Uso de alguns estimadores *ridge* na análise estatística de experimentos em entomologia

Gislene Araujo Pereira[1], Letícia Lima Milani[2], Marcelo Ângelo Cirillo[3]

RESUMO

Inúmeros experimentos em ciências agrárias apresentam variáveis que podem dar origem a problemas de multicolinearidade. Em se tratando da aplicabilidade de modelos de regressão, o problema da multicolinearidade tem como principal consequência o inflacionamento dos erros padrão e, com isso, o valor da estatística t-student é reduzido de tal forma que interfere nos resultados inferenciais. Várias medidas são propostas, na literatura, para resolver o problema de multicolinearidade. Entretanto, o desempenho dessas medidas está sujeito ao grau de multicolinearidade que as variáveis poderão apresentar, bem como ao tamanho amostral. Frente a este problema, este trabalho tem por objetivo avaliar alguns estimadores *ridge*, utilizando simulação Monte Carlo, bem como, apresentar a aplicação desses estimadores em um experimento, com dados reais, na área de entomologia. Mediante esta aplicação, os resultados expressivos alcançados foram obtidos em função da eficiência dos estimadores *ridge* avaliados, em relação ao estimador de mínimos quadrados. Em se tratando dos resultados computacionais, concluiu-se que estimadores *ridge* avaliados são recomendáveis, em experimentos que considerem as variáveis com diferentes graus de multicolinearidade, para amostras maiores do que n=50.

Palavras-chave: multicolinearidade, tamanho amostral, modelos de regressão.

ABSTRACT

Use of some *ridge* estimators in the statistical analysis of experiments in entomology

A large number of experiments in agronomic sciences use variables that may give rise to problems of multicollinearity. About the applicability of regression models, the problem of multicollinearity results mainly in increased standard error, thus, the Student's t-value is reduced, affecting the inferential results. Many actions are proposed in the literature to solve the problems of multicollinearity, however, the performance of these measurements are subject to the degree that multicollinearity of the variables may present, as well as the sample size. To address this problem, this paper aims to evaluate some *ridge* estimators using the Monte Carlo's simulation and demonstrate their application using real data from an entomological experiment. The ridge estimators evaluated were effective, in comparison with the least squares estimator. The results showed that the *ridge* estimators evaluated can be applied to experimenst that consider the variables with different degrees of multicollinearity, for samples greater than n=50.

Key words: multicollinearity, sample size, regression models.

Recebido para publicação em 04/03/2013 e aprovado em 10/10/2013.

[1] Estatística, Doutora. Departamento de Ciências Exatas, Universidade Federal de Lavras, 37200-000, Caixa Postal 3037, Lavras, Minas Gerais, Brasil. gislene.araujo.p@gmail.com

[2] Engenheira- Agrônoma, Doutora. Departamento de Ciências Exatas, Universidade Federal de Lavras, 37200-000, Caixa Postal 3037, Lavras, Minas Gerais, Brasil. rodrigues.milani.l@gmail.com

[3] Estatístico, Pós-Doutor. Departamento de Ciências Exatas, Universidade Federal de Lavras, 37200-000, Lavras, Caixa Postal 3037, Minas Gerais, Brasil. macufla@dex.ufla.br (autor para correspondência).

INTRODUÇÃO

A multicolinearidade é observada, em um modelo de regressão, quando há evidências de um alto grau de correlação entre as variáveis regressoras. A principal consequência é verificada na inferência relacionada com as estimativas dos parâmetros, uma vez que os erros padrões das estimativas são inflacionados, resultando em intervalos de confiança com grandes amplitudes e, naturalmente, menos precisos.

Vários métodos têm sido propostos para resolver o efeito da multicolinearidade em modelos de regressão e maiores detalhes a respeito poderão ser vistos em Guilkey & Murphy (1975), Conniffe & Stone (1974). Contudo, a metodologia proposta por Hoerl & Kennard (1970), na qual se considera a redução na variância das estimativas com a inclusão de um parâmetro *shrinkage*, representado por k, tem sido a mais usual.

Convém ressaltar que o desempenho desse método depende do tamanho amostral e com o grau de multicolinearidade entre as covariáveis envolvidas em um experimento. Assim sendo, para que o método *ridge* (Kibria, 2003) possa ser utilizado adequadamente, é conveniente que o pesquisador tenha conhecimento da relação do tamanho amostral e do efeito da multicolinearidade, supostamente presente nas covariáveis a serem utilizadas no modelo de regressão.

