1. Publicação nº	2. Versão	3. Data	5. Distribuição
INPE-2401-PRE/115		*Maio, 1982*	☐ Interna ☒ Externa
4. Origem *DSE/DES*	Programa *NAS/MEDEP*		☐ Restrita

6. Palavras chaves - selecionadas pelo(s) autor(es)

ESTIMADOR "RIDGE"
MODELOS DE REGRESSÃO LINEAR
MULTICOLINEARIDADE

7. C.D.U.: *519.233.5*

8. Título *INPE-2401-PRE/115*

ESTUDO DE PROBLEMAS DE REGRESSÃO MALCONDICIONADOS OU COM PRESENÇA DE MULTICOLINEARIDADE

10. Páginas: *15*

11. Última página: *14*

12. Revisada por

Paulo Renato de Morais

9. Autoria *Luiz Antonio Nogueira Lorena*
José Roberto Reis

Assinatura responsável José Roberto Reis

13. Autorizada por

Parada

Nelson de Jesus Parada
Diretor

14. Resumo/Notas

Apresenta-se uma aplicação do estimador "ridge" para seleção de variáveis no modelo de regressão linear múltipla. Relaciona-se o problema de multicolinearidade e malcondicionamento do sistema de equações normais, usando-se o conceito de número de condição espectral da matriz associada ao sistema. Apresenta-se um exemplo de estimação da demanda por vestuário a título de comparação deste método apresentado com outros métodos tradicionais para seleção de variáveis.

15. Observações *Submetido para apresentação no 5º Simpósio Nacional de Probabilidade e Estatística a se realizar no período de 26 a 30 de julho de 1982, São Paulo, SP.*

ESTUDO DE PROBLEMAS DE REGRESSÃO MALCONDICIONADOS OU COM PRESENÇA DE MULTICOLINEARIDADE

L.A.N.Lorena (INPE)* e J.R. Reis (INPE)*

RESUMO

Apresenta-se uma aplicação do estimador "ridge" para seleção de variáveis no modelo de regressão linear múltipla. Relaciona-se o problema de <u>multicolinearidade</u> e mal condicionamento do sistema de equações normais, usando-se o conceito de número de condição espectral da matriz associada ao sistema.

Apresenta-se um exemplo de estimação da demanda por vestuário a título de comparação deste método apresentado com outros métodos tradicionais para seleção de variáveis.

ABSTRACT

An application of the ridge estimator to the selection of variables in the linear multiple regression model is presented. The concept of spectral condition number of the matrix associated to the system is used to relate the multicollinearity and ill-conditioned problems of the normal system of equations. An example of the estimation of demand for clothing is presented to compare this method with other traditional methods for the selection of variables.

---

* *Instituto de Pesquisas Espaciais, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Caixa Postal 515, 12200 - São José dos Campos, SP - Brasil.*

1. INTRODUÇÃO

O presente trabalho tem por objetivo utilizar o método de estimação "ridge" na seleção de variáveis relevantes ao modelo da regressão linear múltipla. Este método foi inicialmente sugerido por Hoerl and Kennard (1970), para atenuar os efeitos de forte multicolinearidade nesses modelos.

O trabalho está voltado unicamente para o aspecto de estimação, sendo evidenciado os efeitos da multicolinearidade na solução de sistemas de equações normais malcondicionados.

Supondo-se um modelo de regressão múltipla normalizado, na Seção 2, enfatiza-se a relação existente entre malcondicionamento e multicolinearidade. Apresentam-se alguns aspectos do estimador "ridge" na Seção 3 e a Seção 4 mostra um exemplo de aplicação do método em um problema que usa variáveis econômicas, discutido em Koutsoyiannis (1981), onde foram utilizados testes convencionais para seleção de variáveis, tais como "F" e "critério R^2".

2. O MALCONDICIONAMENTO E A MULTICOLINEARIDADE

Seja o modelo de regressão linear:

$$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}, \qquad (1)$$

onde $X(nxp)$ é a matriz de observações das variáveis independentes do modelo, admitindo-se que X tem característica p e os vetores associados às variáveis são normalizados. $\underset{\sim}{Y}(nx1)$ é o vetor normalizado de observações da variável dependente do modelo, $\underset{\sim}{\beta}(px1)$ é o vetor de parâmetros, e $\underset{\sim}{\varepsilon}(nx1)$ é o vetor de erros, tal que $E(\underset{\sim}{\varepsilon})=\underset{\sim}{0}$, $E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}^t)=\sigma^2 I_n$, onde I_n é a matriz identidade de ordem n, e t denota transposição.

