Comparando equações de regressão em dados de saúde

Terezinha Aparecida Guedes*, Ivan Ludgero Ivanqui e Ana Beatriz Tozzo Martins

Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo, 5790, 87020-900, Maringá, Paraná, Brasil. Author for correspondence. e-mail: taguedes@uem.br

RESUMO. Em muitas situações, a variável resposta Y e o conjunto de variáveis regressoras são medidas em um conjunto composto de grupos distintos. O objetivo é o de examinar como os grupos diferem entre si pela relação entre Y e X_i, determinando se o conjunto de curvas de regressão são paralelas, ou se tem o intercepto comum ou se são idênticas. Foram utilizadas variáveis *dummy* para identificar os grupos no conjunto de dados. Kleinbaum *et al.* (1998), Krzanowski (1998), Neter, J., *et al*. (1996) e Seber, G. A. F. (1977) comparam várias equações de regressão pelo uso de modelos de regressão linear múltipla. Tal método fornece as mesmas informações que os obtidos com a análise de covariância e de variância. Esta metodologia foi aplicada aos dados de uma amostra de doadores do Hemocentro do Hospital Universitário de Maringá no período de 1995 a 1996.

Palavras-chave: regressão linear múltipla, comparação de regressão, variáveis *dummy* em regressão.

ABSTRACT. Comparison of regression equations in health data. In many situations, the variable answer Y and the set of regression variables are measured in a set composed of different groups. Authors' objective is to examine how groups differ among themselves by their Y and X_i relationship and to determine whether the set of regression curves are parallel, whether they have a common intercept or whether they are identical. Dummy of variables is used to identify groups in data set. Kleinbaum *et al.* (1998), Krzanowski (1998), Neter, J., *et al*. (1996) and Seber, G. A. F. (1977) compare several regression equations by models of multiple linear regression. The method supplies the same information as that obtained with covariance and variance analyses. Methodology was applied to data of blood donors at the Blood Bank of the University Hospital of Maringá 1995-96.

Key words: multiple linear regression, regression comparison, variable *dummy* in regression.

Em dados de saúde, geralmente 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. O objetivo é o de examinar como os grupos diferem entre si pela relação entre Y e X_i. Isto pode ser realizado determinando se o conjunto de equações de regressão são paralelas, ou se tem o intercepto comum ou ainda se são idênticas. Será feito o uso de variáveis *dummy* para identificar os grupos no conjunto de dados. Kleinbaum *et al.* (1998), Krzanowski (1998), Neter *et al*. (1996) e Seber (1977) apresentam esta metodologia para comparar várias equações de regressão pelo uso de modelos de regressão linear múltipla.

Segundo Kleinbaum *et al.* (1998), o uso de variáveis *dummy* permite expandir a aplicação de análise de regressão e, ainda, as variáveis *dummy* permitem que a análise de regressão seja empregada para produzir as mesmas informações obtidas por procedimentos analíticos aparentemente distintos tais como análise de covariância e de variância.

Material e métodos

Na maioria das aplicações de análise de regressão as variáveis preditoras são contínuas, mas os métodos de análise de regressão podem ser generalizados para tratar variáveis preditoras categóricas. A generalização é baseada no uso de variáveis *dummy*.

Uma variável *dummy* ou indicadora é qualquer variável na equação de regressão que assume um número finito de valores de forma que diferentes categorias de uma variável nominal podem ser identificadas. A variável *dummy* será utilizada para comparar duas ou mais retas de regressão.

Quando se ajusta reta de regressão para cada um dos grupos formados pelas distintas categorias da variável preditora, pode-se obter um dos seguintes casos:

a) Interceptos diferentes, mas inclinação igual;

b) Interceptos iguais, mas inclinações diferentes;
c) Interceptos e inclinações diferentes;
d) Interceptos e inclinações iguais.

Serão explorados dois métodos de realizar a comparação das retas de regressão denominados método I e método II. Nestes métodos, serão aplicados testes de hipóteses para identificar os casos acima.

Método I. Usando ajustes separados para comparar duas retas.

Considere um conjunto de dados cuja variável independente apresente duas categorias de interesse. Para cada categoria será ajustada uma equação de regressão:

$Y_1 = \beta_{01} + \beta_{11}X + E$ e $Y_2 = \beta_{02} + \beta_{12}X + E$,

onde os erros E são independentes e identicamente distribuídos como uma N(0, σ^2).

a) Teste de paralelismo. A hipótese apropriada para comparar os coeficientes angulares é dada por: $H_0 : \beta_{11} = \beta_{12}$.

Quando H_0 é verdadeira as retas de regressão tornam-se: $Y_1 = \beta_{01} + \beta_1 X + E$ e $Y_2 = \beta_{02} + \beta_1 X + E$.

