_5 = precipitação pluviométrica de junho;
X_6 = temperatura média de junho (efeito linear);
X_7 = temperatura média de junho (efeito quadrático);
X_8 = precipitação pluviométrica de julho;
X_9 = temperatura média de julho;
X_{10} = precipitação pluviométrica de agôsto e
X_{11} = temperatura média de agôsto.

Exceto a primeira variável, fàcilmente codificada e representando o progresso tecnológico (mecanização, técnicas de fertilização e manejo, conservação, híbridos, etc.), tôdas as demais foram vetores de dados não ortogonais.[3]

Thompson (1963) manuseou originalmente o mesmo conjunto de dados e obteve uma equação e análise estatística baseando-se no método dos quadrados mínimos; entretanto, pouco depois, Shaw e Thompson (1964) chamariam a atenção para os perigos de qualquer tentativa de extrapolação para os períodos maiores que cinco anos, com base na equação determinada, em virtude de alguns altos coeficientes de correlação.

No presente trabalho, decidimos executar o método de contrôle de inflação proposto por Hoerl e Kennard (1970b) e comparar os resultados com os do processo dos quadrados mínimos.

Primeiramente foi calculada a matriz de correlação exigida para utilização do "Ridge Trace", sendo necessário executar a seguinte transformação para cada variável:

$$X_i^* = \frac{X_i - \overline{X}_i}{\sqrt{\sum (X_i - \overline{X}_i)^2}}$$

onde X_i^* é o novo vetor de dados transformados,
X_i é o vetor de dados originais e
$\overline{X}_i$ é a média da variável X_i.

Os coeficientes das equações normais obtidas dos dados transformados X_i^* formam a matriz de correlação.

Resultados e Discussão

No Quadro 1 podem-se observar os valores absolutos da correlação entre as variáveis, 0 a 58%, a maioria assumindo valores relativamente baixos. Entretanto, a correlação entre as variáveis 6 e 7 foi de 99%. A causa desta alta correlação foi justamente o argumento de Riley (1955), embora o grau envolvido tenha sido sòmente a segunda potência da variável 6; houve uma proporcionalidade entre esta e seu quadrado, a variável 7.

A determinação dos coeficientes de regressão neste caso foi executada pelo processo de "Ridge Trace", onde também se pode avaliar a solução pelos quadrados mínimos ($k = 0$). A representação gráfica dos mesmos pode ser vista na Fig. 1. No intervalo $0 \geqslant k \geqslant 0{,}1$, os coeficientes sofrem uma variação drástica e desordenada, que se acentua na vizinhança da solução pelos quadrados mínimos. Neste ponto, $k = 0$, os coefici-

[3] Os dados aqui utilizados são parte da tese de mestrado do autor, apresentada ao Departamento de Estatística de Iowa State University em agôsto de 1970.

entes estão altamente superestimados, principalmente os que apresentam índices de correlação mais altos. Para $k > 0,2$, já se observa uma tendência à estabilização.

Observe-se também a linha tracejada apresentando o acréscimo gradual do valor da soma de quadrados residual. A súbita ascensão observada entre os valores de $k = 0$ e $k = 0,02$ pràticamente elimina a possibilidade de escolha válida para valor de k, pois os coeficientes ainda estão bastante instáveis no intervalo.

A seguir foram transformados os dados da variável 6 em tôrno da média. A nova matriz de correlação está representada no Quadro 2, e a correlação com a variável 7 foi controlada de 0,99 para 0,26. Todos os demais valores da matriz permaneceram inalterados, exceto, naturalmente, os que envolveram a variável 7.

A nova determinação gráfica do "Ridge Trace" foi executada segundo a Fig. 2. Após o contrôle da alta correlação existente, tanto as linhas dos coeficientes como a da soma de quadrados residual se mostraram muito mais coordenadas que no caso anterior. As variações bruscas, reflexos de mau condicionamento da matriz com relação à singularidade, deixaram de existir. O contrôle inflacionário oferecido pelo 'Ridge Trace" se mostra muito mais claramente, inclusive para as variáveis que antes estavam altamente correlacionadas.

