TRANSFORMAÇÃO DOS PARÂMETROS DE MODELOS LINEARES NÃO ORTOGONAIS, OBTIDOS PELO MÉTODO RIDGE TRACE, EM COEFICIENTES PRÁTICOS PARA ESTIMATIVA DE PRODUÇÃO [1]

IVAN BARBOSA MACHADO SAMPAIO [2]

SINOPSE.- A descodificação dos coeficientes de regressão obtidos pelo método Ridge Trace, para os casos em que $k > 0$, pode ser feita através de constantes específicas a cada um daqueles coeficientes e determinadas pela relação $C\beta_i$ = coeficiente de regressão normal/coeficiente de regressão padronizado, sendo ambos os termos calculados pelo processo dos quadrados mínimos ($k = 0$). O erro experimental introduzido pelo uso dessas constantes de descodificação $C\beta_i$ é insignificante e fica amplamente compensado pela facilidade de obtenção da equação de produção em unidades originais.

Programa auxiliar em anexo fornece subsídios para computação dos dados em linguagem FORTRAN.

Termos de indexação: Modelos lineares não ortogonais, método de Ridge Trace, estimativa de produção, variável padronizada, constante de descodificação.

INTRODUÇÃO

Em modelos lineares não ortogonais, freqüentemente se constata alta colinearidade entre as variáveis. Neste caso, o sistema formado nos conduz a estimativas inflacionadas dos coeficientes de regressão, ainda que pela solução dos quadrados mínimos. Essa inflação se manifesta em termos de valores absolutos, afetando principalmente as variáveis com elevado grau de correlação.

O método Ridge Trace proposto por Hoerl e Kennard (1970) tenta sucessivamente diminuir esse efeito inflacionário através de um artifício matemático sugerido pelo próprio Hoerl (1962). O resultado é traduzido em coeficientes de regressão mais adequados ao modelo adotado, já que a inflação causada pela colinearidade passa a ser parcialmente controlada. Os valores assim obtidos são tendenciosos sob o ponto de vista estatístico, mas se mostram mais apropriados ao modelo quando se tem interesse em generalizá-lo para outros grupos de dados similares.

Nas opções que o método Ridge Trace oferece como solução, haverá sempre uma considerada melhor pelo experimentador. Independente do problema criado na escolha da melhor opção, o pesquisador encontrará outro na descodificação dos coeficientes da solução escolhida. Isto porque aquele método estipula uma transformação das variáveis de tal ordem que, aliada às operações matriciais posteriores, torna difícil uma tentativa de retornar às unidades originais e obter assim uma equação mais objetiva e prática com fins estimativos.

Este trabalho apresenta um meio expedido para a descodificação dos coeficientes obtidos pelo método Ridge Trace.

MATERIAL E MÉTODOS

O procedimento prescrito pelo método Ridge Trace exige inicialmente a padronização das variáveis, do seguinte teor

$$x_i' = (x_i - \bar{x}_i) / \sqrt{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}$$

onde x_i' = variável padronizada, x_i = variável original, $\bar{x}_i$ = média da variável x_i e n = número de observações de x_i.

Esta padronização simplificará posteriormente a equação final, isenta do coeficiente linear, bem como apresentará as somas de produtos das variáveis sob forma de matriz de correlação. A solução

$$\begin{bmatrix} \text{Matriz de} \\ \text{correlação} \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{} x_{i1}' y_i' \\ \sum_{} x_{i2}' y_i' \\ \vdots \\ \sum_{} x_{ip}' y_i' \end{bmatrix} = \begin{bmatrix} \beta_1' \\ \beta_2' \\ \vdots \\ \beta_p' \end{bmatrix}$$

onde y_i' = iésima variável dependente padronizada, β_i' = coeficiente de regressão padronizado da variável i e p = número de variáveis independentes, é o sistema basicamente utilizado no Ridge Trace. Para destruir os efeitos de altas correlações ou mau condicionamento daquela matriz, o processo incorpora paulatinamente aos valores unitários de sua diagonal, constantes decimais entre 0,0 e 0,9 antes da inversão. Para cada valor k adicionado à diagonal, o sistema oferecerá uma solução. Estas perturbações infligidas à matriz não afetam significativamente a solução de um sistema, segundo Riley (1955), e a nova matriz assim formada é definitivamente mais bem condicionada que a original. Para $k = 0$, teremos a solução dos quadrados mínimos, mas

[1] Aceito para publicação em 20 de dezembro de 1976.