Uma forma de avaliar o grau de multicolinearidade é por meio do fator de inflação da variância, definido por $VIF_j = \frac{1}{1-R_j^2}$, sendo R^2_j o coeficiente de correlação múltipla, resultante da regressão de X_j nos outros p-1 regressores. Quanto maior o grau de dependência de X_j nos regressores restantes e, assim, mais forte a colinearidade, maior será o valor de R^2_j. Percebe-se que esta medida indica que cada variável independente é explicada pelas demais variáveis independentes, de tal forma que a correlação entre as covariáveis é considerada na estimativa do VIF. A questão surge em classificar-se o grau de multicolinearidade em severo ou não. Alguns autores, como, por exemplo, Chatterjee & Hadi (2006), Petrini *et al.* (2012), sugerem que, se qualquer VIF exceder 10, então a multicolinearidade causará efeitos nos coeficientes de regressão. Outros autores, como Myers & Montgomery (2002), sugerem que VIF não deve exceder o valor de 4 ou 5 unidades.

Cabe argumentar que o conhecimento do pesquisador, em consonância com outras metodologias, como, por exemplo, o uso de simulações Monte Carlo, é de grande importância para a classificação do grau de multicolinearidade como severo, ou para a obtenção de um modelo de regressão. Neste sentido, os métodos de simulação Monte Carlo, representam uma contribuição relevante, no que tange à simulação de experimentos em diferentes cenários, que permitam avaliar as propriedades estatísticas de um modelo.

A título de ilustração, cita-se estudo realizado por Oliveira *et al.* (2011), em relação ao desempenho das medidas de curvaturas dos modelos de regressão, de Oswin (1946) e Halsey (1948), em função de diferentes níveis de atividade de água. Neste contexto, os autores concluíram que, em ambos os modelos, os resultados da medida de curvatura extrínseca evidenciaram que, para todas as faixas de atividade de água avaliadas, os modelos carecem de uma parametrização que possa garantir um comportamento mais próximo ao linear.

Em virtude do que foi mencionado, este trabalho tem por objetivo apresentar um estudo de simulação Monte Carlo e uma aplicação em dados entomológicos, em relação à viabilidade do uso de estimadores de regressão *ridge* em experimentos que apresentam efeito de multicolinearidade entre as variáveis regressoras.

MATERIAL E MÉTODOS

A fundamentação metodológica deste trabalho considerou o modelo de regressão linear múltiplo, dado por (1) $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\zeta}$, sendo $\mathbf{Y}_{n\times1}$ o vetor de observações dependentes, $\boldsymbol{\beta}_{p\times1}$ o vetor paramétrico dos coeficientes de regressão a serem estimados, $\mathbf{X}_{n\times p}$ uma matriz conhecida de variáveis explicativas e $\boldsymbol{\zeta}_{n\times1}$ o vetor dos resíduos, em que cada componente $\zeta_i \sim N(0,\sigma^2)$ (i = 1,..., n).

O estimador de mínimos quadrados dos coeficientes de regressão, mencionado por Farrar & Glauber (1967), é dado por, $\mathbf{C=(X^tX)}$. Contudo, os autores ressaltam que, se as variáveis explicativas forem multicolineares, os coeficientes da regressão não poderão ser estimados, uma vez que C^{-1} será singular. Neste caso, a multicolinearidade é considerada severa e o modelo deverá ser revisto.

Diagnosticado o grau de multicolinearidade e frente ao problema de calcular a inversa C^{-1}, procedeu-se à aplicação de estimadores *ridge*, propostos por Kibria (2003), considerando-se o modelo (1) e uma matriz ortogonal **D,** tal que $\mathbf{D^tCD = \Lambda}$, em que $\boldsymbol{\Lambda}$ contém os autovalores da matriz $\mathbf{C = (X^tX)}$. Desta forma, o modelo linear geral (1), na forma canônica, foi reescrito por $\mathbf{Y = X^*\alpha + å}$, sendo $\mathbf{X^* = XD}$ e $\boldsymbol{\alpha} = \mathbf{D^t\beta}$.