Seja o sistema de "equações normais" na forma

$$A \underset{\sim}{x} = \underset{\sim}{b}, \tag{2}$$

fazendo-se $A=X^tX$, $\underset{\sim}{x}=\underset{\sim}{\beta}$, e $\underset{\sim}{b}=X^t\underset{\sim}{Y}$.

Para motivar e caracterizar a definição do número de condição de uma matriz, considere-se a sensibilidade de $\underset{\sim}{x}$ em (2) a pequenas perturbações nos elementos de A e $\underset{\sim}{b}$. Com respeito somente a pequenas variações, $\delta\underset{\sim}{b}$ em $\underset{\sim}{b}$, a nova solução $\underset{\sim}{x} + \delta\underset{\sim}{x}$ de (2) será definida por:

$$A(\underset{\sim}{x} + \delta\underset{\sim}{x}) = \underset{\sim}{b} + \delta\underset{\sim}{b}. \tag{3}$$

Subtraindo-se (2) de (3), resulta $A\,\delta\underset{\sim}{x} = \delta\underset{\sim}{b}$, ou:

$$\delta\underset{\sim}{x} = A^{-1}\,\delta\underset{\sim}{b}. \tag{4}$$

Então,

$$||\delta\underset{\sim}{x}||=||A^{-1}\,\delta\underset{\sim}{b}||\leq||A^{-1}||.||\delta\underset{\sim}{b}||, \tag{5}$$

onde as normas de matriz e de vetor são compatíveis (Wilkinson, 1965). De (2) tem-se $||\underset{\sim}{b}||\leq||A||.||\underset{\sim}{x}||$, e segue-se de (5) que o erro relativo, $\frac{||\delta\underset{\sim}{x}||}{||\underset{\sim}{x}||}$, na solução de (2) satisfaz

$$\frac{||\delta\underset{\sim}{x}||}{||\underset{\sim}{x}||} \leq \frac{||A^{-1}||.||\delta\underset{\sim}{x}||}{||A||^{-1}.||\underset{\sim}{b}||} = ||A||.||A^{-1}||\ \frac{||\delta\underset{\sim}{b}||}{||\underset{\sim}{b}||} \tag{6}$$

Como $||A||.||A^{-1}||$ é um indicador da sensibilidade da solução de (2) a perturbações, define-se $K(A) = ||A||.||A^{-1}||$ como número de condição da matriz A.

Define-se a norma espectral de uma matriz genérica B, e indica-se por $||B||_e$, como: $||B||_e = \sqrt{\xi_{máx}}$, onde $\xi_{máx}$ é o maior autovalor de B^tB. Como a matriz $A = X^tX$ é simétrica e definida positiva, seus autovalores são todos reais e positivos e, então, $||A||_e = \lambda_{máx}$, onde $\lambda_{máx}$ é o maior autovalor de $A^tA = A^2$, e, portanto, o maior autovalor de A (Schwarz et alii, 1973).

Define-se o número de condição espectral de uma matriz genérica B, denotado por $K_e(B)$, como:

$K_e(B) = ||B||_e \cdot ||B^{-1}||_e$. Para a matriz A,

$$K_e(A) = ||A||_e \cdot ||A^{-1}||_e = \frac{\lambda_{máx}}{\lambda_{mín}}, \tag{7}$$

onde $\lambda_{mín}$ é o menor autovalor de A, e, portanto, $1/\lambda_{mín}$ é o maior autovalor de A^{-1}.

Assim, a Relação (7) fornece um número que indica a sensibilidade a perturbações da solução das equações normais. Observa-se que: $K_e(A) = ||A||_e \cdot ||A^{-1}||_e \geq ||AA^{-1}||_e = 1$, com a igualdade valendo somente no caso de A ser ortogonal. Esta última afirmação estabelece a relação existente entre um problema de regressão com variáveis independentes correlacionadas (isto é, com presença de multicolinearidade) e o número de condição espectral para as equações normais, ou seja, conforme $K_e(A) >> 1$ existe multicolinearidade e instabilidades na obtenção de $\hat{\underset{\sim}{\beta}}$.