[2] Eng.º Agrônomo, M.Sc., professor do Dept.º de Zootecnia da Escola de Veterinária da U.F.M.G., Cx. Postal 567, 30.000, Belo Horizonte, MG.

com valores padronizados. A soma de quadrados devida ao erro experimental (SQE) será a menor entre todas as soluções. Para os valores seguintes de k, a SQE irá aumentando gradativamente, porém observa-se uma tendência à estabilização nos valores de β'_i. A melhor escolha será a opção que oferecer coeficientes mais estáveis possíveis, mas ainda com um valor da SQE aceitável. Sampaio (1972) apresenta variações gráficas dos coeficientes que, aliadas à ponderação do experimentador, auxiliam na escolha adequada.

O processo de descodificação para os coeficientes da solução escolhida segue basicamente o utilizado na análise da variância de dados padronizados, sempre do teor abaixo:

Fonte de variação	G. L.	S. Q.
Total	n−1	1,00
Regressão	p	R^2
Erro	n−p−1	$1-R^2$

Neste caso, para obtenção das S.Q. originais, basta multiplicar os valores daquela coluna pela soma de quadrados devido à variável dependente Y, ou seja, pela constante $C_r = \sum_{i=1}^{n} y_i - \bar{y})^2$.

Note-se que o valor codificado da SQ da regressão traduz o coeficiente de determinação R^2, o que, para este tipo de análise, já nos fornece uma informação prática sobre a conveniência do modelo.

Objetivando a constante de descodificação para cada coeficiente, procedemos aos seguintes cálculos, manipulando um grupo de oito variáveis com 48 observações cada[a]:

1) análise seletiva de variáveis pelo processo "stepwise" de regressão, que incorpora variáveis seqüencialmente ao modelo, obtendo a cada passo coeficientes sob as formas normal e padronizada; programa utilizado para esse fim foi fornecido pelo Manual de Sub-rotinas Científicas da IBM (1968);

2) determinação das constantes de descodificação para cada coeficiente, expressa pela fórmula $C\beta_i$ = coeficiente de regressão normal/coeficiente de regressão padronizado;

3) comparação, após a inclusão de cada variável ao modelo, da variação eventual dos $C\beta_i$, ocorrida em conseqüência daquelas inclusões;

4) avaliação da extensão do erro, introduzido pelo uso das constantes de descodificação, nos resultados obtidos pelo Ridge Trace; essa avaliação foi feita comparando-se os valores da SQE obtidos nos seguintes teores:

a) valor real, descodificado da análise da variância do Ridge Trace, traduzido pela fórmula

$$SQE = (1 - R^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 ; \qquad (1)$$

b) valor calculado, obtido a partir dos valores descodificados dos coeficientes, ou seja:

$$SQE = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y}). \qquad (2)$$

Resultados e Discussão

Cada coeficiente de regressão já existente no modelo pode variar bruscamente à medida que outras variáveis são incluídas na equação. Entretanto, sua constante de descodificação permanece a mesma, a despeito dessas inclusões. O Quadro 1 acusa uma variação bastante reduzida, causada pela truncagem operacional de casas decimais. Isso nos permitiria obter qualquer $C\beta_i$ em qualquer passo do "stepwise" que envolvesse a variável específica i.

Essas constantes, quando usadas nas soluções do Ridge Trace para $k > 0$, introduzem um novo erro na estimativa dos coeficientes originais, uma vez que são determinadas para $k = 0$. Entretanto, seguindo o mesmo argumento de Riley (1955) visto anteriormente, podemos utilizá-las para aqueles casos. O Quadro 2 mostra a extensão do erro introduzido. Neste estágio do trabalho já havíamos selecionado apenas três das sete variáveis independentes, justificando os valores de R^2 a partir de apenas 74%. Os valores da SQE no Ridge Trace se elevam juntamente com k, mas não nos deteremos nessa variação e sim nas estimativas de SQE para cada valor de k. A primeira coluna do Quadro 2 representa a soma de quadrados devida à regressão (SQR), correspondente ao segundo termo do lado direito da igualdade (2), e que fornece por subtração da soma de quadrados total, o valor da SQE na coluna seguinte.

