Revista Ceres 31(175):215-229. 1984.

APLICAÇÃO DO MÉTODO DA REGRESSÃO DE CUMEEIRA A UMA FUNÇÃO DE PRODUÇÃO DE LEITE NA ZONA DA MATA DE MINAS GERAIS[1]

Orlando Monteiro da Silva [2]
José Carlos Ribeiro [3]
Albino Sérgio Dias Casali [2]

1. INTRODUÇÃO

A teoria estatística da análise de regressão tem sido utilizada com maior freqüência para testar ou aplicar a teoria econômica.

Na estimação dos coeficientes das funções de regressão, o método mais comumente empregado é o dos mínimos quadrados ordinários, que apresenta as melhores características estatísticas dos coeficientes estimados, tais como consistência, eficiência e não-tendenciosidade. Na utilização desse método deve ser estabelecido o seguinte conjunto de pressuposições:

a) $Y_i = \alpha + \beta X_i + \mu_i$;
b) $E(Y_i/X_i) = \alpha + \beta X_i$ para todo i;
c) VAR $(Y_i/X_i) = \delta^2\mu$;
d) COV $(\mu_1, \mu_j) = 0$ para i = *j*;
e) A variável X permanece fixa em observações sucessivas.

Para o modelo de regressão linear múltiplo, deve-se acrescentar mais um pressuposto;

f) Nenhuma das variáveis independentes deve estar perfeitamente correlacionada com outra variável independente ou com qualquer outra combinação linear de variáveis independentes (*14*).

No entanto, algumas dessas pressuposições não se verificam em muitas situa-

---

[1] Recebido para publicação em 28-3-1984.

[2] Departamento de Administração e Economia da U.F.V. 36570 Viçosa, MG.

[3] Departamento de Engenharia Florestal da U.F.V. 36570 Viçosa, MG.

ções, o que torna necessária a aplicação de outras técnicas de estimação. Um dos problemas comumente encontrados na estimação dos coeficientes de uma regressão, principalmente na área agrícola, é o da alta correlação entre duas ou mais variáveis explicativas. Esse problema é tratado na literatura especializada como multicolinearidade e traz como conseqüências graves perturbações para os estimadores e até a impossibilidade de estimação dos coeficientes da função.

Diz-se que há um problema de forte ou fraca multicolineeridade na análise de regressão se há forte ou fraca correlação linear entre duas variáveis independentes ou entre uma delas e as demais (*19*). Pode-se concluir, então, que multicolinearidade é uma questão de grau, e não de natureza, pois geralmente ocorrerá entre dados de qualquer estudo.

Na análise dos coeficientes de uma função de regressão pelo método dos mínimos quadrados ordinários, a matriz (X'X) invertida é de grande importância na determinação dos $\hat{B}'_i$, desde que $\hat{B}_i = (X'X)^{-1}.\ X'Y$, e de suas variâncias.

Com perfeita multicolinearidade, o determinante de (X'X) é nulo e torna indeterminado o sistema. Na prática, casos como este ocorrem diferentemente, verificando-se, com maior freqüência, situações de forte multicolinearidade, fazendo com que o determinante da matriz (X'X) seja numericamente pequeno. Como as variâncias dos $\hat{B}'_i$ são estimados a partir dos elementos da diagonal principal da matriz $\left[S^2.\ (X'X)^{-1}\right]$ (sendo S^2 uma estimativa da variância da regressão), estas terão valores elevados, fazendo com que os testes de hipóteses sobre os $\hat{B}'_i$ populacionais não sejam confiáveis.

Segundo JOHNSTON (*12*), tal fato leva a estimativas imprecisas dos $\hat{B}'_i$, com grande sensibilidade a pequenas variações nos dados-base da regressão, o que torna difícil isolar a influência relativa dos diversos X's, ocasionando dificuldades na interpretação do modelo.

Para efeito de previsão, desde que a estrutura dos dados não se tenha modificado sensivelmente no período de previsão, a presença de multicolinearidade não traz maiores prejuízos ao modelo, visto que a imprecisão dos $\hat{B}'_i$ não interfere necessariamente, como um possível bom ajustamento global (medido pelo R^2) do modelo. Apesar dos problemas causados pela multicolinearidade, os estimadores de mínimos quadrados ordinários permanecem não-tendenciosos (*8*).

Para corrigir os problemas de multicolinearidade, têm sido propostos:

a) aumento do tamanho da amostra;
b) abandono da variável mais atingida;
c) combinação de dados de «Cross-Section» com dados de série temporal;
d) uso de componentes principais,
e) uso de «Ridge Regression» (regressão de cumeeira).

Esta última técnica, desenvolvida por HOERL e KENNARD (*6*), é de uso recente e produz estimadores tendenciosos, porém dotados de variâncias menores que os produzidos pelo método dos mínimos quadrados ordinários. Sua grande vantagem é gerar erros quadrados médios (tendência ao quadrado + variância) menores que os obtidos com os mínimos quadrados ordinários.

MADALLA (*15*) afirma que o método da regressão de cumeeira é olhado de modo diferente por estatísticos e economistas. Estes últimos sugerem que o artifício utilizado no método seria puramente matemático, sem considerar os aspectos econômicos da relação que vai ser estimada.

