COMPARAÇÃO ENTRE MODELOS DE REGRESSÃO LINEARES APLICADOS À ÁREA MÉDICA

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

[1]Universidade Vale do Rio Verde de Três Corações - UNINCOR / Instituto de Ciências da Saúde (INCIS), Rua Capri, 27 – Arquipélago Verde 32000-000– Betim – MG – Brasil.
sergio@betim.unincor.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 a eficiência dos métodos estatísticos da Identidade de Modelos e das Variáveis Dummy (binárias), usados para comparar modelos de regressão. Utilizaram-se dados coletados no período de 2004 a 2005 provenientes de uma amostra de doadores de sangue de ambos os sexos de um Hospital Universitário e ajustaram-se retas de regressão da variável pressão sanguínea sistólica versus idade através do programa de análise estatistica SAS®. Foram considerados quatro casos de regressão lineares; a) Interceptos diferentes e inclinações iguais, b) Interceptos iguais e inclinações diferentes, c) Interceptos e inclinações diferentes e d) Interceptos e inclinações iguais. Concluiu-se que a aplicação dois métodos sinalizaram a resultados equivalentes.

Palavras-chave: Regressão linear, Identidade de Modelos, Variáveis Dummy.
Área do Conhecimento: Ciências Exatas e da Terra

Introdução

Segundo Draper e Smith (1998), a análise de regressão é uma técnica de análise de dados muito utilizada, nos casos em que se deseja estudar a relação entre variáveis quantitativas - uma variável resposta e uma ou mais variáveis explicativas.

Com muita freqüência, o estudo é feito em diferentes tratamentos para comparar seus efeitos, podendo ser realizado em diferentes fatores de classificação como épocas ou locais. Para cada uma dessas situações a análise de regressão pode ser aplicada separadamente obtendo-se tantas equações quantas forem o número de situações distintas.

Em dados da área médica, normalmente a variável resposta Y e o conjunto de variáveis regressoras X_i, i=1, 2,...,n são medidas em um conjunto composto de grupos distintos, visando a comparação de como estes diferem segundo a relação de X_i e Y, conforme ressaltou Seber (1977). Isto pode ser realizado, a partir da geração de modelos de regressão para cada grupo e, em seguida verificando se as equações correspondentes são paralelas, ou se tem intercepto comum, ou ainda, se são idênticas.

Deste modo, o presente estudo teve como objetivo aplicar os métodos estatísticos da Identidade de Modelos proposto por Graybill (1976) e das Variáveis Dummy discutido por Gujarati (1970a), utilizados nas comparação de modelos de regressão, a uma amostra de dados de doadores de sangue.

Materiais e Métodos

Utilizaram-se dados coletados no período de 2004 a 2005 provenientes de uma amostra de doadores de sangue do Hemocentro do Hospital Universitário Mário Penna da Universidade Vale do Rio Verde de Belo Horizonte, em pacientes de ambos os sexos.

Para a realização da comparação das metodologias propostas, ajustaram-se retas de regressão da variável pressão sanguínea sistólica versus idade, para uma amostra dse 1500 homens e 1500 mulheres, através do programa de de análise estatistica SAS®[5] e, considerou-se os casos a seguir:

a) Interceptos diferentes e inclinações iguais;
b) Interceptos iguais e inclinações diferentes;
c) Interceptos e inclinações diferentes;
d) Interceptos e inclinações iguais.

Explorou-se dois métodos para a comparação de Modelos de regressão linear, aos quais foram aplicados testes de hipóteses para identificação das situações acima.

Método da Identidade de Modelos

Trata-se de um teste bastante geral que verifica a igualdade de duas regressões lineares, cujo algoritmo segue os seguintes passos:

1. Dadas as seguintes relações lineares:

$$y_{1i} = a_1 + b_1 x_{1i} + e_{1i} \qquad i = 1,...,n_1$$
$$y_{2i} = a_2 + b_2 x_{2i} + e_{2i} \qquad i = 1,...,n_2 \qquad (1)$$

referentes a dois conjuntos de observações.

