AVALIAÇÃO DE MÉTODOS PARA COMPARAÇÃO DE MODELOS DE REGRESSÃO POR SIMULAÇÃO DE DADOS

Sérgio Ricardo Silva Magalhães[1],Márcio Magini[2]

[1]Universidade Vale do Rio Verde de Três Corações - UNINCOR / Departamento de Ciências Exatas
Rua Capri, 27 – Arquipélago Verde 32000-000– Betim – MG – Brasil.
sergio@ufla.br
[2] Universidade do Vale do Paraíba -UNIVAP / Instituto de Pesquisa e Desenvolvimento - IP&D
Av. Shishima Hifumi, 2911 - Urbanova 12244-000 - São José dos Campos -SP – Brasil.
magini@univap.br

Resumo - O presente estudo teve como objetivo avaliar os métodos da Identidade de Modelos, das Variáveis Dummy (binárias) e da Análise de Variância, usados para a comparação de modelos de regressão por meio de simulação de dados em computador. Foram considerados quatro casos de regressão linear e cinco casos de regressão polinomial quadrática. Utilizando-se os recursos do Interactive Matrix Language (IML), do sistema SAS®, foram desenvolvidas rotinas apropriadas para a metodologia de comparação de modelos de regressão. Realizou-se uma simulação de dados composta de 10.000 experimentos, considerando os diferentes tamanhos de amostras (10, 50 e 100 observações) para cada uma dos nove casos. Os resultados de todas os casos simulados pelos três métodos foram semelhantes, apresentando baixos percentuais de Erro Tipo I e Erro Tipo II. O Método das Variáveis Dummy foi o mais eficiente para os três tamanhos de amostra, pois, apresentou os menores percentuais de Erro Tipo I e Erro Tipo II.

Palavras-chave: Identidade de Modelos, Variáveis Dummy, Análise de Variância, Simulação.
Área do Conhecimento: I - Ciências Exatas e da Terra

Introdução

Quando se têm várias equações predizendo valores de uma mesma variável em condições distintas, algumas situações podem ser consideradas: As equações de regressão podem ser consideradas idênticas? Existirá uma equação comum para representar o conjunto? Os coeficientes de regressão dos vários conjuntos são estimadores de um mesmo coeficiente populacional? De que forma diferem as equações? Análises referentes a essas situações são comuns e de fundamental importância nas áreas de experimentação agropecuária, econometria e biometria florestal. Para realizar comparações entre equações de regressão, existem diversos métodos. Entre eles, destacam-se Identidade de Modelos, Variáveis Dummy (binárias), Análise de Variância e Comparações Múltiplas.

O presente trabalho teve por objetivo avaliar os métodos da Identidade de Modelos, das Variáveis Dummy (binárias) e da Análise de Variância, utilizados para a comparação entre equações de regressão lineares e quadráticas e/ou de seus coeficientes, empregando a simulação de dados. Pela padronização de rotinas e de teste, pretende-se verificar se existem divergências entre os métodos em estudo e suas aplicações práticas.

A comparação de H modelos lineares simples foi proposta através de um método detalhado para a verificação da identidade de modelos [1]. Considerando este trabalho e ajustando dados de observações relativos a H equações de regressão polinomial do segundo grau, utilizou-se, posteriormente, a técnica dos polinômios ortogonais [2].

Muitos autores priorizam a utilização de variáveis binárias, também mencionadas como variáveis *dummy*, indicadoras ou classificatórias, para testar a igualdade de equações ou coeficientes. As variáveis binárias podem ser definidas de várias formas e que a escolha da definição, ou da forma mais conveniente, depende das características do problema e das hipóteses que se deseja testar [3].

Alguns autores utilizaram a análise de variância seguida de procedimentos para comparações múltiplas no estudo da comparação de modelos de regressão. Os testes de comparações múltiplas, ou testes de comparações de médias, servem como um complemento do teste F e são adequados para detectar quais são as diferenças existente entre os tratamentos [4]. A análise de variância seguida de teste de aplicação

de médias, também foi utilizada para verificar a formação de grupos semelhantes, ao ajustar equações de volume para quatro espécies com alto valor de importância para uma floresta semidecídua montana na região de Lavras, MG [5].

Algumas comparações de coeficientes de regressão em situações com três ou mais grupos foram feitas através da simulação de dados de rotinas desenvolvidas no sistema computacional SAS® [6].