De acordo com DRAPER e SMITH (*6*), há muita controvérsia no uso do método da regressão de cumeeira, mas recomendam a utilização do método para o tratamento do problema da multicolinearidade em duas situações: na formulação de um problema de regressão, se tivesse conhecimento «prévio» das possíveis magnitudes dos parâmetros e se o problema fosse tomado como um problema de míni-

mos quadrados, sujeito a algum tipo específico de restrição de parâmetros.

No entanto, diversos trabalhos têm utilizado o método da regressão de cumeeira (*1, 2, 3, 13, 16, 17, 18*). BROWN e BEATTLE (*3*) e KALIL (*13*) mostraram que esse método fornece resultados melhores quando aplicado a funções de produção e, talvez por esse motivo, a maioria dos trabalhos utiliza essas funções (*1, 3, 13, 18*), sem, contudo, deixar de ter aplicação em modelos de mercado (*2, 16, 17*).

O grande número de trabalhos realizados, assim como a introdução da teoria da regressão de cumeeira em livros técnicos (*5, 6*), sugere a importância do método na correção dos problemas de multicolinearidade encontrados em numerosas funções econômicas.

O objetivo geral do presente trabalho foi aplicar o método da regressão de cumeeira a uma função de produção de leite para a Zona da Mata de Minas Gerais. Especificamente, pretendeu-se analisar a coerência teórica dos resultados encontrados com a aplicação da regressão de cumeeira e compará-los com os resultados obtidos por CASALI (*4*), que utilizou o método dos mínimos quadrados ordinários.

2. METODOLOGIA

2.1. *Os Dados*

Os dados deste trabalho foram os mesmos utilizados por CASALI (*4*) na estimação de uma função de produção de leite para a Zona da Mata de Minas Gerais, nos períodos da seca e das águas. Referem-se a valores físicos da produção de leite e dos fatores de produção empregados na pecuária leiteira em 65 propriedades da região, no ano agrícola 1977/78. Os dados são primários e foram coletados por extensionistas da Empresa de Assistência Técnica e Extensão Rural do Estado de Minas Gerais, com a participação do Centro Nacional de Pesquisa de Gado de Leite (CNPGL), com o objetivo de comparar os resultados obtidos nessas fazendas com os alcançados no modelo experimental daquele centro. Visando aferir a eficiência do setor no uso de recursos, CASALI (*4*) estimou funções de produção, na forma Cobb-Douglas, com esses dados, e deparou com problemas de multicolinearidade entre as variáveis explicativas área de propriedade e número de vacas em lactação.

Os resultados encontrados e uma análise desses dados são encontrados no item resultados e discussão deste trabalho.

2.2. *O Método da Regressão de Cumeeira*

Em notação matricial, um modelo de regressão linear múltiplo pode ser apresentado do seguinte modo:

$$Y = XB + U$$

sendo Y um vetor com os valores da variável dependente, X uma matriz dos valores das variáveis independentes, com dimensão N x P, B um vetor dos parâmetros que vão ser estimados e U um vetor dos erros. Admite-se que $E(U) = 0$ e que $E(UU') = I_n\sigma^2U$, sendo σ^2U a variância dos erros e I_n uma matriz identidade.

Satisfeitas essas condições, o método dos mínimos quadrados ordinários for-

nece os melhores estimadores lineares não-tendenciosos dos coeficientes B, que são obtidos por meio da expressão:

$$\hat{B} = (X'X)^{-1}.X'Y$$

Pelo método da regressão de cumeeira, costuma-se utilizar as variáveis padronizadas, ou seja:

$$Z_i = \frac{Yi - \bar{Y}}{S(Y)} \text{ e } V_i = \frac{Xi - \bar{X}}{S(X)}$$

sendo S(Y) e S(X) os desvios-padrão de Y e X, respectivamente.

Dessa forma evita-se a influência das unidades de medida das variáveis e as matrizes X'X e X'Y tornam-se matrizes de correlações simples. Então, como o método de regressão de cumeeira tem por objetivo diminuir a multicolinearidade, o estimulador de cumeeira é obtido pelo aumento dos elementos da diagonal principal da matriz X'X, na forma de matriz-correlação, por pequenas quantidades (K). O estimador torna-se, então:

$$B^* = (X'X + KIp)^{-1}.X'Y$$

Como X'X é uma matriz de correlações simples, só se consideram valores de K para o intervalo $0 \leq K \leq 1$. O valor de K escolhido será um valor arbitrário nesse intervalo; o ideal seria utilizar um K para o qual a esperança do quadrado do desvio fosse mínima.

Se K for igual a zero, B* = B, e a estimativa de cumeeira será igual à dos mínimos quadrados. Para cada valor de K utilizado, obtêm-se novos coeficientes B* (K), de forma que se pode, com eles, traçar um gráfico conhecido como «gráfico de cumeeira». Tal gráfico é a base para a escolha do nível K com o qual se obterá a regressão estimada. De posse dele, podem-se analisar os efeitos da multicolinearidade sobre as estimativas dos parâmetros a partir de um grupo de dados. Em seus primeiros estudos sobre a regressão de cumeeira, HOERL e KENNARD (*9, 10*) sugeriram a escolha do valor de K com o uso do método gráfico, observando os seguintes aspectos: a) o valor de K deve coincidir com o ponto a partir do qual as estimativas dos parâmetros são relativamente estáveis; b) os sinais dos coeficientes, a partir desse ponto, deverão permanecer inalterados; c) o valor da soma dos quadrados dos resíduos da regressão não deverá ter sofrido aumento excessivo nesse nível, em relação ao nível K = 0. Por ser esse método de escolha, um tanto subjetivo, os mesmos autores, em artigo publicado em 1975 (*11*), sugerem como valor de K o resultado da aplicação da seguinte fórmula:

$$K^* = rs^2 / \left\{ B^*(0) \right\}' \left\{ B^*(0) \right\}$$

sendo — r o número de parâmetros do modelo, sem contar o intercepto;
— S^2 o quadrado médio residual, obtido do quadro de análise de variância no ajustamento por mínimos quadrados e

$$\left\{ B^*(0) \right\} = \left\{ B_1^*(0), B_2^*(0), ..., B_r^*(0) \right\}$$

Para a escolha do valor do K «ótimo», VINOD (*20*) propôs uma nova estatísti-

ca, a que chamou de Índice de Estabilidade das Magnitudes Relativas, definida como

$$IEMR = \sum_i [(p\sigma_i^2/\overline{S}\lambda_i - 1]^2$$

sendo p = o número de parâmetros, λ = os valores característicos da matriz X'X, $\delta_i = \lambda_i/(\lambda_i + K_i)^2$ e $S = \sum_i \lambda_i + K)^2$

que será igual a zero para um sistema completamente ortogonal. Assim, o valor de K que produzir o menor IEMR fará com que o sistema chegue o mais próximo possível da ortogonalidade.

No presente trabalho, o método da regressão de cumeeira, assim como as fórmulas para a escolha do valor de K, foi empregado no modelo conhecido como função Cobb-Douglas. A função mostra a dependência da produção aos fatores por meio da expressão

$$Y = B_0 X_1^{B_1} . X_2^{B_2} ... X_n^{B_n}$$

Essa expressão se torna linear com a aplicação de logaritmos:

$$\log Y = \log B_0 + B_1 \log X_1 + B_2 \log X_2 + ... + B_n \log X_n,$$

sendo Y = variável dependente (produção)
B_0 = constante
B_i = coeficientes de regressão
X_i = variáveis independentes (fatores)

Supondo que os erros tenham distribuição normal, com E(U) = 0 e $E(UU') = I_n\sigma^2 U$, o estimador B apresentará distribuição multinormal e a aplicação dos testes de significância comumente usados em análise de regressão será válida. Os estimadores de cumeeira são tendenciosos, e os valores de t e F calculados da forma usual não podem ser utilizados, aqui, como testes de significância. Os valores de t devem ser interpretados apenas como medidas descritivas, que mostram quanto o valor da estimativa do parâmetro é maior que o valor do respectivo desvio-padrão. O R^2, como medida descritiva, continua válido para o método da regressão de cumeeira. A demonstração e a discussão das propriedades dos estimadores de cumeeira podem ser encontradas no trabalho de KALIL (*13*).

3. RESULTADOS E DISCUSSÕES

Antes da análise dos resultados obtidos na estimação pelo método da regressão de cumeeira, optou-se pela apresentação dos resultados obtidos por CASALI (*4*) e discussão dos problemas encontrados.

Naquele estudo estavam disponíveis, a princípio, 13 variáveis, que expressavam fatores de produção de leite. Em ajustamentos preliminares, foram eliminadas as variáveis mão-de-obra, estoque de capital e juros de empréstimos agropecuários; agregaram-se as variáveis capim e cana picada (volumoso) e sal comum e minerais (minerais totais), restando para a análise da produção as seguintes: número de vacas em lactação, área em pastagem, silagem, concentrado, grau de sangue, capim e cana picada, minerais totais e gastos com sanidade.

Utilizando o programa SPSS e o processo de Análise de Regressão por Seqüência (Stepwise), em que a inclusão das variáveis no modelo resulta da sua im-

portância «estatística» (correlações parciais) na explicação da variável dependente, obteve-se, naquele estudo, a seguinte ordem de entrada das variáveis nas equações:

Período das Águas: número de vacas em lactação, área em pastagem, sanidade, concentrado, grau de sangue e minerais.

Período da Seca: número de vacas, silagem, concentrado, grau de sangue, área em pastagem, capim e cana, minerais e sanidade.

Buscando «consistência teórica» da produção com as relações empiricamente conhecidas e observando os sinais e «coerência» dos coeficientes de regressão, as correlações simples e a significância estatística da regressão, foram selecionadas as equações do Quadro 1, para os dois períodos:

Pode-se notar que nos modelos escolhidos para representar a estrutura de produção de leite na Zona da Mata não aparece a variável área em pastagem, apesar de ter entrado em 2.º e 4.º lugar nos modelos de produção, para os períodos das águas e da seca, respectivamente.

Sua retirada foi justificada pela sua alta correlação simples com variável número de vacas em lactação e também pelo sinal negativo de seu coeficiente na função, indicando, «provavelmente», a presença de multicolinearidade. Sabe-se que um alto coeficiente de correlação entre duas variáveis, num modelo de regressão múltipla, é condição suficiente, mas não necessária, para a ocorrência de um alto grau de multicolinearidade. Portanto, optou-se, neste trabalho, por verificar de maneira mais correta o problema da multicolinearidade antes da aplicação do método de regressão de cumeeira.