3. ALGUNS ASPECTOS DO ESTIMADOR "RIDGE"

Seja o modelo de regressão linear (1) e o estimador de mínimos quadrados de $\underset{\sim}{\beta}$, $\hat{\underset{\sim}{\beta}} = (X^tX)^{-1} X^t\underset{\sim}{Y}$.

Define-se o quadrado da distância entre $\underset{\sim}{\beta}$ e $\hat{\underset{\sim}{\beta}}$ por :

$$D^2 = ||\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}||^2 = \langle\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}, \hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}\rangle = (\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})^T (\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}). \tag{8}$$

Então (Searle, 1971):

$$E(D^2) = \sigma^2 \cdot \text{Traço } (X^tX)^{-1}, \tag{9}$$

ou $$E(\hat{\underset{\sim}{\beta}}^t \hat{\underset{\sim}{\beta}}) = \underset{\sim}{\beta}^t \underset{\sim}{\beta} + \sigma^2 \cdot \text{Traço}(X^tX)^{-1}, \tag{10}$$

e, quando $\underset{\sim}{\varepsilon}$ é normalmente distribuído, tem-se:

$$\text{Var}(D^2) = 2\sigma^4 \cdot \text{Traço}(X^tX)^{-2}. \tag{11}$$

Como X^tX é simétrica e definida positiva, pois tem característica p, logo ela possui p autovalores reais positivos. Sejam: $\lambda_{máx} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \lambda_{mín} > 0$, os p autova lores de X^tX, então, de (9), o valor médio do quadrado da distância entre $\hat{\underset{\sim}{\beta}}$ e $\underset{\sim}{\beta}$ é dado por:

$$E(D^2) = \sigma^2 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right) , \tag{12}$$

e, de (11), sua variância, quando o erro é normal, será:

$$\text{Var}(D^2) = 2\sigma^4 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right)^2 . \tag{13}$$

Portanto, limites inferiores para a média e variância são, respectivamente: $\sigma^2/\lambda_{mín}$ e $2\ \sigma^4/\lambda_{mín}$.

Sabe-se que quando $0 < \lambda_{mín} << 1, K_e(\tilde{A}) >> 1$ existirá multicolinearidade. Como consequência, verifica-se que a distância entre $\hat{\underset{\sim}{\beta}}$ e $\underset{\sim}{\beta}$ tende a ser grande e, de acordo com (10),

$$E(\hat{\underset{\sim}{\beta}}^t \underset{\sim}{\beta}) = E(||\hat{\underset{\sim}{\beta}}||^2) = ||\underset{\sim}{\beta}||^2 + \sigma^2/\lambda_{mín} , \tag{14}$$

e $||\hat{\underset{\sim}{\beta}}||$ estará "inflacionada".

Neste caso, o malcondicionamento implica um aumento em $||\hat{\underset{\sim}{\beta}}||$. Quando o valor do coeficiente de correlação simples, r_{ij}, $i \neq j$, é próximo a 1 (um) para as variáveis x_i e x_j, o reflexo em $\hat{\underset{\sim}{\beta}}$ pode ser sério, com grandes instabilidades nos coeficientes do Modelo (1).

Para contornar o problema da "inflação" e instabilidade associadas ao estimador $\hat{\underset{\sim}{\beta}}$, Hoerl and Kennard (1970) sugeriram usar outro estimador $\hat{\underset{\sim}{\beta}}^*$, dado por:

$$\hat{\underset{\sim}{\beta}}^* = [X^tX + KI]^{-1}\ X^t\underset{\sim}{Y}\ ;\ K \geq 0. \tag{15}$$

Tal estimador recebeu o nome de estimador "ridge".

A seguir são apresentadas algumas propriedades do estimador "ridge":

a) $||\hat{\underset{\sim}{\beta}}^*||$ é uma função monótona decrescente contínua em K, tal que quando $K \to \infty$, $||\hat{\underset{\sim}{\beta}}^*|| \to 0$, ou seja, o problema da "inflação" estará atenuado para $\hat{\underset{\sim}{\beta}}^*$.

b) O estimador $\hat{\underset{\sim}{\beta}}^*$ é uma transformação linear de $\hat{\underset{\sim}{\beta}}$, e que depende unicamente de X e K. Seja: $\hat{\underset{\sim}{\beta}}^* = (X^tX+KI)^{-1}\ X^t\underset{\sim}{Y}$, mas $X^t\underset{\sim}{Y} = (X^tX)\hat{\underset{\sim}{\beta}}$, então $\hat{\underset{\sim}{\beta}}^* = (X^tX+KI)^{-1}(X^tX)\hat{\underset{\sim}{\beta}}$, ou $\hat{\underset{\sim}{\beta}}^* = Z_K\hat{\underset{\sim}{\beta}}$. Segue

-se, imediatamente, que $\hat{\beta}^*$ é um estimador viciado de β , pois $E(\hat{\beta}^*) = Z_K\beta$.

c) Se $\beta^t\beta = ||\beta||^2$ é limitada, então existe um K tal que $E(D^2)$ para $\hat{\beta}^*$ é menor que $E(D^2)$ para $\hat{\beta}$.