Com estas especificações, o estimador de mínimos quadrados, para o modelo na forma canônica, apresentado por Kibria (2003), é dado por $\hat{\boldsymbol{\alpha}} = \Lambda^{-1}\left(\mathbf{X}^*\right)^t\mathbf{Y}$ e os estimadores de regressão generalizada *ridge* são apresentados por (2)

$$\hat{\boldsymbol{\alpha}}(k) = \left[\left(\mathbf{X}^*\right)^t\mathbf{X}^* + \mathbf{K}\right]^{-1}\left(\mathbf{X}^*\right)^t\mathbf{Y} \qquad (2)$$

sendo **K** uma matriz diagonal definida por K = diag (k_1, k_2,..., k_p), $k_j > 0$ (j = 1, ..., p)

Tendo por base esta regressão, Kibria (2003) propôs os estimadores para k, utilizando as média aritmética (3) e geométrica (4) e a mediana (5).

$$K_m \frac{1}{p}\sum_{i=1}^{p}\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_j^2}\right) \quad (3)$$

$$K_g = \left(\frac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p}\hat{\alpha}_j^2\right)^{\frac{1}{p}}}\right) \quad (4)$$

$$K_{med} = mediana\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_j^2}\right) \quad (5)$$

Em todas as situações, α_i indicou o j-ésimo elemento de $\boldsymbol{\alpha}$ e o quadrado médio residual obtido no modelo (1). Desta forma, com a matriz diagonal K=diag(K_m); K=diag(K_g) e K=diag(K_{med}) redefinida, respectivamente, para cada estimador, e substituindo no modelo (2), obtiveram-se os estimadores da regressão generalizada para os parâmetros do modelo de regressão múltipla definido em (1).

Para avaliar o desempenho desses estimadores, procedeu-se a um estudo de simulação, seguindo-se o procedimento especificado por Gibbons (1981), o que permitiu especificar os diferentes graus de correlação entre as variáveis explicativas, assumindo-se a relação (6).

$$X_{ji} = (1 - \gamma^2)^{1/2} Z_{ij} + \gamma Z_{ip} \text{ em que} \quad (6)$$

z_{ij} correspondeu aos valores gerados por uma distribuição normal padrão e γ^2 representou o grau de correlação ($0 \leq \gamma^2 \geq 1$) entre duas variáveis explicativas. Especificando-se o resíduo ζ_i (i = 1,...,n) ~ N(0,1) e os coeficientes $\beta_1,\beta_2,...,\beta_{p,\ e}$ assumindo-se cada valor do autovetor normalizado, correspondente ao maior autovalor da matriz $\mathbf{X'X}$, tornou-se possível gerar a variável resposta.

Seguindo-se este procedimento, os valores paramétricos assumidos no processo de simulação foram arbitrariamente definidos como $\sigma^2 = 4$ e 20; p = 4; γ = 0,1; 0,5; 0,7 e 0,9 e, por fim, consideraram-se os tamanhos amostrais, definidos em n = 15, 50 e 100. Desta forma, dado os cenários envolvendo grau de multicolinearidade (γ) e os tamanhos amostrais (n), computou-se a distribuição empírica do erro quadrático médio (EQM), considerando-se os estimadores (3)-(5).

O valor esperado do (EQM), em 2.000 simulações Monte Carlo para cada estimador, foi obtido para os estimadores de regressão de mínimos quadrados (MQ) e *ridge*, mencionados anteriormente.

A fim de ilustrar a aplicação dos estimadores *ridge* em contexto agrário, considerou-se um experimento, cujo objetivo foi estudar a melhor combinação dos componentes (Tabela 1) a serem utilizados para a formulação de uma dieta energética, que proporcionasse maior tempo de vida das operárias de abelhas *Apis mellifera*. (Brighenti *et al*. 2010).

O delineamento experimental adotado foi extreme-vértice (Piepel & Cornell, 1987), pelo fato de que as proporções referentes aos componentes utilizados para a formação da mistura, no experimento, foram submetidas a restrições (Tabela 1). O modelo de regressão ajustado foi dado pela equação $Y_i = \hat{a}_0 + \hat{a}_1 X_{i1} + \hat{a}_2 X_{i2} + \hat{a}_3 X_{i3} + \hat{a}_4 X_{i4} + \mathring{a}_{i,}$ para a i-ésima unidade amostral (i=1,...,n=54), Y_i correspondeu ao número de abelhas vivas, considerando-se a transformação raiz quadrada, X_{1i} é o tempo (horas) de submissão à dieta, X_{2i} a proporção de água na dieta, X_{3i} é a proporção de açúcar granulado na dieta e X_{4i} é a proporção de suco de limão Tahiti na dieta.