QUADRO 1. *Matriz de correlação das variáveis, antes da transformação da variável 6 em tôrno da média*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Y
X_1	1,00											
X_2	0,00	1,00										
X_3	0,04	0,04	1,00									
X_4	—0,04	0,16	—0,12	1,00								
X_5	—0,07	—0,45	0,32	—0,39	1,00							
X_6	—0,26	0,17	—0,45	0,29	—0,35	1,00						
X_7	—0,27	0,18	—0,46	0,29	—0,35	0,99	1,00					
X_8	0,40	0,07	—0,03	0,08	—0,28	—0,08	—0,08	1,00				
X_9	—0,54	0,11	—0,30	0,15	—0,16	0,26	0,26	—0,51	1,00			
X_{10}	0,06	0,03	0,20	0,00	—0,05	0,17	0,16	0,02	—0,09	1,00		
X_{11}	—0,08	0,07	0,09	0,12	0,08	—0,23	—0,23	—0,42	0,37	—0,32	1,00	
Y	0,74	0,04	0,17	—0,13	—0,13	—0,14	—0,15	0,57	—0,58	0,21	—0,35	1,00

QUADRO 2. *Matriz de correlação das variáveis, depois da transformação da variável 6 em tôrno da média*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Y
X_1	1,00											
X_2	0,00	1,00										
X_3	0,04	0,04	1,00									
X_4	—0,04	0,17	—0,12	1,00								
X_5	—0,07	—0,45	0,33	—0,39	1,00							
X_6	—0,26	0,17	—0,45	0,29	—0,35	1,00						
X_7	—0,22	0,11	—0,24	—0,01	—0,22	0,26	1,00					
X_8	0,40	0,07	—0,03	0,08	—0,28	—0,08	0,04	1,00				
X_9	—0,54	0,11	—0,30	0,15	—0,16	0,26	0,03	—0,51	1,00			
X_{10}	0,06	0,03	0,20	0,00	—0,05	0,17	—0,15	0,02	—0,09	1,00		
X_{11}	—0,08	0,07	0,09	0,12	0,08	—0,23	—0,14	—0,42	0,37	—0,32	1,00	
Y	0,74	0,04	0,17	—0,13	—0,13	—0,14	—0,41	0,57	—0,58	0,21	—0,35	1,00

QUADRO 3. *Valores dos coeficientes de regressão e da soma de quadrados residual para Iowa, antes da transformação da variável 6 em tôrno da média*

Variáveis	Valores de k										
	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
1	0,51	0,54	0,48	0,44	0,41	0,38	0,35	0,33	0,32	0,30	0,29
2	0,00	0,00	0,00	0,01	0,01	0,01	0,01	0,01	0,02	0,02	0,02
3	0,17	0,16	0,13	0,11	0,09	0,08	0,08	0,07	0,06	0,06	0,06
4	—0,20	—0,12	—0,11	—0,10	—0,09	—0,08	—0,07	—0,06	—0,06	—0,06	—0,05
5	—0,19	—0,10	—0,09	—0,08	—0,08	—0,08	—0,07	—0,07	—0,06	—0,06	—0,06
6	12,81	0,08	0,04	0,03	0,02	0,01	0,00	0,00	0,00	0,00	0,00
7	—12,74	—0,02	—0,02	0,00	—0,01	—0,01	—0,01	—0,01	—0,01	—0,02	—0,02
8	0,24	0,23	0,22	0,21	0,20	0,19	0,18	0,18	0,17	0,17	0,16
9	—0,09	—0,07	—0,10	—0,12	—0,13	—0,13	—0,14	—0,14	—0,14	—0,14	—0,14
10	—0,02	0,07	0,08	0,08	0,08	0,08	0,08	0,07	0,07	0,07	0,07
11	—0,17	—0,12	—0,11	—0,11	—0,11	—0,10	—0,10	—0,10	—0,10	—0,09	—0,09
SQR	0,15	0,27	0,29	0,30	0,31	0,32	0,34	0,35	0,36	0,37	0,38

Também o aumento da soma de quadrados do resíduo se apresenta gradativo, permitindo uma conveniente escolha de k.

Ao serem comparadas as Fig. 1 e 2, nota-se que na primeira o "Ridge Trace" pràticamente estabilizou todos os coeficientes para os mesmos valores alcançados na segunda, exceto o correspondente à variável 7. Isso demonstra a grande necessidade de serem controladas as altas correlações que, na ausência dêsse contrôle, se refletiriam, indubitàvelmente, mesmo sôbre o poder de contrôle do "Ridge Trace".