Uma estimativa para β_1 é dada por:

$$\hat{\beta}_1 = \frac{(n_1 - 1)s_1^2\hat{\beta}_{11} + (n_2 - 1)s_2^2\hat{\beta}_{12}}{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}$$

A estatística de teste é: $T = \dfrac{\hat{\beta}_{11} - \hat{\beta}_{12}}{s_{\hat{\beta}_{11} - \hat{\beta}_{12}}}$,

onde,

$\hat{\beta}_{11}$ é a estimativa de β_{11}, usando as n_1 observações do grupo 1;

$\hat{\beta}_{12}$ é a estimativa de β_{12}, usando as n_2 observações do grupo 2;

$S_{\hat{\beta}_{11} - \hat{\beta}_{12}}$ é a estimativa do desvio-padrão da diferença entre os coeficientes. Este desvio-padrão é igual a raiz quadrada da seguinte variância:

$$S^2_{\hat{\beta}_{11} - \hat{\beta}_{12}} = S^2_{P,\,Y/X}\left[\frac{1}{(n_1 - 1)S^2_{X_1}} + \frac{1}{(n_2 - 1)S^2_{X_2}}\right],$$

onde,

$$S^2_{P,\,Y/X} = \frac{(n_1 - 2)S^2_{Y/X_1} + (n_2 - 2)S^2_{Y/X_2}}{n_1 + n_2 - 4},$$

onde,

S^2_{Y/X_1} é o quadrado médio do resíduo para o grupo 1;

S^2_{Y/X_2} é o quadrado médio do resíduo para o grupo 2;

$S^2_{X_1}$ é a variância dos X para o grupo 1;

$S^2_{X_2}$ é a variância dos X para o grupo 2.

A estatística de teste, T, sob as suposições usuais de regressão, terá distribuição t de Student com n_1+n_2-4 graus de liberdade quando H_0 for verdadeira.

b) Teste de Igualdade de Interceptos. A hipótese nula para comparar dois interceptos é dada por: $H_0 : \beta_{01} = \beta_{02}$.

Se H_0 é verdadeira as retas de regressão tornam-se: $Y_1 = \beta_0 + \beta_{11}X + E$ e $Y_2 = \beta_0 + \beta_{12}X + E$.

Uma estimativa do intercepto é dada por: $\hat{\beta}_0 = \dfrac{n_1\hat{\beta}_{01} + n_2\hat{\beta}_{02}}{n_1 + n_2}$.

A estatística de teste é: $T = \dfrac{\hat{\beta}_{01} - \hat{\beta}_{02}}{S_{\hat{\beta}_{01} - \hat{\beta}_{02}}}$,

onde,

$\hat{\beta}_{01}$ é a estimativa de β_{01}, usando as n_1 observações do grupo 1;

$\hat{\beta}_{02}$ é a estimativa de β_{02}, usando as n_2 observações do grupo 2;

$S_{\hat{\beta}_{01} - \hat{\beta}_{02}}$ é a estimativa do desvio-padrão da diferença entre os interceptos e deve ser calculada por:

$$S^2_{\hat{\beta}_{01} - \hat{\beta}_{02}} = S^2_{P,\,Y/X}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{\overline{X}_1^2}{(n_1 - 1)S^2_{X_1}} + \frac{\overline{X}_2^2}{(n_2 - 1)S^2_{X_2}}\right].$$

c) Teste de coincidência. As hipóteses nulas, neste caso, são:

$H_0 : \beta_{01} = \beta_{02}$ e $H_0 : \beta_{11} = \beta_{12}$

Se as hipóteses nulas são verdadeiras as retas de regressão tornam-se: $Y = \beta_0 + \beta_1 X + E$.

Usando as retas de regressão ajustadas separadamente, se uma ou ambas as hipóteses nulas forem rejeitadas, conclui-se que não há evidencia amostral suficiente de que as duas retas sejam coincidentes.

Críticas sobre este procedimento de teste recaem sobre o poder do teste, pois envolve dois testes separados ao invés de um único teste originando as dificuldades:

i) O procedimento não é precisamente um teste para coincidência;
ii) Se α é o nível de significância de cada teste separado, o nível de significância geral para os dois testes combinados é maior que α; isto é, existe mais chance de rejeitar H_0 quando esta é verdadeira (cometer o erro tipo I).

Método II. Usando uma simples equação de regressão.

Utiliza-se uma ou mais variáveis *dummy* para distinguir os grupos.

Modelo geral: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$ (1)

onde Z é uma variável *dummy* indicando grupo 1 e grupo 2 e os erros E são independentes e identicamente distribuídos como uma $N(0, \sigma^2)$.