O valor da SQE obtido pela equação (1), na terceira coluna, permite calcular a percentagem de erro incorporada no resíduo pelo uso das constantes de descodificação. Esse aumento percentual sobre o valor real é indicado na última coluna.

[a] Referentes ao estudo do grau de infestação de *Phythophtora infestans* em tomateiro, em função de variáveis climatológicas e algumas de suas interações.

Quadro 1. *Valores das constantes de descodificação em cada passo do Stepwise*

Variáveis	Número de variáveis no modelo						
	1	2	3	4	5	6	7
X_4	0,0245	0,0245	0,0245	0,0245	0,0245	0,0245	0,0245
X_2		0,4955	0,4961	0,4960	0,4960	0,4961	0,4959
X_1			0,2698	0,2698	0,2698	0,2698	0,2699
X_7				0,0036	0,0037	0,0038	0,0039
X_6					0,0640	0,0639	0,0639
X_3						0,5672	0,5670
X_5							0,0382

QUADRO 2. *Avaliação do erro introduzido com o uso das constantes de descodificação* [a]

K	Valor calculado		Valor real da SQE_c	R_2 (%)	$\frac{b - c}{c} \cdot 100$ (%)
	SQR	SQR_b			
0	55,70	19,28	19,26	74	0,08
1	49,61	25,37	25,35	66	0,05
2	45,41	29,57	29,56	61	0,03
3	42,16	32,82	32,79	56	0,08
4	39,53	35,45	35,42	53	0,06
5	37,32	37,66	37,64	50	0,03

[a] Todos os valores obtidos com cinco decimais e posteriormente truncados.

CONCLUSÕES

Para transformar os valores codificados dos coeficientes de regressão, obtidos pelo método Ridge Trace, em parâmetros com unidades originais, o uso de constantes de descodificação calculadas a partir da solução dos quadrados mínimos e de teor $C\beta_i$ = coeficiente de regressão normal/coeficiente de regressão padronizado, fornece uma solução expedita e prática na elaboração de um modelo estimativo. A elevação do erro experimental neste caso é insignificante e compensa as inconveniências de uma equação codificada.

O coeficiente linear será obtido da maneira usual, utilizando-se valores já descodificados, ou seja

$$a = \bar{y} - \sum_{j=1}^{p} \beta_j \bar{x}_j$$

AGRADECIMENTOS

Agradecemos à Eng.ª Agrônoma Lucila Marshall de Araújo, do Setor de Horticultura do IPEAME, pela cessão dos dados de epifitologia de *Phythophtora infestans* em tomateiro, sob condições climatológicas da região de Curitiba.

REFERÊNCIAS

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

Hoerl A.E. & Kennard R.D. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometric 12: 55-67.

IBM do Brasil 1968. Stepwise – statistical system. User's manual H 20-0333-1:7-30.

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.

Sampaio I.B.M. 1972. Controle inflacionário na determinação dos coeficientes de regressão em problemas envolvendo variáveis não ortogonais. Pesq. agropec. bras., Sér. Agron., 7: 65-69.

ABSTRACT.- Sampaio, I.B.M. [*Decoding the results of the Ridge Trace method for nonorthogonal linear models into practical predicting coefficients*]. Transformação dos parâmetros de modelos lineares não ortogonais, obtidos pelo método Ridge Trace, em coeficientes práticos para estimativa de produção. *Pesquisa Agropecuária Brasileira, Série Zootecnia* (1976), *11*, 57-63 [Pt, en] Univ. Fed. Minas Gerais, Cx. Postal 567, 30.000, Belo Horizonte, MG, Brazil.