Hoerl and Kennard (1970) sugerem ainda uma representação bidimensional, em K e $\hat{\beta}^*(K)$, denominada traço "ridge", em que, considerando-se K variando no intervalo $[0,1]$, um certo número (geralmente quinze, para K = 0; 0,02; 0,04 ; 0,06; 0,08; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1) de estimação de $\hat{\beta}^*(K)$ são realizadas. A cada valor de $K>0$, $K\to1$, e de acordo com os valores de r_{ij}, $i \neq j$, as instabilidades no valor dos coeficientes são diminuídas e os autores sugerem uma seleção visual para o melhor valor de K. Evidentemente, para K = 0, tem-se $\hat{\beta}^*(0) = \hat{\beta}$, o estimador por mínimos quadrados.

Na Seção 4, apresenta-se um exemplo de traço "ridge" na seleção de variáveis, aplicado a uma função demanda por vestuário, em que se usa o critério visual para a seleção do "melhor" estimador "ridge".

Usa-se também, no exemplo, o critério de comparar cada estimativa com a anterior, ou seja:

$$|\hat{\beta}_i^* (K_{atual}) - \hat{\beta}_i^* (K_{anterior})| \leq \Delta, \qquad (16)$$

onde Δ é um pequeno valor prefixado.

4. EXEMPLO

O exemplo apresentado é discutido por Koutsoyiannis (1981), páginas 240 a 242, para mostrar os efeitos da multicolinearidade nos coeficientes estimados em modelos de regressão múltipla.

Os dados, apresentados na Tabela 1, relativos aos anos de 1959 a 1968 incluem Gastos com Vestuário (G.V.), Ren da Disponível (R.D.), Ativo Líquido (A.L.), Índice de Preços para Vestuário (I.P.V.) e Índice Geral de Preços (I.G.P.).

O modelo estimado por mínimos quadrados é:

$$(G.V.) = -13.53 + 0,097(R.D.) + 0.015(A.L.) - 0,199(I.P.V.) + 0.34(I.G.P.), \quad (17)$$

e tem-se ainda que: $R^2 = 0,998$, $d = 3,4$

$$\Sigma \hat{Y}^2 = 28,15, \; \Sigma e^2 = 0,33,$$

$$F^* = \frac{\Sigma \hat{Y}^2/(K-1)}{\Sigma e^2/(n-K)} = \frac{28,15/4}{0,33/5} = 15,6,$$

(Johnston, 1963).

Como $F_{0,05}$, com $v_1 = K-1 = 5$ e $v_2 = n-K = 5$ graus de liberdade, é 5,19, rejeitou-se a hipótese nula, o que indica que existe um relacionamento significativo ($p<0,05$) entre os gastos com vestuário e as variáveis explanatórias.

TABELA 1

DADOS PARA ESTIMAÇÃO DA FUNÇÃO DEMANDA POR VESTUÁRIO

ANO	GASTOS COM VESTUÁRIO (£m) (G.V.)	RENDA DISPONÍVEL (£m) (R.D.)	ATIVO LÍQUIDO (£m) (A.L.)	ÍNDICE DE PREÇOS PARA VESTUÁRIO (1963 = 100) (I.P.V.)	ÍNDICE GERAL DE PREÇOS (1963=100) (I.G.P.)
1959	8,4	82,9	17,1	92	94
1960	9,6	88,0	21,3	93	96
1961	10,4	99,9	25,1	96	97
1962	11,4	105,3	29,0	94	97
1963	12,2	117,7	34,0	100	100
1964	14,2	131,0	40,0	101	101
1965	15,8	148,2	44,0	105	104
1966	17,9	161,8	49,0	112	109
1967	19,3	174,2	51,0	112	111
1968	20,8	184,7	53,0	112	111

Entretanto, as variáveis explanatórias são seriamente multicolineares, como pode ser visto na matriz de correlações abaixo:

	R.D.	A.L.	I.P.V.	I.G.P.
R.D.	1.	0.988315	0.980356	0.987666
A.L.	0.988315	1.	0.969962	0.969477
I.P.V.	0.980356	0.969962	1.	0.991796
I.G.P.	0.987666	0.969477	0.991796	1.