Materiais e Métodos

A metodologia apresentada neste trabalho foi aplicada por meio de um estudo de simulação de dados, com a geração de distribuições comportadas nas suas propriedades. O objetivo foi o de comparar os três métodos estatísticos, ou seja, identidade de modelos, variáveis dummy e análise de variância, que são muito utilizados na comparação de coeficientes e/ou equações de regressão.

Por meio de comparações mais detalhadas entre as metodologias, com o objetivo de padronizar as rotinas de testes e estimativas que são realizadas na prática, pretendeu-se verificar se existem divergências entre os métodos aplicados. Procedeu-se à verificação e comparação dos percentuais de taxas de Erro Tipo I e Erro Tipo II, em todas as situações de regressão linear simples e de regressão linear quadráticas consideradas.

A notação foi apresentada de forma matricial e foram utilizados os recursos do módulo Interactive Matrix Language (IML), do sistema Statistical Analysis System (SAS), para implementação da metodologia, proporcionando uma maior facilidade na sua aplicação.

Para a avaliação dos métodos, foram considerados quatro casos de regressão linear, representados por (a) o caso mais geral, quando todos os coeficientes são diferentes; (b) regressões paralelas, quando as inclinações são iguais, mas os interceptos são diferentes; (c) regressões concorrentes, quando os interceptos são iguais, mas as inclinações são diferentes e (d) regressões coincidentes, quando todas as retas coincidem. Consideraram-se também cinco casos de regressão polinomial quadrática, sendo (a) o caso mais geral, quando todos os coeficientes são diferentes; (b) regressões que possuem o mesmo intercepto; (c) regressões que possuem o mesmo coeficiente relativo ao termo de 1° grau; (d) regressões que possuem o mesmo coeficiente referente ao termo de 2° grau e (e) regressões coincidentes, quando todas as curvas coincidem.

Identidade de Modelos

Testa a hipótese de igualdade de um conjunto de modelos lineares utilizando o teste F, em que define-se um modelo completo e um modelo reduzido, sendo que o teste é aplicado sobre a redução que o modelo reduzido provoca no modelo completo.
Considera H modelos de regressão e testa as seguintes hipóteses:

(a)H_0:as H equações são idênticas;
(b)H_0:as H equações têm uma constante de regressão comum;
(c)H_0:as H equações têm um e/ou mais coeficientes de regressão iguais.

Variáveis dummy

A inclusão de variáveis binárias aditivas ou multiplicativas, permite verificar se duas equações lineares diferem em intercepto, em inclinação, ou ainda em ambos.
Seja a seguinte relação, referente a dois conjuntos de dados:

$$y_i = \alpha_0 + \alpha_1 D + \alpha_2 x_i + \alpha_3 (D x_i) + e_i \qquad i = 1,...,(n_1 + n_2)$$

em que:

D=1 para observações do primeiro conjunto (n_1 observações)

D=0 para observações do segundo conjunto (n_2 observações)

As variáveis binárias foram introduzidas na forma aditiva e multiplicativa. Os coeficientes a_1 e a_3 são diferenças de interceptos e inclinações, respectivamente.

Se H_0: a_1=0 é rejeitada, ou seja, a_1 é significativo, então o valor do intercepto do primeiro conjunto é obtido por a_1+a_0 , neste caso a_0 é o intercepto do segundo conjunto. Se H_0: a_1=0 não é rejeitada, ou seja, a_1 é não significativo, então a_0 representa o intercepto comum para ambos os conjuntos.

Se H_0: a_3=0 é rejeitada, então o valor da inclinação do primeiro conjunto é obtido por a_2 + a_3, neste caso a_2 é a inclinação do segundo conjunto. Se H_0: a_3=0 não é rejeitada, então a_2 representa a inclinação comum para ambos os conjuntos.