A variação numérica dos coeficientes para cada valor de k, bem como o valor da soma de quadrados residual correspondente, podem ser seguidos nos Quadros 3 e 4, para ambas as Fig.

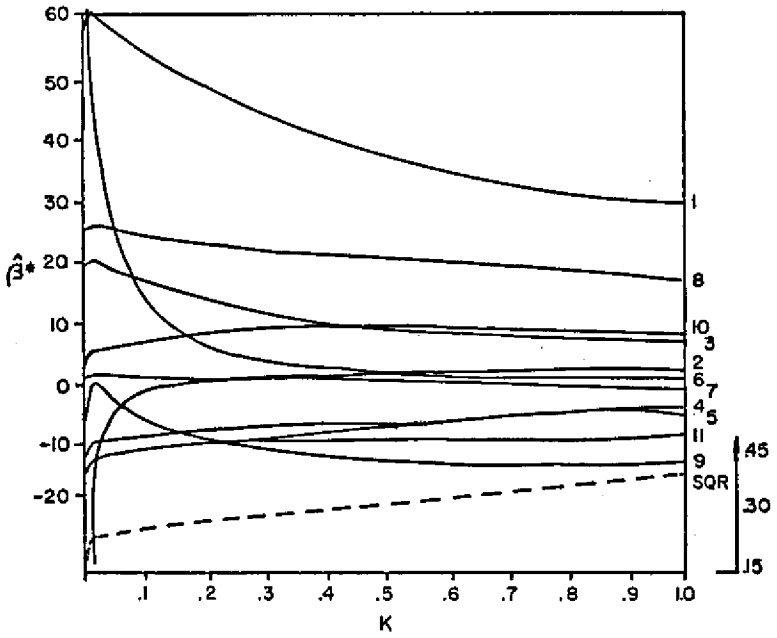

FIG. 1. *Variação dos coeficientes e da SQR da regressão antes da transformação da variável 6.*

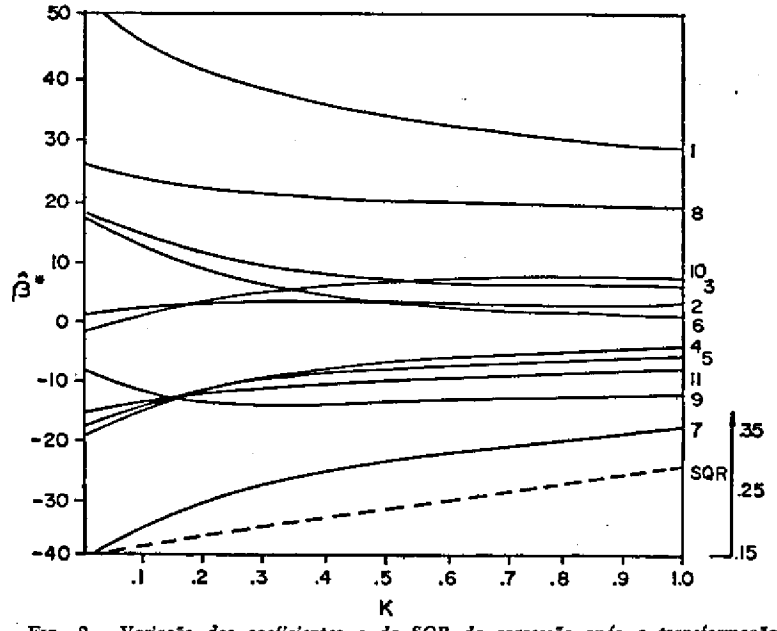

FIG. 2. *Variação dos coeficientes e da SQR da regressão após a transformação da variável 6.*

QUADRO 4. *Valores dos coeficientes de regressão e da soma de quadrado residual para Iowa, após a transformação da variável 6*