The troublesome task of decoding the coefficients obtained from the Ridge Trace method for nonorthogonal models when $k > 0$ can be easily resolved by using the decoding factor $C\beta_i$ = normal coefficient/standardized coefficient, both terms calculated by the least square procedure. A non-significant amount of error is then introduced but recovering the coefficients in their original unities for a practical predicting equations surmounts this objection. The intercept of the final equation is obtained in the usual way.

A FORTRAN program on the method is added allowing quick data computation and further basis for decoding coeficients.

Index terms: Nonorthogonal linear models, Ridge Trace method, predictivy coefficients, standardized variable, decoding factor.

Anexo Computacional

Programação FORTRAN para o cálculo das constantes de descodificação.

1. Especificações técnicas

Linguagem: FORTRAN IV.

Capacidade mínima de memória exigida: 8 K.

Meio executor: computador IBM 1130.

Número máximo de variáveis: 15.

Número máximo de observações por variável: 50.

Sub-rotinas chamadas pelo programa: LOC, ARRAY, MATA, XCPY, MINV.

"Cores" exigidos: 2.948 para variáveis e 1.040 para o programa.

2. Especificações operacionais

2.1. Substituições de cartões:

Apenas dois cartões serão substituídos, de acordo com o tamanho da matriz C(I, J) utilizada pelo usuário:

2.1.1. Cartão de número de comando 2:

2 FORMAT ( ), que deverá conter entre os parênteses o formato específico utilizado para perfuração dos dados nos cartões.

2.1.2. Cartão de número de comando 6:

6 FORMAT((2X, F6.2)), onde o espaço reservado entre os dois primeiros parênteses deve conter o número total de variáveis estudadas, NV.

2.2. Leitura dos dados:

Após o último cartão do programa, //XEQ, segue-se o primeiro cartão de dados que identificará o número de repetições (NR) para cada variável (perfurado nas duas primeiras colunas) e o número total de variáveis (NV), perfurado nas duas colunas seguintes. Seguindo esse primeiro cartão, os demais fornecerão em fluxo contínuo as repetições seqüenciais para cada variável, devendo ser a última delas a variável dependente Y. Tal procedimento permitirá o registro total dos dados na matriz C(I, J), I = 1, NR e J = 1, NV.

2.3. Leitura das sub-rotinas científicas:

Se as sub-rotinas exigidas pelo programa não estiverem gravadas no disco em operação, deverão ser lidas antes da execução do programa. Essas sub-rotinas (LOC, ARRAY, XCPY, MATA, MINV) são encontradas nos manuais da IBM e normalmente já se encontram disponíveis nos centros de processamento sob forma de cartões perfurados.

2.4. Resultados impressos pelo programa:

Para as variáveis em unidades originais:

Matriz da soma de produtos corrigida,

Coeficientes de regressão,

Soma de quadrados devida à regressão,

Soma de quadrados do erro.

Para as variáveis codificadas (método Ridge Trace), k = 0,0 a 0,5:

Matriz de correlação,

Coeficientes de regressão,

Soma de quadrados devida à regressão,

Soma de quadrados do erro, e finalizando a impressão,

Médias das variáveis em unidades originais.

2.5. Manipulação dos resultados:

O acompanhamento da evolução dos valores dos coeficientes e da SQ do erro, na parte do controle inflacionário pelo método Ridge Trace, indicará um valor de k conveniente. Neste estágio, a descodificação de cada coeficiente de regressão poderá ser obtida pelo uso da constante proposta $C\beta_i$ = BETA I (estimativa para a solução dos quadrados mínimos)/BETA I (estimativa para k = 0,0). O valor do coeficiente linear.

$$a = \overline{Y} - \sum_{i=1}^{n} \beta_i \overline{X}_i$$

utilizará os valores descodificados dos coeficientes e as médias impressas no final do programa.

A matriz das somas de produtos corrigidos e a de correlação fornecem informações adicionais, seguindo a seqüência normal da leitura das variáveis (última coluna sempre referente a Y).