Por fim, para a realização deste trabalho, procedeu-se à elaboração de uma rotina computacional no programa R (Development Core Team, 2012).

RESULTADOS E DISCUSSÃO

Estudos de simulação Monte Carlo

Os resultados descritos na Tabela 2 evidenciaram que, nas situações simuladas, envolvendo fraca multicolinearidade, isto é, (γ = 0,1), para todos os tamanhos amostrais, os estimadores *ridge* apresentaram um erro quadrático bem inferior ao erro proporcionado pelo método de mínimos quadrados.

Reportando-se esses resultados a uma situação real, dado um modelo com p-variáveis regressoras, há evidências estatísticas de que, independentemente do tamanho amostral, na situação de fraca multicolinearidade o pesquisador não deverá utilizar o método de mínimos quadrados ordinários, uma vez que os estimadores *ridge* apresentaram resultados indicativos de melhor precisão e acurácia. Esta afirmativa é também verificada ao se analisarem os resultados obtidos em uma situação de grau de multicolinearidade considerado moderado, conforme resultados apresentados para (γ = 0,5). Porém, ao se considerar um grau elevado (γ = 0,7) e severo (γ = 0,9) de multicolinearidade, os valores do erro quadrático médio proporcionado pelos estimadores *ridge* foram aumentados, mas inferiores aos valores obtidos pelo estimador de mínimos quadrados, com exceção do tamanho amostral n

Tabela 1. Restrições dos componentes utilizados na formulação da dieta energética

Componentes	Valor mínimo	Valor máximo
Açúcar	0,40	0,50
Água	0,45	0,50
Suco de Limão	0,00	0,10

Fonte: Brigenti *et al.*(2010)

= 15 dado (γ = 0,9), no qual, foram observados valores exorbitantes para o erro quadrático médio, ao se comparar com o valor obtido no método de mínimos quadrados. (Tabela 2).

Em consequência dos resultados obtidos, a comparação dos estimadores *ridge*, em considerando diferentes níveis de multicolinearidade torna-se relevante, pelo fato de que a multicolinearidade provoca uma redução da precisão das estimativas, condicionada aos dados. Como principal consequência, as variâncias são inflacionadas. Wetherill (1986) explica que dados mal condicionados são provenientes de covariáveis que estão escritas como combinações lineares das demais, portanto, apresentam pouca contribuição em qualquer estatística que envolva a variância das estimativas. Exemplos a serem considerados seriam os testes de significância para os coeficientes.

Aplicação em um experimento com dados reais

Com base nos resultados obtidos no experimento descrito na metodologia, foi estimado para cada variável o fator de inflação da variância e os valores estão apresentados na Tabela 3.

Aplicando-se os estimadores *ridge* descritos em (3), (4) e (5) e o estimador de mínimos quadrados para esse experimento, observou-se que todos os estimadores foram eficientes, em relação ao método dos mínimos quadrados, com destaque para o estimador K_{med}, que apresentou maior precisão que os demais (Tabela 4). Desta forma, nota-se que o estimador *ridge* dado por K_{med} corrigiu com maior precisão o efeito da multicolinearidade, proporcionando maior estabilidade, no sentido de que, em caso de alguma perturbação nos dados, o impacto nas estimativas dos parâmetros será pequeno.

Tabela 2. Erro quadrático médio para os estimadores *ridge* e mínimos quadrados, para os cenários de simulação avaliados via Monte Carlo

γ	n	σ	MQ	K_m	K_g	K_{med}
0,1	15	4	0,8635	0,2720	0,1836	0,1865
		20	0,9119	0,2707	0,1920	0,1939
	50	4	0,9241	0,0809	0,0482	0,0476
		20	0,9769	0,0595	0,0411	0,0401
	100	4	0,9336	0,0517	0,0272	0,0271
		20	0,9869	0,0290	0,0199	0,0195
0,5	15	4	0,8792	0,4842	0,3487	0,3596
		20	0,9103	0,4739	0,3536	0,3645
	50	4	0,9309	0,1231	0,0789	0,0780
		20	0,9750	0,1021	0,0747	0,0753
	100	4	0,9431	0,0691	0,0398	0,0395
		20	0,9874	0,0481	0,0348	0,0350
0,7	15	4	0,8776	0,8199	0,6174	0,6414
		20	0,9066	0,8453	0,6547	0,6847
	50	4	0,9405	0,1984	0,1374	0,1398
		20	0,9747	0,1850	0,1397	0,1434
	100	4	0,9524	0,1064	0,0676	0,0673
		20	0,9887	0,0837	0,0627	0,0637
0,9	15	4	0,9799	2,9365	2,3067	2,3458
		20	0,9141	2,8680	2,2882	2,2911
	50	4	0,9490	0,6038	0,4572	0,4604
		20	0,9766	0,6170	0,4859	0,4868
	100	4	0,9609	0,2957	0,2144	0,2160
		20	0,9890	0,2810	0,2196	0,2167