Desta forma o autor usa o "critério R^2", descrito por Neter and Wasserman (1974), página 376, e também considera a significância dos coeficientes estimados através de seus desvios padrões, concluindo pela escolha do modelo G.V. = f (R.D., I.P.V., I.G.P.) como o melhor (veja a Tabela 2); portanto, excluindo a variável A.L. do Modelo (17) por não apresentar coeficiente significativo.

O professor Koutsoyiannis faz considerações a respeito do acréscimo em R^2, quando são incluídas variáveis independentes até compor o Modelo (17), e verifica, para cada relação, a significância dos coeficientes de regressão.

b) A Figura 1 mostra o traço "ridge" para o exemplo. Conforme $K \to 1$ as instabilidades diminuem, sendo que, visualmente, para K=0,16 as instabilidades estão praticamente eliminadas. Neste caso, pode-se usar o seguinte modelo estimado:

$$(G.V.) = -20{,}2186+0{,}0340(R.D.)+0{,}0802(A.L.)+0{,}0967(I.P.V.)$$
$$+0{,}1674(I.G.P.) \qquad (18)$$

como alternativo ao Modelo (17). Observa-se pela Tabela 1 que as variáveis I.P.V. e I.G.P. apresentam dados muito próximos, altamente correlacionados, o que evidencia que numericamente deveriam ser uma única variável. A mudança de sinal do coeficiente da variável I.P.V. é justificada pela relação quase direta entre ela e a variável dependente - G.V. - (veja a Tabela 1). A variável R.D. é a mais estável conforme mostra a Figura 1, embora ligeiramente superestimada para K = 0 (solução de mínimos quadrados), enquanto a variável A.L. apresenta-se subestimada neste caso.

c) Considerando-se o critério (16) para $\Delta = 0{,}01$, seleciona-se como estimador "ridge":

$$(G.V.) = -18{,}5820+0{,}0466(R.D.)+0{,}0689(A.L.)-0{,}0107(I.P.V.)$$
$$+0{,}2200(I.G.P.). \qquad (19)$$

Neste caso, o coeficiente de I.P.V. é muito próximo a zero, sendo esta variável uma séria candidata a ser excluída do modelo. A forma correta para fazer tal exclusão seria considerar, em cada predição de G.V., o valor médio de I.P.V., que é equivalente a desconsiderar o efeito desta variável na predição de G.V.. A figura 2 mostra o traço "ridge" para o caso em que foi eliminada a variável A.L., conforme sugere o professor Koutsoyiannis. Nota-se que as instabilidades continuam, o que evidencia que o problema ainda é malcondicionado.

d) Os números de condição para os Modelos (18) e (19) foram 25,07 e 62,91, respectivamente e evidenciam a melhor estabilidade destas soluções.

6. CONCLUSÕES

Devido ao estimador "ridge" ser um estimador viciado, suas propriedades estatísticas são desconhecidas, e não se deve pensar em critérios tais como R^2 para comparar o estimador "ridge" e o de mínimos quadrados. Hoerl and Kennard (1970) mostraram que existe $K > 0$ inversamente proporcional aos coeficientes estimados de regressão (para $K=0$), tal que o erro médio quadrático para o estimador "ridge" é menor que para o estimador por mínimos quadrados. A determinação de tal valor de K tem preocupado pesquisadores (veja o trabalho de Wichern and Churchill (1978) e suas referências bibliográficas). Contudo, no aspecto de condicionamento das equações normais, a solução "ridge" é no mínimo mais estável que a de mínimos quadrados.

Deve-se observar que, para o exemplo na Seção 4, os resultados obtidos pela análise "ridge" refletem os dados apresentados na Tabela 1, não sendo considerados, neste trabalho, os aspectos de teoria econômica referente ao sinal dos coeficientes de regressão, possivelmente explicáveis à luz dos dados disponíveis.

