

# SELEÇÃO DE VARIÁVEIS EM MODELOS LINEARES ORTOGONAIS<sup>1</sup>

IVAN BARBOSA MACHADO SAMPAIO<sup>2</sup>

**SINOPSE.**— Uma revisão nos métodos já existentes para seleção de variáveis é feita, buscando-se uma forma combinatória que venha ao encontro do interesse prático do experimentador. Baseados no controle do “bias” da equação reduzida e conseqüente redução de sua variância, os processos aqui empregados são extensivos a todos os tipos de regressão múltipla que, pelo grande número de variáveis ortogonais utilizadas, exigem seleção das mesmas para obtenção de uma equação de produção mais reduzida e prática.

## INTRODUÇÃO

O emprêgo de variáveis cujos valores podem ser determinados “a priori” conduz freqüentemente à formação de vetores ortogonais, onde o experimentador, para posterior facilidade de cálculo, simplesmente determina iguais intervalos entre os diversos níveis das variáveis independentes.

A experimentação agropecuária utiliza um número relativamente alto de variáveis ortogonais, acrescido na equação de produção pelos termos extras de interação. Sem dúvida, a minimização da soma de quadrados do erro e conseqüente obtenção de coeficientes “unbiased”, se processará estimando-se a equação completa. Muitas vezes isso não é desejável, não só pelo volume de cálculos exigidos, como também pela impraticabilidade de uma equação muito extensa.

O interesse do experimentador e objetivo deste trabalho é determinar uma equação reduzida que reúna apenas as variáveis mais importantes, mantendo, entretanto, um poder estimativo semelhante ao da equação completa.

## MATERIAL E MÉTODOS

Para ilustrar os métodos que iremos discutir, lançamos mão dos resultados obtidos na Fazenda Experimental de Matão, São Paulo, no primeiro trimestre de 1968. Estudamos na ocasião o efeito dos nutrientes limitantes da produção de sorgo (*Sorghum vulgare*), N e P, em solo de cerrado. O experimento foi executado em casa de vegetação, em esquema fatorial  $5 \times 5$  completamente casualizado, com 4 repetições, sendo as dosagens, para ambos os elementos, 0, 100, 200, 300 e 400 kg/ha. Adubação básica de potássio e micronutrientes foi adicionada a todos os vasos e os dados aqui utilizados são os referentes ao péso seco de 16 plantas por vaso, colhidas rente ao solo, aos 27 dias.

A equação completa de produção, no caso, envolveria 25 termos, ou seja, 24 coeficientes de regressão. Para se reduzir o número de variáveis na equação a um

grupo considerado melhor, é necessário que se adote um critério que caracterize a condição de “melhor grupo”.

Beale *et al.* (1967) definiram o melhor conjunto de variáveis como sendo o que maximiza o valor da correlação múltipla entre as variáveis selecionadas e a variável dependente Y.

Realmente, Garside (1965) havia citado que, para um determinado número r de variáveis, se tal valor fôsse alcançado, a soma de quadrados residual correspondente seria a menor de todas as das demais combinações de r variáveis. Este, então, seria o melhor grupo de r variáveis.

Para a obtenção de tal grupo, Draper e Smith (1966) apresentaram os métodos distintos de eliminação e incorporação seqüencial, ambos baseados nos coeficientes de correlação da variável dependente com as demais. No primeiro método parte-se da equação completa, eliminando-se a variável menos importante toda vez que o valor do teste F parcial correspondente fôr inferior a um valor F<sup>0</sup> pré-determinado. Na incorporação seqüencial, introduz-se uma variável de cada vez, partindo-se da de coeficiente de correlação mais alto e testando-se seu respectivo F parcial, bem como os de todas as variáveis previamente incorporadas. O processo incorpora a variável cujo F ultrapassa F<sup>0</sup>, também um valor pré-determinado, e elimina toda variável que posteriormente se torna insignificante.

Embora ambos os métodos sigam critérios similares de seleção, quando aplicados ao mesmo conjunto de dados, podem conduzir a diferentes combinações de variáveis. Além disso, mesmo utilizando-se apenas um dos métodos, pode-se encontrar dois ou mais grupos igualmente qualificados, a um mesmo nível de significância.

Sem dúvida, os métodos de eliminação e incorporação seqüenciais são uma solução mais prática que o cálculo das  $2^n - 1$  equações possíveis, com a posterior escolha do melhor grupo de r variáveis. Entretanto, cumpre observar que para cada valor de r haverá uma equação considerada “melhor”. Como um número relativamente alto de “melhores grupos” pode ser obtido, deve-se adotar um critério que indique a melhor equação entre as “melhores”.

Mallows (1964) sugeriu que este critério fôsse o quadrado do erro total padronizado e desenvolveu a estatística  $C_p$  para caracterização do mesmo, cuja determinação se obtém facilmente, como se segue.