Simulação dos métodos

Dadas as seguintes relações lineares

$$y_{1i} = \beta_{01} + \beta_{11} x_{11i} + \varepsilon_{1i}$$
$$y_{2i} = \beta_{02} + \beta_{12} x_{12i} + \varepsilon_{2i}$$
:
$$y_{hi} = \beta_{0h} + \beta_{1h} x_{1hi} + \varepsilon_{hi} \quad \text{em que} \quad h=1,2.$$

e polinomiais quadráticas

$$y_{1i} = \beta_{01} + \beta_{11} x_{11i} + \beta_{21} x_{21i} + \varepsilon_{1i}$$
$$y_{2i} = \beta_{02} + \beta_{12} x_{12i} + \beta_{22} x_{22i} + \varepsilon_{2i}$$
:
$$y_{hi} = \beta_{0h} + \beta_{1h} x_{1hi} + \beta_{2h} x_{2hi} + \varepsilon_{hi} \quad \text{em que h=1,2.}$$

em que:

y_{hi}: i-ésima observação da variável resposta do h-ésimo modelo, sendo i = 1, 2,..., n_h o número de observações e h = 1,2 o número de modelos;

x_{1hi}, x_{2hi}: i-ésimo valor das variáveis regressoras do h-ésimo modelo;

β_{0h}, β_{1h}, β_{2h}: coeficientes do h-ésimo modelo;

ε_{hi}: erro aleatório, associado à i-ésima observação do h-ésimo modelo, sendo supostos independentes e normalmente distribuídos, com média zero e variância comum, isto é,

ε_{hi} ~ NID (0, σ^2) , $\sum_{h=1}^{H} n_h = N$.

Realizou-se uma simulação de dados composta de 10.000 experimentos, cada qual com 10, 50 e 100 observações para cada uma das situações descritas anteriormente.

Para cada experimento, foram gerados modelos de regressão nos quais os valores das variáveis independentes foram obtidas em um intervalo fechado de 0 a 10, aleatoriamente, pela função RANUNI do sistema SAS®.

Para a geração dos resíduos de cada modelo, foi necessário estimar a variância dos mesmos. Fixando-se o coeficiente de determinação R^2 em 90 %, e conhecida a relação $R^2 = \frac{\delta^2_{\text{modelo}}}{\delta^2_{\text{modelo}} + \delta^2_{erro}}$, em que δ^2_{modelo} corresponde à média dos valores das variáveis dependentes, estimou-se a variância dos resíduos δ^2_{erro} .

Estimada a variância dos resíduos δ^2_{erro}, geraram-se pela função RANNOR do sistema SAS®, os resíduos aleatórios de cada modelo. Estes são, supostamente, independentes e normalmente distribuídos, com média zero e variância comum, isto é, ε_{hi} ~ NID (0, δ^2_{erro}).

Com base nos modelos de regressão considerados, e fixando-se os parâmetros de cada modelo para cada uma das situações descritas anteriormente para a comparação dos três métodos, foram implementados computacionalmente os métodos da identidade de modelos, variáveis dummy e análise de variância, pelo módulo IML do sistema SAS® [7].

Ressalta-se que, para a implementação computacional do Método da Análise de Variância, foram considerados dois tratamentos conforme a Tabela 1, utilizou-se o Procedimento GLM do sistema SAS®. Os tratamentos para a implementação deste método foram constituídos pelos dois modelos (h=2) e os dados analisados foram os valores preditos pelas equações. Com base nesses dados foi realizada a Análise de Variância, considerando o delineamento inteiramente casualizado.

Tabela 1 – Tratamentos definidos para a metodologia da análise de variância

Tratamentos	Descrição
1	Equações ajustadas ao primeiro modelo
2	Equações ajustadas ao segundo modelo

Resultados

Os resultados foram analisados com base nos procedimento FREQ do módulo BASE, do Statistical Analysis System (SAS).

Para os casos de regressão linear simples e de regressão polinomial quadrática foram determinadas as freqüências dos resultados obtidos para os níveis de significância nominal. Estes resultados foram encontrados para os valores do teste F nos modelos para amostras de tamanho 10, 50 e 100, respectivamente.

A avaliação dos métodos da Identidade de Modelos, das Variáveis Dummy e da Análise de Variância baseou-se no nível nominal de 5 % dos percentuais das taxas de ocorrência do Erro Tipo I, que consiste na rejeição de uma hipótese H_0 tida verdadeira e nos percentuais das taxas de ocorrência do Erro Tipo II, que consiste na não-rejeição de uma hipótese inicial H_0, tida como falsa.

A Tabela 2 ilustra todas as nove situações simuladas utilizando-se os três métodos em estudo. Pode-se notar que, de modo geral, foram percebidas maiores taxas de Erro Tipo I e Erro Tipo II nos casos em tamanho da amostra é igual a 50 observações, com uma aparente vantagem para o Método das Variáveis Dummy.