Por outro lado, na formulação do modelo, buscou-se um relacionamento mais técnico, em que a produção de leite fosse explicada pelas variáveis: número de vacas em lactação, área em pastagem, concentrado, capim e cana, minerais (medidas em quantidades físicas) e grau de sangue, que é um índice médio de cada rebanho, variando de zero, para os rebanhos sem nenhum arraçoamento e um (1), para os rebanhos com total pureza de sangue. A diferença entre essa função e a selecionada por CASALI (4) em nada interfere nos objetivos da análise, na medida em que, detectado o problema da multicolinearidade, procurou-se manter, no modelo estimado pelo método da regressão de cumeeira, uma variável freqüentemente retirada, com base num sinal incoerente com o esperado.

O modelo especificado foi, primeiramente, estimado pelo método dos mínimos quadrados, e o resultado encontra-se nos Quadros 2 e 3.

Na detecção do problema de multicolinearidade em modelos de regressão múltipla, vários métodos são sugeridos, e a maioria deles foi analisada neste trabalho.

Partiu-se da verificação, feita por CASALI (4), de que o coeficiente de correlação simples entre as variáveis área em pastagem e número de vacas em lactação era relativamente alto nos dois períodos (águas, 0,8162, e seca, 0,8066) e que a variável área em pastagem apresentava sinal contrário ao esperado para o coeficiente. Segundo FARRAR e GLAUBER (7), uma medida prática da existência de multicolinearidade prejudicial ao modelo pode ser dada quando o coeficiente de correlação simples, $r_{x_i x_j}$, for maior que o coeficiente de correlação múltipla, R_y, o que não foi verificado ($r_{x_i x_j} = 0{,}8162$ e $R_y = 0{,}9351$ para o período das águas e $r_{x_i x_j} = 0{,}8066$ e $R_y = 0{,}9415$ para o período da seca). Além disso, quando ocorre um problema mais sério de multicolinearidade, as variâncias dos coeficientes são elevadas, fazendo com que no teste de hipóteses esses coeficientes sejam não significantes, o que também não foi verificado (os coeficientes das variáveis área em pastagem e número de vacas em lactação foram significativos a 1%, nos dois períodos). Realmente, o que chamou mais a atenção para o alto grau de multicolinearidade

QUADRO 1 - Modelos selecionados por CASALI (4) das funções de produção total de leite nas estações seca e das águas, Zona da Mata de Minas Gerais, ano agrícola 1977/78

Variáveis independentes[a]	Coeficientes de regressão	
	Estação seca	Estação das águas
Número médio de vacas em lactação	0,8650* (0,0521)	0,8445* (0,0780)
Silagem	0,0205 (0,0059)	...
Concentrado	0,0395* (0,0112)	0,0199** (0,0085)
Grau de sangue	0,2236* (0,0750)	0,2125* (0,0821)
Sanidade	...	0,2696** (0,1178)
Minerais	...	0,0777** (0,0402)
Coeficiente de determinação (R^2)	0,8865	0,8745
$\bar{R}^2$	0,8789	0,8638
Valor de F	117,1866	82,2208

a: Todas as variáveis são expressas em logaritmos decimais.

*: Significância a 1%; **: significância a 5%.

QUADRO 2 - Funções de produção de leite da estação seca. Estimativas por mínimos quadrados e por cumeeira nos níveis de K = 0,0128, K = 0,05 e K = 0,30

Variáveis independentes [a]	MQO	K = 0,0128	K = 0,05	K = 0,30
Número médio de vacas em lactação	1,0212	0,9767	0,8731	0,5849
Área em pastagem	- 0,1478	- 0,1179	- 0,0495	0,1043
Silagem	0,0210	0,0215	0,0226	0,0234
Concentrado	0,0294	0,0302	0,0318	0,0328
Capim e cana	0,0101	0,0098	0,0090	0,0064
Minerais	0,0221	0,0205	0,0169	0,0119
Grau de sangue	0,2139	0,2157	0,2189	0,2099
Coeficiente de determinação (R^2)	0,9011	0,8895	0,8604	0,7391
Valor de F	74,2487	72,9244	67,0651	39,9512

a: Variáveis não padronizadas expressas em logaritmos decimais.

QUADRO 3 - Funções de produção de leite da estação das águas. Estimativas por mínimos quadrados e por cumeeira nos níveis de K = 0,0118, K = 0,05 e K = 0,30

Variáveis independentes [a]	MQO	K = 0,0118	K = 0,05	K = 0,30
Número médio de vacas em lactação	1,1176	1,1205	0,6863	0,6390
Área em pastagem	- 0,1872	- 0,1524	- 0,0669	0,1082
Concentrado	0,0256	0,0256	0,0231	0,0202
Capim e cana	- 0,0031	- 0,0028	0,0005	0,0024
Minerais	0,0471	0,0557	0,0730	0,1074
Grau de sangue	0,2114	0,2176	0,2231	0,2306
Coeficiente de determinação (R^2)	0,8730	0,8606	0,8271	0,7018
Valor de F	66,4910	65,2262	59,4371	36,0304

a: Variáveis não padronizadas expressas em logaritmos decimais.

foi a ocorrência do sinal oposto ao esperado para a variável área em pastagem, visto que, para a amostra estudada, a área média em pastagem está em torno de 73 ha. É difícil acreditar que um aumento dessa área não trouxesse um aumento

na produção de leite, ainda mais se se conhece a baixa capacidade de suporte desses pastos (0,8 UA/ha).