<sup>1</sup> Recebido 22 jan. 1971, aceito 26 mar. 1971.

<sup>2</sup> Eng.º Agrônomo, M.Sc., do Setor de Estatística Experimental e Análise Econômica do EPE, 9.º andar, Ministério da Agricultura, Brasília, DF., e bolsista do Conselho Nacional de Pesquisas (CNPq 14450/70).

Em uma equação de  $p$  termos, estimada de  $N$  dados o quadrado do erro total (erro ao acaso + bias) é

$$\sum_{i=1}^N (v_i - \eta_i)^2 + \sum_{i=1}^N \text{Var} [\tilde{Y}_{pi}] \quad (1)$$

onde  $v_i = v(X_{1i}, X_{2i}, \dots)$  é o valor esperado do vetor das variáveis, obtido da equação real,

$\eta_i = \sum_{j=1}^p \beta_j X_{ji}$  é o valor esperado do vetor das variáveis, obtido da equação utilizada, e

$\text{Var} [\tilde{Y}_{pi}]$  é a variância da estimativa de  $Y_i$  através da equação de  $p$  termos.

Portanto,  $v_i - \eta_i$  é, por definição, o "bias" do  $i$ ésimo dado. Chamando-se à soma de quadrados, devido ao "bias", SQB, e ao quadrado do erro total padronizado,  $\Gamma_p$ , a equação (1) torna-se

$$\Gamma_p = \frac{\text{SQB}}{\sigma^2} + \frac{1}{\sigma^2} \sum_{i=1}^N \text{Var} [\tilde{Y}_{pi}] \quad (2)$$

Mas Gorman e Toman (1966) mostram que

$$\sum_{i=1}^N \text{Var} [\tilde{Y}_{pi}] = p\sigma^2$$

e, portanto,

$$\Gamma_p = \frac{\text{SQB}}{\sigma^2} + p \quad (3)$$

O valor esperado da soma de quadrados do resíduo de uma regressão de  $p$  termos é, por sua vez,

$$E(\text{SQR}_p) = \text{SQB} + (N-p)\sigma^2, \text{ ou}$$

$$\text{SQB} = E(\text{SQR}_p) - (N-p)\sigma^2, \text{ que substituindo em (3),}$$

$$\Gamma_p = \frac{E(\text{SQR}_p)}{\sigma^2} - (N-2p)$$

Com uma boa estimativa de  $\sigma^2$ ,  $C_p$  torna-se a estimativa de  $\Gamma_p$ , ou seja,

$$C_p = \frac{\text{SQR}_p}{\sigma^2} - (N-2p) \quad (4)$$

Se uma equação de  $p$  termos possuir um "bias" desprezível, SQB é aproximadamente zero e  $\text{SQR}_p$  é estimada por  $(N-p)\sigma^2$  apenas. Neste caso o valor de  $C_p$  torna-se

$$C_p = \frac{(N-p)\sigma^2}{\sigma^2} - (N-2p)$$

Entretanto  $\sigma_p^2$  é aproximadamente igual a  $\sigma^2$ , se considerarmos a homogeneidade da variância e

$$C_p \approx (N-p) - (N-2p) \cdot$$

$$C_p \approx p.$$

O critério de escolha sugerido por Mallows (1964) esteia-se no fato de que o grupo de variáveis, cujo  $C_p$  estiver mais próximo do número de termos da equação reduzida correspondente,  $p$ , será o conjunto que apresentará menor "bias" e, portanto, melhor combinação das variáveis.

Para utilização desse método é necessária uma estimativa de variância,  $\sigma^2$ , que pode ser obtida através do cálculo do quadrado médio do erro para a equação completa. Nesta ocasião, o teste  $t$  para cada coeficiente pode ser facilmente obtido, e uma ordenação decrescente das variáveis com base no teste  $t$  dará uma idéia da importância de cada uma delas. Evidentemente, aquelas que forem significantes entrarão obrigatoriamente na equação reduzida final. A necessidade de inclusão de algumas das não significantes só será respondida através do cálculo dos  $C_p$ 's respectivos. Muitas vezes a relação  $C_p = p$  é alcançada utilizando-se apenas as variáveis realmente significativas. Outras vezes, a inclusão das demais, uma a uma, ou em combinações adequadas, é necessária para se conseguir uma equivalência ideal. De qualquer modo, o número de grupos a serem pesquisados será sempre muito mais reduzido que as  $2^n - 1$  equações possíveis, combinando-se  $n$  variáveis.