2. Combinam-se todas as $n_1 + n_2$ observações e calcula-se a estimativa de quadrados mínimos de a e b na regressão combinada $y = a + bx + e$. Desta equação obtém-se a soma de quadrados de resíduo (S_1) com grau de liberdade igual a $n_1 + n_2 - p$, em que p é o número de parâmetros a ser estimado. Neste caso, p = 2.

3. Obtém-se a soma de quadrados de resíduo para as duas equações, ou seja, S_2 e S_3, com os graus de liberdade $n_1 - p$ e $n_2 - p$, respectivamente. Somam-se estas duas somas de quadrados de resíduo, isto é, $S_4 = S_2 + S_3$ e seus graus de liberdade $n_1 + n_2 - 2p$.

4. Obtém-se $S_5 = S_1 - S_4$.

5. Calcula-se a estatística F como:

$$F_c = \frac{S_5 / p}{S_4 / (n_1 + n_2 - 2p)} \qquad (2)$$

com p e $n_1 + n_2 - 2p$ graus de liberdade.

Se F_c > F tabelado, para um determinado nível de significância α, rejeita-se a hipótese de que os parâmetros $a's$ e $b's$ são os mesmos para os dois conjuntos de observações.

Método das 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:

$$y_i = \alpha_0 + \alpha_1 D + \alpha_2 x_i + \alpha_3 (D x_i) + e_i$$
$$i = 1,...,(n_1 + n_2) \qquad (3)$$

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.

Resultados

Para aplicar o Método da Identidade de Modelos, ajustou-se as retas para cada um dos sexos:

Masculino: $\hat{Y}_{mas} = 99{,}81 + 0{,}48x$

Feminino: $\hat{Y}_{fem} = 105{,}14 + 0{,}37x$

E, as estimativas dos parâmetros para os dois sexos foram registrados na Tabela 1.

Tabela 1: Estimativas dos parâmetros para os modelos estimados idade *versus* pressão-sistólica

Grupo	$\hat{\beta}_0$	$\hat{\beta}_1$	$\bar{x}$	s_x^2	$s_{Y/X}^2$
Masculino	99,81	0,48	31,08	105,21	328,25
Feminino	105,14	0,37	31,05	115,44	254,81

Para o Método das Variáveis Dummy, ajustou-se um modelo de regressão de todo o conjunto e, em seguida, este foi separado, originando um modelo para cada sexo, através da inclusão das variáveis dummy.

$$D = \begin{cases} 0, & \text{Se o indivíduo é homem} \\ 1, & \text{Se o indivíduo é mulher} \end{cases} \qquad (4)$$

Reta Geral: $\hat{Y} = 100{,}11 + 0{,}52x + 12{,}67D - 0{,}04xD$

Reta Ajustada - Masculino:

$\hat{Y}_{mas} = 100{,}01 + 0{,}52x \ \ (D = 0)$

Reta Ajustada - Feminino:

$$\hat{Y}_{fem} = 113,41 + 0,49x \quad (D=1)$$

Na Tabela 2 é apresentada a análise de variância do ajuste de retas para esta situação.

Tabela 2: ANOVA pelo Método das Variáveis Dummy para a variável idade *versus* pressão-sistólica

Fonte de variação	GL	SQ	QM	F
Regressão (x)	1	77071,12	7707,12	20,60
Resíduo	3005	1124445,00	374,19	
Regressão (x,D)	2	926547,00	463273,5	245,81
Resíduo	3004	1428954,00	475,70	
Regressão (x,d,xD)	3	155768,00	51912,67	165,84
Resíduo	3003	926158,00	308,41	

Discussão

Considerando-se o Método da Identidade de Modelos, procedeu-se a identificação dos casos em que as retas estimadas se enquadravam nos testes do paralelismo e no da igualdade dos parâmetros conforme detalhamento abaixo.

a) Teste do paralelismo:

$$H_0 = \beta_{1mas} = \beta_{1fem}$$

$$S^2_{P,Y/X} = 301,25 \text{ e } S^2_{\hat{\beta}_{1mas} - \hat{\beta}_{1fem}} = 0,04.$$