Tabela 3. Valores do VIF da variável independente X_j em relação às demais variáveis independentes

Variáveis X_j	X_1	X_2	X_3	X_4
	Tempo de submissão	**Proporção de água na dieta**	**Proporção de açúcar na dieta**	**Proporção de limão na dieta**
VIF_j	30,584	6,265	21,110	30,552

Tabela 4. Estimativas do erro quadrático médio (EQM) estimados, considerando o método de mínimos quadrados e os estimadores *ridge* propostos por Kibria (2003)

Estimadores	MQ	K_m	K_g	K_{med}
EQM	0,3476	0,0046	0,0046	0,0016

Convém ressaltar que a maior precisão observada para o estimador $K_{med,}$ de certa forma foi concordante com os resultados simulados (Tabela 2). Por se tratar de estudo de simulação, no qual a variância dos dados previamente fixada, e por causa das oscilações do erro do método Monte Carlo, naturalmente alguma diferenciação nos valores obtidos é justificável, por este estudo.

CONCLUSÃO

Os estimadores *ridge*, propostos por Kibria (2003), podem ser utilizados em experimentos que considerem as variáveis com diferentes graus de multicolinearidade, para amostras maiores do que n=50, em Entomologia.

AGRADECIMENTOS

Os autores agradecem o auxílio financeiro recebido do CNPq.

REFERÊNCIAS

Brighenti DM, Brighenti CRG, Cirillo MA & Santos CMB (2010) Optimization of the components of an energetic diet for africanized bees through the modelling of mixtures. Journal of Apicultural Research, 49:326-333.

Chatterjee S & Hadi AS (2006) Regression analysis by example. 4ªed. New York, John Wiley. 408p.

Conniffe D & Stone J (1974) A critical review of *ridge* regression. The Statistician, 22:181-187.

Farrar DE & Glauber RR (1967) Multicollinearity in regression analysis: The Problem Revisited. The review of economics and statistics, 49:92-107.

Gibbons DG (1981) A simulation study of some *ridge* estimators. Journal of the American Statistical Association, 76:131-139.

Guilkey DK & Murphy JL (1975) Directed *ridge* regression techniques in cases of multicollinearity. Journal of the American Statistical Association, 70:769-775.

Halsey G (1948) Physical adsorption on non-uniform surfaces. Journal of chemistry physics, 16:931-937.

Hoerl AE & Kennard RW (1970) *Ridge* regression: biased estimation for nonorthogonal problems. Technometrics, 12:55-67.

Kibria BMG (2003) Performance of Some New Regression Estimators. Communications in Statistic. Simulation and Computation, 32:419-435.

Myers RH & Montgomery DC (2002) Response surface methodology: process and product optmization using designed experiments. 2ªed. Nova York, John Wiley. 798p.

Oliveira IA, Cirillo MA & Borges SV (2011) Estudo da não linearidade dos modelos de Oswin e Halsey aplicados na construção de isotermas. Revista Ceres, 58:735-739.

Oswin CR (1946) The kinetics of package life. III. The isotherm. Journal of Chemistry and Industry, 56:419-423.

Petrini J, Dias RAP, Pertile SFN, Eler JP, Ferraz JBS & Mourão GB (2012) Degree of multicollinearity and variables involved in linear dependence. Pesquisa Agropecuária Brasileira, 47:1743-1750.

Piepel GF & Cornell JA (1987) Designs for mixture-amount experiments. Journal of Quality Technology, 19:11-28.

R Development Core Team (2012) R: A Language and environment for statistical computing. Vienna, R Foundation for Statistical Computing. Disponível em: <http://www.r-project.org/ >. Acessado em: 01 de janeiro de 2012.

Wetherill GB (1986) Regression analysis with applications. New York, Chapman and Hall. 408p.