## RESULTADOS E DISCUSSÃO

As produções médias obtidas podem ser vistas no Quadro 1 e a estimativa de  $\sigma^2$  para a equação completa foi 98,91. Nessa ocasião foi executado o teste  $t$  para os coeficientes e estes ordenados segundo a magnitude do teste. Esta ordenação pode ser vista no Quadro 2, onde também figuram os valores de  $C_p$  e  $p$  após a inclusão sucessiva de cada variável. O valor de  $C_p$  decresce, à proporção que novas variáveis vão sendo incluídas, até um certo ponto, para depois aumentar gradativamente até a perfeita igualdade  $C_p = p$  da equação completa. Dentro desse campo de variação há muitas oportunidades de se obterem valores aproximados para  $C_p$  e  $p$ . Note-se que o valor de  $p$  inclui sempre o termo  $b_0 X_0$ . Este método de inclusão seqüencial baseada no teste  $t$  foi proposto por Gorman e Toman (1966).

A exclusão de qualquer variável significativa fará o valor de  $C_p$  destoar do de  $p$  correspondente. Portanto, estas variáveis obrigatoriamente deverão estar presentes na

QUADRO 1. Produção média dos tratamentos (gramas de peso seco)

Doses de N	Doses de P				
	P <sub>0</sub>	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>
N <sub>0</sub>	9,8	44,8	54,3	45,5	52,3
N <sub>1</sub>	8,0	56,5	83,9	86,3	88,0
N <sub>2</sub>	8,3	48,8	91,0	110,0	97,5
N <sub>3</sub>	8,5	48,3	101,8	108,8	119,0
N <sub>4</sub>	8,0	38,5	91,3	121,0	126,0

QUADRO 2. Busca seqüencial de  $C_p$  pelo teste  $t$

Variável	t	$C_p$	p
P'	31,86*	471,5	2
P''	12,01*	329,2	3
N'	11,88*	189,7	4
N'P'	11,16*	67,0	5
N''	5,66*	37,0	6
N'P''	4,08*	22,3	7
N''P''	2,62*	17,4	8
P'''	2,49*	13,2	9
N'P'''	2,38*	9,5	10
N'''P'''	1,98	7,6	11
N''''	1,55	7,1	12
N''''P''''	1,24	7,6	13

\* Significativo pelo menos a 5%.

equação reduzida. No nosso caso, o Quadro 2 revela que à inclusão das nove primeiras variáveis ( $p = 10$ ) corresponde um  $C_p$  de 9,5, sendo esses dois últimos valores os mais próximos possíveis entre as colunas de  $C_p$  e p.

Se representarmos graficamente os valores de  $C_p$  no eixo vertical e de p no horizontal, a bissetriz do quadrante formado será a linha  $C_p = p$ , como pode ser visto na Fig. 1. A representação dos valores de  $C_p$  de todas as  $2^n - 1$  equações possíveis daria certamente uma idéia geral dos melhores grupos de variáveis, que estariam representados por pontos incrustados na bissetriz. Entretanto, a determinação de todas essas equações foge ao objetivo principal deste trabalho. Desta maneira, a busca seqüencial utilizada no Quadro 2 oferece uma excepcional oportunidade de redução de cálculos.

No nosso caso, apenas a inclusão das variáveis significativas parece suficiente, mas se houvesse um meio de igualar os valores de  $C_p$  e p, para uma equação reduzida, obteríamos uma equação melhor e mais representativa por ser "unbiased".

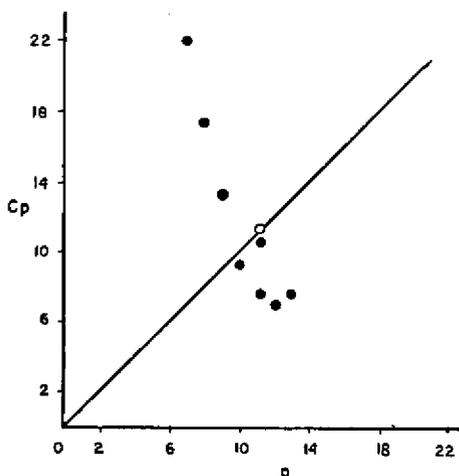


FIG. 1. Representação dos valores de  $C_p$  para algumas equações reduzidas.

O critério adotado pelo experimentador e o conhecimento que este tiver sobre as variáveis poderão auxiliá-lo na escolha acertada das demais variáveis a serem incluídas. O trabalho combinatório levará à determinação de pontos, como na Fig. 1. Aqui, todos os pontos representam equações reduzidas que incluem todas as variáveis significativas mais combinações das demais, a critério do experimentador.

Pontos situados fora da bissetriz conterão "bias" à proporção que se afastam da mesma, e sendo  $C_p$  função do quadrado do erro total, este aumentará com o afastamento do ponto em relação ao eixo horizontal. Portanto, a inclusão de variáveis pode diminuir o "bias" da equação, mas, por outro lado, aumenta gradativamente a variância.