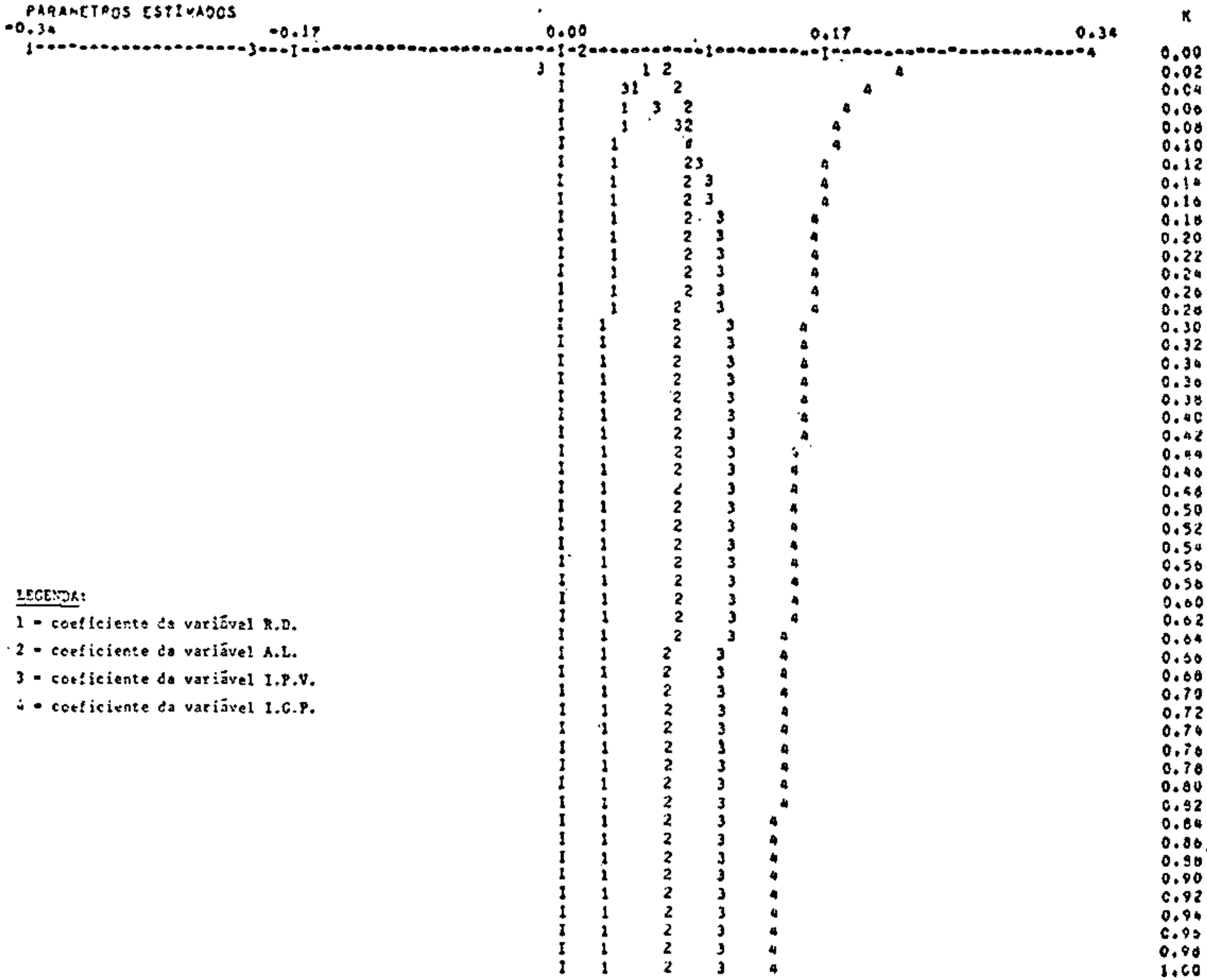

Fig. 1 - Traço "ridge" do modelo completo.