Uma análise da matriz de correlação indica uma associação linear positiva entre área em pastagem e produção de leite nos dois períodos, a saber: estação da seca (0,6478) e a das águas (0,6517).

Estimaram-se então, para os dois períodos, regressões em que as variáveis área de pastagem e número de vacas em lactação foram consideradas dependentes, em relação às demais variáveis explicativas. Estas apresentaram coeficientes de determinação de 0,70 e 0,67 para o modelo VACAS e ÁREA, respectivamente, no período das águas, o que indica alto grau de dependência entre cada uma delas e as demais.

Outro indicador de multicolinearidade é o determinante da matriz inversa das observações das variáveis explicativas. Para esse modelo, como as variáveis estão padronizadas, a amplitude de variação do valor do determinante está entre 0 e 1, e os valores encontrados para os dois períodos foram 0,165, na seca, e 0,135, nas águas. Estando próximos de zero, constituem indicadores do problema.

Utilizou-se também o teste proposto por FARRAR e GLAUBER (*7*) para mostrar que, se a população dos X'_S, tem distribuição normal multivariada e é ortogonal para a amostra da regressão, a expressão

$$\chi^2 = \left\{ n - 1 - 1/6\,(2\,K + 5) \right\} \ln |R^*|.$$

tem distribuição de χ^2, com $\varnothing = 1/2\,K\,(K-1)$ graus de liberdade, sendo K o número de variáveis independentes e $|R^*|$ o determinante da matriz de correlação dos X'_S.

Como $|R^*|$ assume valores entre 0 e 1, conforme os dados amostrais apresentem características de singularidade ou ortogonalidade, compara-se o valor de χ^2 calculado por essa fórmula com o valor obtido da distribuição teórica de χ^2 a determinado nível de significância e com os graus de liberdade citados. Esse teste, a 5% de significância e 28 graus de liberdade, forneceu um valor de χ^2 igual a 32,6706. Os valores de χ^2 calculados pela fórmula foram iguais a 109,33 e 137,70, para os períodos da seca e das águas, sendo, portanto, bem maiores que o valor tabelado, fazendo com que seja aceita a hipótese da existência de multicolinearidade.

O quadrado da distância entre $\hat{B}$ e B foi tido por HOERL e KENNARD (*9*) como dado pela fórmula $E[L^2_1] = \sigma^2 \sum_{i=1}^{K} (1/\lambda_i)$ sendo λ_i os valores característicos ou autovalores da matriz $(X'X)^{-1}$. O somatório do inverso dos valores característicos foi igual a 13,05 e 13,04, para os períodos da seca e das águas, respectivamente, mostrando que os coeficientes estimados pelo método dos mínimos quadrados têm variâncias aproximadamente duas vezes maiores do que deveriam ter, caso o sistema fosse ortogonal.

Os resultados obtidos nos diversos testes sugerem a existência de um grau de multicolinearidade prejudicial ao modelo estimado e, portanto, reforçam a utilização do método da regressão da cumeeira como possível solução para o problema. Na estimação das funções para os dois períodos, adotaram-se valores de K com variação de 0,05, no intervalo de 0 a 1.

A questão, agora, seria a determinação de um valor de K, nesse intervalo, que fornecesse estimativas dos parâmetros da regressão livres do problema da multicolinearidade e, portanto, estáveis no gráfico de cumeeira.

Utilizando o teste proposto por HOERL e KENNARD (*11*), os valores de K encontrados para os dois períodos foram de 0,0128, para a seca, e de 0,0118, para as

águas. Os coeficientes da função estimados para esses valores de K são apresentados nos Quadros 2 e 3. Pode-se verificar que o coeficiente da variável área em pastagem, nos dois períodos, apresentou-se ainda negativo, o que indica que, nesse ponto, persiste o efeito da alta correlação dessa variável com a variável número de vacas em lactação. A comparação dos coeficientes, nesses níveis de K, com os de mínimos quadrados não mostra diferenças significantes de valores que cheguem a alterar as elasticidades parciais dos fatores ou os retornos à escala naquelas funções. Observando os índices de estabilidade das magnitudes relativas (IEMR) propostos por VINOD (*20*), tem-se que os menores índices foram obtidos para valores de K iguais a 0,05, nos dois períodos. Os coeficientes das variáveis, nesse nível de K, são apresentados nos Quadros 2 e 3.

As alterações sofridas pelos coeficientes, nesse nível de K, são bem maiores agora, com mudança de sinal da variável capim e cana no modelo da estação das águas. No entanto, parece que esse valor de K ainda não é ideal, visto que permanece o sinal trocado da variável área em pastagem. Vale ressaltar que os IEMR atingiram os valores 2,97 e 1,39, para a estações da seca e das águas, respectivamente. Valores menores poderiam ser encontrados caso os intervalos de K adotados fossem menores que 0,05.

Para cada valor de K foram obtidas diferentes estimativas dos coeficientes das variáveis na função, o que possibilitou traçar os gráficos de cumeeira das Figuras 1 e 2.