Como a soma de quadrados devida a cada variável é um dado necessariamente calculado para execução do teste t e estimativa de  $\sigma^2$ , sugerimos que o processo combinatório se faça por um método racional lógico que abrevie o trabalho.

Em nosso caso, para  $p = 10$ , o valor de  $C_p$  foi 9,5. A inclusão da próxima variável mais importante, N'P', elevou p a 11, mas reduziu o respectivo  $C_p$  a 7,6. Evidentemente não é compensador incluí-la na equação, que teria um "bias" substancial. Se o nosso interesse fôsse colocar alguma outra das variáveis restantes, obtendo  $C_p = p$ , o novo p seria ainda igual a 11. Utilizando os valores conhecidos de  $\sigma^2$ , a soma de quadrados do resíduo para a equação das variáveis significativas e a condição de  $p = 11$ , teremos, segundo a equação (4),

$$\frac{SQR_{vs} - SQ_i}{98,91} - (100 - 2 \times 11) = 11,$$

onde  $SQR_{vs} = 8857,77$ , correspondendo à soma de quadrados do resíduo para a equação contendo apenas as variáveis significativas. Devido à ortogonalidade condicional do problema, esse valor corresponde ao somatório das SQ devidas a todas as variáveis significativas, subtraído da SQ total.

$SQ_i$  é a soma de quadrados devida à variável i a ser incluída, incógnita a ser resolvida para a nossa conveniência.

Resolvendo para  $SQ_i$ , temos  $SQ_i = 54,78$ .

Uma rápida busca nos valores anteriormente calculados das somas de quadrados para as variáveis, revela que o valor mais próximo à  $SQ_i$  foi 51,00, correspondente ao efeito cúbico de P. Tal efeito foi dos menos importantes na análise estatística, porém, sua inclusão na equação, agora com 11 termos, conduz a um  $C_p$  de exatamente 11. A nova equação determinada satisfaz nossas condições, por ser "unbiased" e possuir um número reduzido de variáveis. A nova estimativa da variância será  $(SQR_{vs} - 51,00) : 89 = 98,95$ , sendo 89 o novo valor dos graus de liberdade do resíduo. O aumento da variância em relação à equação completa foi de 0,04 apenas.

Este exemplo, entretanto, mostrou-se muito conveniente, conduzindo-nos à uma solução quase imediata. Temos encontrado casos em que a igualdade  $C_p = p$  é dificilmente conseguida e valores aproximados têm que ser adotados. Aqui torna-se mais flexível a seleção de novas variáveis, pois certamente serão necessárias mais de uma delas, convenientemente escolhidas, para se alcançar bons valores aproximados de p e  $C_p$ . O raciocínio para a re-

solução do problema continua o mesmo. O valor de  $SQ_1$  na fórmula será a soma de 2, 3, r valores da soma de quadrados das 2, 3, r variáveis a serem escolhidas. Cabe ao experimentador selecionar, num critério pessoal justificável, as variáveis adequadas cujas SQ somadas alcancem o valor calculado para  $SQ_1$ . Também pode ocorrer que duas ou mais equações reduzidas estejam habilitadas a bem representar a completa. O gráfico constitui excepcional auxílio para decisões entre estas equações que, dependendo da localização dos pontos pelos quais estão representadas, indicarão a magnitude de seus "bias" e variância.

## REFERÊNCIAS

- Beale, E.M.L., Kendall, M.G. & Man, D.V. 1967. The discarding of variables in multivariate analysis. *Biometrika* 54:337-366.
- Draper, N.R. & Smith, H. 1966. *Applied regression analysis*. Wiley, New York.
- Garside, M.J. 1965. The best subset in multiple regression analysis. *Appl. Statist.* 14:196-200.
- Gorman, J.W. & Toman, R.J. 1966. Selection of variables for fitting equation to data. *Technometrics* 8:27-51.
- Mallows, C.L. 1964. Choosing variables in a linear regression: a graphical aid. Central Regional Meeting of the Institute of Mathematical Statistics em Manhattan, Kansas. (Mimeo., não publicado)

ABSTRACT. — Sampaio, I.B.M. 1972. *Selecting variables in orthogonal linear models*. *Pesq. agropec. bras., Sér. Agron.*, 7:71-74 (Escrit. Pesq. Exp., Min. Agricultura, 9.º andar, Brasília, DF, Brazil)

Reducing the number of variables of an orthogonal problem may lead to a more practical predicting equation. Based on the standardized total squared error, or the  $C_p$  statistic, this can be attained through several methods. A direct search in the sums of squares due to each variable may save computational work in the combinatory trials required by the  $C_p$  statistic method. The result is a reduced equation yet as unbiased as possible.