Variáveis	Valores de k										
	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
1	0,52	0,45	0,41	0,38	0,36	0,34	0,32	0,30	0,29	0,28	0,27
2	0,00	0,01	0,02	0,02	0,20	0,02	0,02	0,02	0,02	0,02	0,02
3	0,17	0,12	0,09	0,08	0,07	0,06	0,06	0,05	0,05	0,05	0,04
4	—0,20	—0,16	—0,14	—0,12	—0,11	—0,09	—0,09	—0,08	—0,07	—0,07	—0,06
5	—0,19	—0,16	—0,14	—0,12	—0,11	—0,10	—0,10	—0,09	—0,08	—0,08	—0,07
6	0,17	0,11	0,08	0,06	0,04	0,03	0,02	0,02	0,01	0,01	0,00
7	—0,39	—0,35	—0,32	—0,29	—0,27	—0,25	—0,24	—0,22	—0,21	—0,20	—0,19
8	0,24	0,23	0,22	0,21	0,20	0,19	0,19	0,18	0,18	0,17	0,17
9	—0,09	—0,13	—0,14	—0,15	—0,15	—0,15	—0,15	—0,15	—0,15	—0,15	—0,14
10	—0,02	—0,01	0,03	—0,04	—0,04	—0,04	0,05	0,05	0,05	0,05	0,05
11	—0,17	—0,15	—0,14	—0,13	—0,13	—0,12	—0,12	—0,11	—0,11	—0,11	—0,10
SQR	0,15	0,16	0,17	0,19	0,21	0,21	0,23	0,25	0,26	0,27	0,29

Pelo Quadro 4, ou mais fàcilmente pela Fig. 2, um valor para k pode ser escolhido levando-se em conta a estabilidade dos coeficientes e a menor soma de quadrados residual possível. Evidentemente, valores de k compreendidos entre 0,2 e 0,3 alcançam êsse objetivo, uma vez que os valores das SQR correspondentes são 0,16 e 0,17, comparados a 0,15, dado pela solução dos quadrados mínimos.

Uma subdivisão neste intervalo pode ser levada a efeito pelo experimentador, se houver interêsse para tal. Trabalhando-se com o auxílio de computadores, isto não constituirá obstáculos. Na infinidade de situações que podem ocorrer num determinado intervalo elegido de k, a ponderação do experimentador terá grande importância.

CONCLUSÕES

A correlação elevada de variáveis pode conduzir a matrizes operacionais quase singulares, que por sua vez fornecerão coeficientes de regressão superestimados em seus valores absolutos.

Torna-se, portanto, compensador o estudo de correlação entre as variáveis não ortogonais existentes em um modêlo linear. O contrôle de valores altos de correlação deve ser procurado, sempre que possível, através da transformação de uma das variáveis envolvidas, em tôrno da média. Devido à experiência obtida com outros experimentos similares, aconselhamos que a referida transformação seja efetuada sempre que a correlação se mostrar superior a 75%.

Por outro lado, mesmo baixos valores de correlação podem estar condicionados de tal forma na matriz que esta esteja também próxima da singularidade. A solução do "Ridge Trace" controlará a inflação dos coeficientes de regressão causada por esta quase singularidade, ou, se fôr o caso, pelas próprias interações complexas das variáveis correlacionadas. Aqui, os coeficientes possuem uma parcela de "bias", sendo, entretanto, estimativas mais próximas dos valores reais.

REFERÊNCIAS

Hoerl, A.E. 1962. Application of ridge analysis to regression problems. Chem. Eng. Progr. 58(3):54-59.

Hoerl, A.E. & Kennard, R.W. 1970a. Ridge regression: biased estimation for monorthogonal problems. Technometrics 12: 55-67.

Hoerl, A.E. & Kennard, R.W. 1970b. Ridge regression: application to monorthogonal problems. Technometrics 12:69-82.

Riley, J.D. 1955. Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. Mathematical Tables and other Aids to Computation 9:96-101.

Shaw, F.S. & Thompson, L.N. 1964. Grain yields and weather fluctuation. Iowa State Univ. Sci. Technol. Center Agric. Econ. Development 20:9-32.

Thompson, L.N. 1963. Weather and Technology in the production of corn and soybeans. Iowa State Univ. Sci. Technol. Center Agric. Econ. Development 17:1-31.

ABSTRACT.- Sampaio, I.B.M. 1972. *Controlling overestimation of the regression coefficients in nonorthogonal linear models.* Pesq. agropec. bras., Sér. Agron., 7:65-69. (Escrit. Pesq. Exp., Min. Agricultura, 9.º andar, Brasília, DF, Brazil)

Dealing with nonorthogonal problems, overestimation of the regression coefficients may be caused either by the presence of correlated variables or by the nature of the X'X matrix which may be near singular itself. Adding small values to the diagonal elements of the correlation matrix seems to control satisfactorily this inflation without losing too much precision. The coefficients obtained from the new pertubed system are slightly biased, however, their values tend to be closer to their actual values.