Numa simples visualização desses gráficos, percebe-se a grande modificação dos coeficientes de algumas variáveis, principalmente para pequenos acréscimos em K e para as variáveis que, presume-se, causam maiores problemas ao modelo (área em pastagem e número de vacas em lactação).

No entanto, observando, nos gráficos, os pontos referentes aos valores de K selecionados pelos testes de HOERL e KENNARD e VINOD, nota-se que eles fornecem estimativas dos coeficientes que não se estabilizaram, permanecendo a inversão de sinal da variável área em pastagem.

Considerando que HOERL e KENNARD (*11*), em seu primeiro trabalho, sugeriram a escolha de um valor de K que fornecesse estimativas relativamente estáveis, cujos sinais deveriam permanecer sem modificação, e que, nesse nível, a soma dos quadrados dos resíduos não sofreu acréscimo excessivo, em relação ao nível K = 0, o método gráfico, apesar de subjetivo, facilita a visualização de um nível de K em que as estimativas dos coeficientes obedeçam a essas considerações.

Pelas Figuras 1 e 2, esse nível de K ficaria em torno de 0,3, para os dois períodos. As estimativas dos coeficientes, nesse nível, encontram-se nos Quadros 2 e 3.

Verifica-se, agora, mudança bastante acentuada nos valores dos coeficientes das variáveis número de vacas em lactação e área em pastagem, tendo a primeira se reduzido praticamente à metade do valor inicial e a segunda se tornado positiva, sinal que permanece para os níveis de K subseqüentes.

As demais variáveis, no valor de K escolhido, não sofreram alterações muito grandes, o que mostra que esse método realmente elimina os problemas de multicolinearidade do modelo, atuando nas variáveis que apresentam alta correlação entre si.

Nos Quadros 2 e 3 são apresentados também os valores dos coeficientes de determinação (R^2) e a estatística «F», para as estimativas de mínimos quadrados e de cumeeeira. O coeficiente de determinação, como medida descritiva das estimativas de cumeeira, continua válido e, apesar de sua diminuição para valores crescentes de K, nos níveis escolhidos, permanece bem alto. Quanto aos valores da estatística «F», apesar de apresentados, não podem ser utilizados como testes de significância.

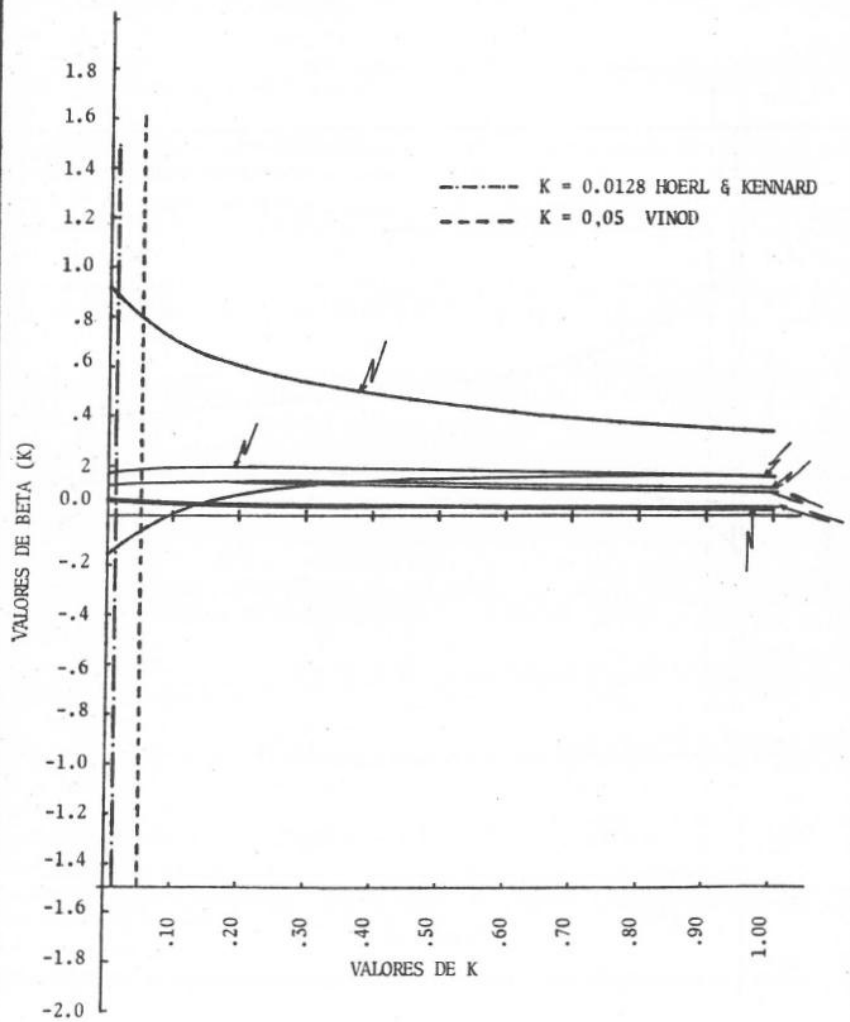

FIGURA 2 - Gráfico de cumeeira para a função de produção de leite no período das águas: 1 - número de vacas em lactação; 2 - área em pastagem; 3 - concentrado; 4 - capim e cana; 5 - minerais; 6 - grau de sangue.