2.6. Apresentação do programa (cartões perfurados):

```
/ / FOR
*ONE WORD INTEGERS
*IOCS(CARD)
*IOCS(1132PRINTER)
*IOCS(DISK)
*LIST SOURCE PROGRAM
      DIMENSION C(50, 15), B(14, 14), SMVAR(15), XBAR(15), SQX(15), BETA(14) L(15),
     M(5), A(15, 15), R(15, 15)
      READ(2,1)NR,NV

    1 FORMAT (212)
      READ (2,2) ((C(I,J),J = 1, NV), I = 1, NR)

    2 FORMAT (9X, 3F4.0, 12X, F4.0)
      SELEC = 1.
      DO 10 J = 1, NV
      AUX 1 = 0.
      Do 10 I = 1, NR
      SMVAR (J) = AUX 1 + C (I, J)

   10 AUX 1 = SMVAR (J)
      Do 11 J = 1, NV

   11 XBAR (J) = SMVAR (J) / NR
      DO 12 J = 1, NV
      AUX 2 = 0.
      DO 12 I = 1, NR
      C (I, J) = C (I, J) — XBAR (J)
      SQX (J) = C (I, J) ** 2 + AUX 2

   12 AUX 2 = SQX (J)
      NV 1 = NV — 1
   64 CALL ARRAY (2, NR, NV, 50, 15, C, C)
      CALL MATA (C, A, NR, NV, 0)
      CALL XCPY (A, R, 1, 1, NV, NV, NV, NV, 1)
      CALL ARRAY (1, NV, NV, 15, 15, R, R)
      IF (SELEC-1.) 51, 51, 52

   51 WRITE (3, 53)

   53 FORMAT (1H1,' SOLUCAO PELO METODO DOS QUADRADOS MINIMOS EM
      UNIDADES
     1ORIGINAIS' /// 20X 'MATRIZ DAS SOMAS DE PRODUTOS CORRIGIDAS')
      GO TO 55

   52 WRITE (3, 54)

   54 FORMAT (//// '(CONTROLE INFLACIONARIO DOS COEFICIENTES DE REGRESSAO
     1PELO METODO RIDGE TRACE' /// 20X 'MATRIZ DE CORRELACAO'/)

   55 WRITE (3, 6) ((R(I, J), J = 1, NV), I = 1, NV)

    6 FORMAT ( 4(2X, F12.4))
      IF (SELEC-1.) 57, 57, 56

   56 DO 14 K = 1,6
      C1 = 0.1 * (K-1)
      DO 20 J = 1, NV1
      DO 20 I = 1, NV1

   20 B (I, J) = R (I, J)
      DO 15 I = 1, NV1

   15 B (I, I) = R (I, I) + C1
      GO TO 58

   57 DO 59 J = 1, NV1
      DO 59 I = 1, NV1
```

```
 59  B (I, J) = R (I, J)

 58  CALL ARRAY (2, NV1, NV1, 14, 14, B, B)
     CALL MINV (B, NV1, D, L, M)
     CALL ARRAY (1, NV1, NV1, 14, 14, B, B)
     DC 16 I = 1, NV1
     AUX 3 = 0.
     DO 16 J = 1, NV1
     BETA (I) = B (I, J) * R (J, NV) + AUX 3

 16  AUX 3 = BETA (I)
     AUX 4 = 0.
     DO 17 I = 1, NV1
     RSS = BETA (I) * R (I, NV) + AUX 4

 17  AUX 4 = RSS
     IF (SELEC-1.) 60, 60, 61

 60  SSE = SQX (NV) — RSS
     WRITE (3, 7)

  7  FORMAT (/// 20X 'ESTIMATIVAS PARA A SOLUCAO DOS QUADRADOS MINIMOS')
     GO TO 63

 61  SSE = 1.-RSS
     WRITE (3, 62) C1

 62  FORMAT (/// 20X 'ESTIMATIVAS PARA O CASO DE K =' F3.1)

 63  WRITE (3, 8) (I, BETA (I), I = 1, NV1)

  8  FORMAT (/ 10X 'BETA' 12' =' F11.5)
     WRITE (3, 9) RSS, SSE

  9  FORMAT (/ 8X' SQ DA REGRESSAO =' F14.5, 10X, 'SQ DO ERRO =' F14.5)
     IF (SELEC-1.) 71, 71, 14

 14  CONTINUE
     GO TO 72

 71  CALL ARRAY (1, NR, NV, 50, 15, C, C)
     DO 13 J = 1, NV
     DO 13 I = 1, NR

 13  C (I, J) = C (I, J) / SQRT (SQX(J))
     SELEC = 2.
     GO TO 64

 72  WRITE (3, 73) (XBAR(J), J = 1, NV)

 73  FORMAT (// 20X' MEDIAS DAS VARIAVEIS EM UNIDADES ORIGINAIS' /// 10 F11.5
    1//5F11.5)
     CALL EXIT
     END
//   XEQ
```