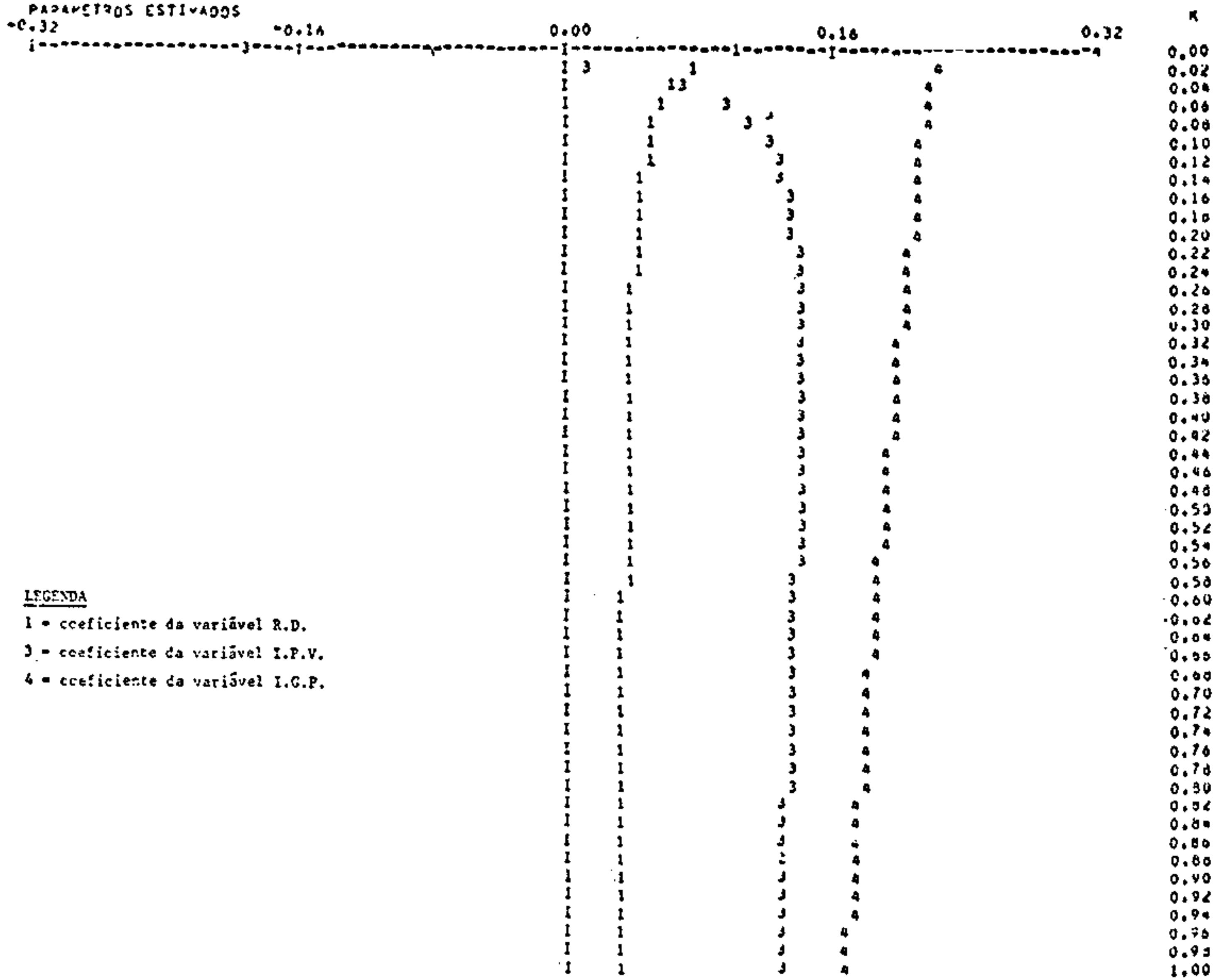

Fig. 2 - Traço "ridge" do modelo em três variáveis.

7. REFERÊNCIAS BIBLIOGRÁFICAS

HOERL, A.E.; KENNARD, R.W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12 (1): 55-67, Feb. 1970.

JOHNSTON, J. *Econometrics methods*. New York, NY, McGraw-Hill, 1963.

KOUTSOYIANNIS, A. *Theory of econometrics*. 2 ed. London, UK, The Macmillan, 1981.

NETER, J.; WASSERMAN, W. *Applied linear statistical models*. Homewood, IL, Richard D. Irwin, 1974.

SCHWARZ, H.R.; RUTISHAUSER, H.; STIEFEL, E. *Numerical analysis of symmetric matrices*. Englewood Cliffs, NJ, Prentice-Hall, 1973.

SEARLE, S.R. *Linear models*. New York, NY, John Wiley, 1971.

WICHERN, D.W.; CHURCHILL, G.A. A comparison of ridge estimators. *Technometrics*, 20 (3): 301-311, Aug. 1978.

WILKINSON, J.H. *The algebraic engenvalue problem*. Oxford, UK, Clarendon, 1965.