Tabela 2 – Distribuição de freqüências de Erro Tipo I e Erro Tipo II para os métodos utilizados nos 10.000 experimentos simulados

MÉTODOS									
Casos	Identidade Modelos (Nº.de Observações)			Variáveis Dummy (Nº.de Observações)			Análise de Variância (Nº.de Observações)		
	10	50	100	10	50	100	10	50	100
Linear									
a	281	126	119	215	101	103	102	39	139
b	132	47	29	103	38	15	47	61	36
c	132	48	31	104	47	22	48	59	190
d	153	102	1	301	85	0	257	83	8
Subtotal	698	323	180	723	271	140	454	242	373
Quadrático									
a	95	85	976	106	72	861	508	367	143
b	31	174	467	26	156	396	174	138	118
c	39	50	39	33	41	21	166	202	296
d	96	26	142	21	14	118	26	668	579
e	46	41	42	39	43	18	14	56	69
Subtotal	307	376	1666	225	326	1414	888	1431	1205
Total	1005	699	1846	948	597	1554	1342	1673	1578
Total Geral		3550			3099			4593	

Discussão

Esperava-se que, com o aumento do número de observações uma redução nas taxas de Erro Tipo I e Tipo II. Mas, este fato, em geral, não ocorreu. Por exemplo, para o Método das Variáveis Dummy, verificaram-se menores taxas com tamanho de amostra de 50 observações. Em geral, amostras com 50 observações apresentaram menores taxas de erros, mas estes valores não são bem diferentes dos valores dos outros tamanhos de amostras. Isto porque seus valores médios foram 1,22 % para amostra de tamanho 10; 1,09 % para amostra de tamanho 50 e 1,84 % para amostra de tamanho 100.

Pôde-se também verificar que para todas as nove situações estudadas, em todas elas foram observados indícios uma boa precisão para os três métodos estudados. Contudo, deve-se ressaltar que, para o Método das Variáveis Dummy, obteve-se menor probabilidade de ocorrência de Erro Tipo I e de Erro Tipo II.

Conclusão

Os métodos da Identidade de Modelos, das Variáveis Dummy e da Análise de Variância sinalizam para a resultados bem semelhantes, devido a baixos percentuais de Erro Tipo I e Erro Tipo II.

Deve-se ressaltar que para todas as nove situações simuladas, para os três tamanhos de amostras, o Método das Variáveis Dummy, apresentou-se mais eficiente. Pois, o mesmo apresentou os menores percentuais de Erro Tipo I e Erro Tipo II.

Sugere-se um estudo bem mais detalhado, no qual deve-se aumentar o número de amostras, com o objetivo de encontrar um tamanho mínimo de amostra que minimize os percentuais de erros. Deve-se também estender a comparação entre os métodos da Identidade de Modelos, das Variáveis Dummy e da Análise de Variância a outros modelos, como, por exemplo, modelos não-lineares e modelos aplicados a algum comportamento biológico.

Referências

[1] GRAYBILL, F. A. Theory and application of the linear model. Belmont: Duxbury Press, 1976. 704p.

[2] REGAZZI, A. J. Teste para verificar a identidade de modelos de regressão e a igualdade de alguns parâmetros num modelo polinomial ortogonal. Revista Ceres, Viçosa, v. 40, n. 228, p. 176-195, mar./abr. 1993.

[3] HOFFMANN, R.; VIERA, S.Análise de regressão: uma introdução à econometria. 3 ed. São Paulo: Hucitec, 1998. 379p.

[4] BANZATTO, D. A.; KRONKA, S. N. Experimentação agrícola. Jaboticabal: Funep, 1995. 247p.

[5] SCOLFORO, J. R.; MELLO, J. M. de; LIMA, C. S. Obtenção de relações quantitativas para estimativa do volume de fuste em floresta estacional semidecídua montana. Revista Cerne, Lavras, v. 1, n. 1, p. 123-134, 1994.

[6] MITCHELL, M. How can I compare regression coefficients across 3 (or more) groups. 2000. Disponível em: <http://www.ats.ucla.edu/stat/sas/faq>. Acesso em: 18 set. 2000.

[7] SAS® INSTITUTE. SAS Procedures guide for computers. 6. ed. Cary N. C.: SAS® Institute, 1999. v. 3, 373p.