2.7. Observações:

2.7.1. O sétimo cartão * LIST SOURCE PROGRAM será omitido se não for desejada a impressão do programa em si, mas apenas os resultados.

2.7.2. A inclusão dos cartões das sub-rotinas, sempre que for o caso, deve ser feita entre os cartões //FOR e * ONE WORD INTEGER, do início do programa.

2.7.3. O programa é flexível quanto ao volume de variáveis e número de repetições, desde que, aumentada a capacidade de memória do computador, se redimensionem as matrizes operacionais do contexto do programa.

2.8. Exemplo dos resultados impressos pelo programa anterior:

SOLUCAO PELO METODO DOS QUADRADOS MINIMOS EM UNIDADES ORIGINAIS
MATRIZ DAS SOMAS DE PRODUTOS CORRIGIDOS

304.6657	25.9999	3857.3271	68.3332
25.9999	233.2500	2415.3730	45.1249
3857.3271	2415.3730	124219.3595	2600.6020
68.3332	45.1249	2600.6020	74.9791

ESTIMATIVAS PARA A SOLUCAO DOS QUADRADOS MINIMOS

BETA 1 = —0.08054
BETA 2 = —0.05040
BETA 3 = 0.02441
SQ DA REGRESSÃO = 55.71990 SQ DO ERRO = 19.25920

CONTROLE INFLACIONARIO DOS COEFICIENTES DE REGRESSAO PELO METODO

RIDGE TRACE

MATRIZ DE CORRELACAO

1.0000	0.0975	0.6270	0.4521
0.0975	0.9999	0.4487	0.3412
0.6270	0.4487	1.0000	0.8521
0.4521	0.3412	0.8521	1.0000

ESTIMATIVAS PARA O CASO DE K = 0.0

BETA 1 = —0.16236
BETA 2 = —0.08889
BETA 3 = 0.99383
SQ DA REGRESSÃO = 0.74313 SQ DO ERRO = 0.25686

ESTIMATIVAS PARA O CASO DE K = 0.1

BETA 1 = —0.04853
BETA 2 = —0.01534
BETA 3 = 0.80859
SQ DA REGRESSÃO = 0.66185 SQ DO ERRO = 0.33814

ESTIMATIVAS PARA O CASO DE K = 0.2

BETA 1 = 0.01153
BETA 2 = 0.02340
BETA 3 = 0.69533
SQ DA REGRESSÃO = 0.60572 SQ DO ERRO = 0.39427

ESTIMATIVAS PARA O CASO DE K = 0.3

BETA 1 = 0.04669
BETA 2 = 0.04596
BETA 3 = 0.61710
SQ DA REGRESSÃO = 0.56265 SQ DO ERRO = 0.43734

ESTIMATIVAS PARA O CASO DE K — 0.4

BETA 1 = 0.06849
BETA 2 = 0.05985
BETA 3 = 0.55880
SQ DA REGRESSÃO = 0.52757 SQ DO ERRO = 0.47242

ESTIMATIVAS PARA O CASO DE K = 0.5

BETA 1 = 0.08247
BETA 2 = 0.06862
BETA 3 = 0.51308
SQ DA REGRESSÃO = 0.49792 SQ DO ERRO = 0.50207

MÉDIAS DAS VARIAVEIS EM UNIDADES ORIGINAIS

7.66666 3.12500 72.27084 0.64583