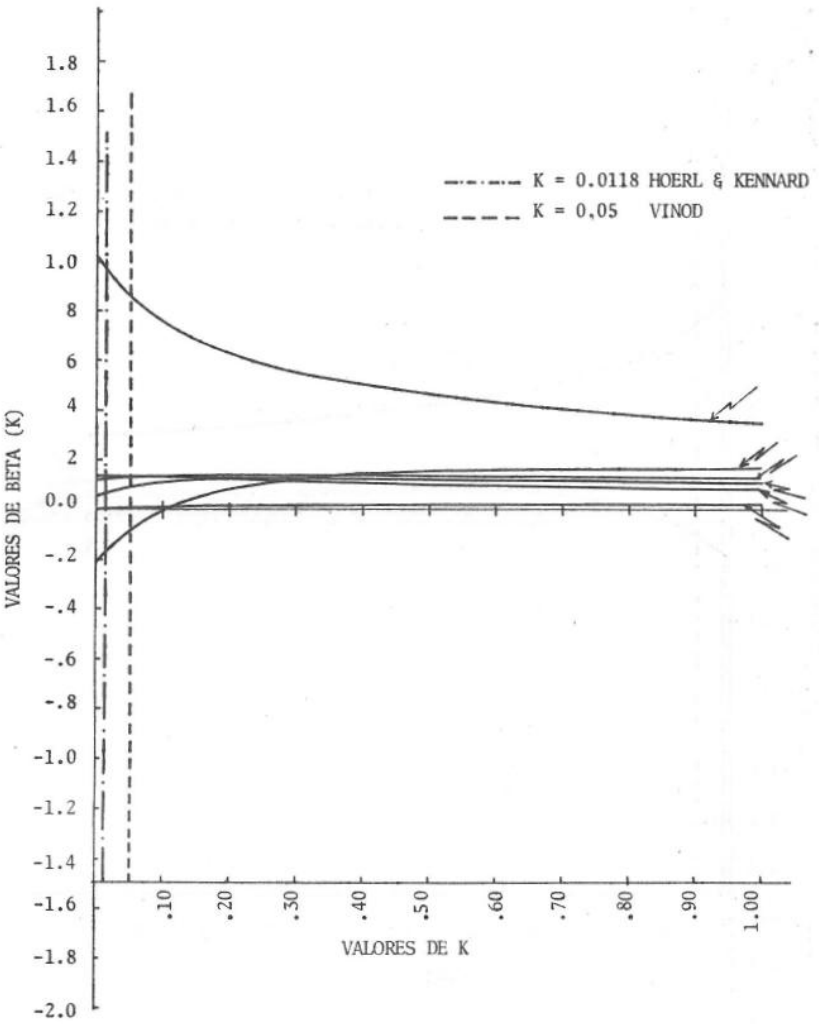

FIGURA 1 - Gráfico de cumeeira para a função de produção de leite no período de seca: 1 - número de vacas em lactação; 2 - área em pastagem; 3 - silagem; 4 - concentrado; 5 - capim e cana; 6 - minerais; 7 - grau de sangue.

A análise global dos dois modelos estimados indica que todos os fatores de produção são utilizados no estadio racional e que a análise individual desses fatores dá os mesmos resultados obtidos por CASALI (*4*), para as variáveis incluídas em seu modelo. A diferença encontra-se justamente na inclusão, no modelo estimado pelo método da regressão de cumeeira, de variáveis importantes e que não foram incluídas no aludido estudo.

Com relação aos testes propostos por HOERL e KENNARD (*11*) e por VINOD (*20*) para determinar o nível ótimo de K, as estimativas dos coeficientes da função continuam com sinais diferentes do esperado; portanto, o método gráfico, nesse caso, mostrou-se o mais indicado.

Contudo, antes de adotar o método da regressão de cumeeira em modelos que apresentem problemas de multicolinearidade, deve-se ter senso crítico para conciliar seus resultados com a relação funcional teórica analisada.

4. RESUMO E CONCLUSÕES

Um dos grandes problemas da estimação de funções de regressão é a alta correlação entre variáveis, o que traz graves perturbações para os estimadores, com dificuldades na interpretação dos parâmetros.

Neste trabalho, utilizou-se um dos métodos propostos para a correção dos problemas da multicolinearidade (regressão de cumeeira), com o objetivo de comparar os resultados com os encontrados por CASALI (*4*) na estimação de duas funções de produção de leite para a Zona da Mata de Minas Gerais, utilizando o método dos mínimos quadrados numa função Cobb-Douglas.

Inicialmente, reestimando as funções pelo método dos mínimos quadrados, testou-se, de várias formas, a existência de um grau de multicolinearidade que fosse prejudicial ao modelo. Comprovado o problema, partiu-se para a aplicação do método da regressão de cumeeira, utilizando dois procedimentos propostos por HOERL e KENNARD (*11*) e VINOD (*20*) para determinar o nível de K que forneceria estimativas tidas como ótimas. Esses níveis de K foram 0,0128 e 0,0118, pelo teste de HOERL e KENNARD (*11*), nos períodos da seca e das águas, respectivamente, e iguais a 0,05, pelo teste de VINOD (*20*), nas funções estabelecidas para os dois períodos.

Esses valores de K forneceram estimativas dos parâmetros das funções que não sofreram modificações relevantes. Para a variável área em pastagem, nos dois períodos, o sinal permaneceu trocado, fato evidenciado pelos gráficos de cumeeira traçados para as funções.