A estatística de teste foi T=0,61. Para esta estatística, o valor crítico bilateral dado pelo p-value=2P(T≥|0,61|)=0,55. Considerando-se o nível de significância nonimal α igual a 5%, observou-se que o valor p-value > α. Portanto, a hipótese de nulidade não foi rejeitada, ou seja, tiveram evidências amostrais suficientes para que a hipótese de paralelismo não fosse rejeitada.

b) Teste da igualdade de interceptos:

$$H_0 = \beta_{0mas} = \beta_{0fem}$$

$$S^2_{P,Y/X} = 301,25 \text{ e } S^2_{\hat{\beta}_{0mas} - \hat{\beta}_{0fem}} = 5,01.$$

A estatística de teste foi T=-5,61. Para esta estatística, o valor crítico bilateral dado pelo p-value=2P(T≥|-5,09)≅0. Portanto, a hipótese de nulidade foi rejeitada para quaisquer níveis de significância nominal de α. Houveram fortes evidências amostrais de que a hipótese igualdade de interceptos não seja verdadeira.

Em contrapartida, considerando-se o Método das Variáveis Dummy, procedeu-se a identificação dos casos em que as retas estimadas se enquadravam nos testes do paralelismo, no da igualdade dos parâmetros e no teste da coincidência, discriminados a seguir.

a) Teste do paralelismo:

$$H_0 = \beta_3 = 0$$

A estatística de teste foi F(XD/X,D)=0,52. O p-value com 1 e 3003 graus de liberdade foi igual a 0,46. Portanto, não foi rejeitada a hipótese de nulidade H_0 para quaisquer valores nominais de α. Logo, não existiram evidências amostrais para que a hipótese de paralelismo das regressões lineares fosse rejeitada

b) Teste da igualdade de interceptos:

$$H_0 = \beta_2 = 0$$

A estatística de teste foi F(D/X,XD)=253,25. O p-value com 1 e 3003 graus de liberdade foi aproximadamente igual a zero. Portanto, foi rejeitada a hipótese de nulidade H_0 para quaisquer valores nominais de α diferentes de zero. Logo, perceberam evidências amostrais para que a hipótese de igualdade de interceptos das equações lineares dos dois sexos não fosse verdadeira.

c) Teste da coincidência:

$$H_0 = \beta_2 = \beta_3 = 0$$

A estatística de teste foi F(D/X)=121,68. O p-value com 2 e 3003 graus de liberdade foi aproximadamente igual a zero. Portanto, foi rejeitada a hipótese de nulidade H_0 para quaisquer valores nominais de α diferentes de zero. Logo, não notaram evidências amostrais para que a hipótese de coincidência das regressões lineares estimadas para ambos os sexos.

Conclusão

Verificou-se que para a amostra de dados referentes a pressão sanguínea sistólica e idade, submetidos às duas metodologias do estudo, revelaram que as retas estimadas para o sexo masculino e para o sexo feminino não foram coincidentes. As mesmas foram paralelas, com

interceptos diferentes e admitiram a forma $Y=\beta_0+\beta_1 x+\varepsilon$.

Foi possível verificar que a aplicação do Método da Identidade de Modelos foi equivalente ao Método das Variáveis Dummy.

Agradecimentos

Agradecemos ao Hemocentro do Hospital Escola Mário Penna, da Universidade Vale do Rio Verde de Belo Horizonte, pela cessão do dados analisados neste estudo.

Referências

- DRAPER, N. R.; SMITH, H. **Applied regression analysis.** 2. ed. New York: John Wiley & Sons, 1998. 709p.

- GRAYBILL, F. A. **Theory and application of the linear model**. Belmont: Duxbury Press, 1976. 704p.

- GUJARATI, D. "Use of dummy variables in testing for equality between sets of coefficients in linear regressions: a generalization." **The American Statistician**. Washington, v. 24, n. 5, p. 18-22, Dec. 1970a.

- SEBER, G. A. F. **Linear regression analysis.** New York: John Wiley, 1977. 465p.

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