A análise dos referidos gráficos sugeriu que o valor de K estabilizador das estimativas dos parâmetros estaria em torno de 0,30. Os coeficientes das variáveis explicativas, nesse nível, apresentaram-se positivos, conforme esperado.

Além disso, somente os coeficientes das variáveis área em pastagem e número de vacas em lactação sofreram alterações significativas para esse valor de K, justamente as variáveis que apresentavam alta correlação entre si.

Em razão da subjetividade da escolha do nível de K pelo método gráfico e de não serem muito convincentes os resultados obtidos da aplicação dos dois testes propostos, conclui-se que a utilização do método da regressão de cumeeira deve ser precedido de um conhecimento profundo do fenômeno que se quer estudar, de forma que o pesquisador possa discernir entre um resultado estatístico e a coerência teórica do relacionamento proposto. Ficou evidenciado que, realmente, o método tende a solucionar o problema da alta correlação entre variáveis explicativas, mas que há necessidade de avaliar alguns testes propostos como solução ou,

talvez, desenvolver outros que reúnam as características matemáticas e teóricas do modelo que vai ser analisado.

5. SUMMARY

(APPLICATION OF THE RIDGE REGRESSION METHOD TO A FUNCTION OF MILK PRODUCTION IN THE «ZONA DA MATA» REGION OF THE STATE OF MINAS GERAIS)

The estimated function of milk production for «Zona da Mata» region State of Minas Gerais, was determined by two different methods: (a) Ridge regression; and (b) Ordinary Least Squares.

The tests proposed by HOERL & KENNARD and VINOD were used to determine the level of K that generates optimal estimates of the parameters; however, high correlation among variables still persisted.

An alternative level of K was suggested using ridge graphs.

Due to the subjectivity of this procedure to determine the level of K, it was concluded that prior to the use of ridge regression it is desirable to have a deep knowledge about the problem being studied so that. The researcher can discriminate betwen the statistical results and te related theory.

It was clear that ridge regression method tends to solve the high correlation problems among exogenous variables, but there is a necessity to develop other tests that analyze the mathematical and theoretical characteristics of the model being studied.

6. LITERATURA CITADA

1. BARE, B. & HANN, D.W. Applications of ridge regression in forestry. *Forest Science, 27*(2):339-348. 1981.

2. BELONGIA, M. An application of ridge regression with verification of new procedures. *Agricultural Economics Review, 31*(2):36-39. 1979.

3. BROWN, W.G. *Effect of omitting relevant variables versus use of ridge regression in economic research*. Corvallis, Oregon State University, 1973. 40 p. (Special Report 394).

4. CASALI, A.S.D. *Análise da estacionalidade da produção de leite na Zona da Mata — MG; ano agrícola 1977/78*. Viçosa, U.F.V., Imprensa Universitária, 1981. 73 p. (Tese M.S.).

5. CHATTERJEE, S. & PRICE, B. *Regression analysis by example*. New York. John Wiley & Sons, 1977. 229 p.

6. DRAPER, N.R. & SMITH, H. *Applied regression analysis*. New York, John Wiley & Sons, 1977. 709 p.

7. FARRAR, D.E. & GLAUBER, R.R. Multicolinearity in regression analysis: the problem revisited. *Review of Economics and Statistics, 49*(2):92-107. 1967.

8. GUJARATI, D. *Basic econometrics*. New York, McGraw-Hill, 1978. 426 p.

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

10. HOERL, A.E. & KENNARD, R.W. Ridge regression: Applications to nonorthogonal problems. *Tecnometrics, 12*(1):69-82. 1970.

11. HOERL, A.E. & KENNARD, R.W. Ridge regression: some simulations. *Communication in Statistics, 4*:105-123. 1975.

12. JOHNSTON, J. *Métodos econométricos*. São Paulo, Atlas, 1977. 318 p.

13. KALIL, M.N. *Aplicação do método de regressão de cumeeira (ridge regression) na estimação de funções da demanda e de produção*. Piracicaba, ESALQ, 1977. 153 p. (Tese M.S.).

14. KMENTA, J. *Elementos de Econometria*. São Paulo, Atlas, 1978. 670 p.

15. MADDALA, G.S. *Econometrics*. New York, McGraw-Hill, 1977, 516 p.

16. MAHAJAN, N., JAIN, A.K. & BERGIER, M. Parameter estimation in marketing models in the presence of multicolinearity: an application of ridge regression. *Journal of Marketing Research, 14*(4):586-591., 1977.

17. MARQUARDT, D.W. & SNEE, R.D. Ridge regression in practice. *American Statistician, 29*(1):3-19. 1975.

18. MITTELHAMMER, R.C. Young, D.L., TASANASANTA, D. and DONNELLY, J.T. Mitigating the effects of multicolinearity using exact and stochastic restrictions: the case of an aggregate agricultural production function in Thailand. *American Journal of Agricultural Economics, 62*(2):199-209. 1980.

19. ROSSI, J.W. A matriz de correlação revisitada. *Revista Brasileira de Estatística, 38*(152):379-384. 1977.

20. VINOD, H.D. Application of new ridge regression methods to a study of Bell system scale economies. *Journal of American Statistics Association 71*(356): 835-841. 1976.