Assim, $\begin{cases} Z = 0, & y_1 = \beta_0 + \beta_1 X + E \\ Z = 1, & y_2 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + E \end{cases}$

a) Teste de paralelismo. A hipótese nula de que as duas retas de regressão são paralelas é:

$H_0 : \beta_3 = 0$.

O teste estatístico é o teste F parcial para a significância da adição da variável XZ ao modelo já contendo X e Z.

$$F(XZ/X, Z) = \frac{SQReg(X,Z,XZ) - SQReg(X,Z)}{QMRes(X,Z,XZ)}.$$

b) Teste de igualdade de intercepto. A hipótese nula para o modelo geral é :$H_0 : \beta_2 = 0$.

1) Procedimento 1. O teste compara o modelo geral ao modelo reduzido:

$Y = \beta_0 + \beta_1 X + \beta_3 XZ + E$.

A estatística de teste é: $F(Z/X, XZ) = \dfrac{SQReg(X,Z,XZ) - SQReg(X,XZ)}{QMRes(X,Z,XZ)}$.

2) Procedimento 2. O teste envolve a comparação do modelo

$Y = \beta_0 + \beta_1 X + \beta_2 Z + E$ ao modelo reduzido $Y = \beta_0 + \beta_1 X + E$.

Este procedimento presume igualdade de coeficientes angulares, assim é um teste para coincidência, assumindo paralelismo.

A estatística de teste é: $F(Z/X) = \dfrac{SQReg(Z,X) - SQReg(X)}{QMRes(X,Z,XZ)}$.

c) Teste para coincidência. A hipótese que as duas retas de regressão coincidem para o modelo geral é: $H_0 : \beta_2 = \beta_3 = 0$.

Quando H_0 é verdadeira, o modelo para o grupo 2, $Y_2 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + E$, reduz- se a $Y_1 = \beta_0 + \beta_1 X + E$, o modelo para o grupo 1.

O teste para a hipótese nula é um teste F parcial múltiplo, pois envolve um subconjunto de coeficientes de regressão. Os dois modelos que serão comparados são:

$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$ e $Y = \beta_0 + \beta_1 X + E$.

A estatística de teste será: $F(Z/X) = \dfrac{[SQReg(X,Z,XZ) - SQReg(X)]/2}{QMRes(X,Z,XZ)}$.

Comparando os dois métodos

Os dois métodos produzem as mesmas estimativas para coeficientes de regressão das duas retas. Isto é, se no modelo geral o método de mínimos quadrados for utilizado para obter as estimativas dos coeficientes, as equações obtidas fazendo Z=0 e Z=1, neste modelo, serão iguais as obtidas pela estimação das retas ajustadas separadas.

Como os testes estatísticos nos dois métodos envolvem coeficientes estimados, os pontos seguintes são válidos:

a) Os testes de paralelismo são equivalentes; isto é, a estatística F calculada pelo método II é a mesma que a estatística T calculada pelo método I.
b) Os testes para coincidência são diferentes. O procedimento de ajuste de equações de regressão separadas testa a hipótese H_0: $\beta_{11} = \beta_{12}$ e H_0: $\beta_{01} = \beta_{02}$ separadamente e então rejeita a hipótese nula de retas coincidentes se uma ou ambas as hipóteses forem rejeitadas. Isto é equivalente a realizar dois testes separados de H_0: $\beta_2 = 0$ e H_0: $\beta_3 = 0$ e usar a mesma regra de decisão para procedimento com variáveis *dummy*, mas não é equivalente a testar uma única hipótese nula H_0: $\beta_2 = \beta_3 = 0$, ou seja, testar se ambos são simultaneamente zero.

Resultados e discussão

Aplicação. A idade e a pressão sangüínea sistólica dos doadores de sangue do Hemocentro do Hospital Universitário de Maringá, Estado do Paraná, no

período de 1995 a 1996 foram coletadas com objetivo de verificar se estas variáveis têm relação linear semelhante para ambos os sexos. Para realizar a comparação, foram ajustadas retas de regressão da pressão sistólica versus idade, para uma amostra aleatória de 809 homens e 2200 mulheres, através do programa estatístico SAS (Statistical Analysis System), seguindo a metodologia acima discutida.

Para aplicar o **Método I**, foram ajustadas retas para cada um dos sexos:

Masculino: $\hat{Y}_M = 100{,}863 + 0{,}520X$

Feminino: $\hat{Y}_F = 113{,}610 + 0{,}479X$.

As informações consistindo das estimativas dos parâmetros para cada um dos sexos se encontram na Tabela 1.

Tabela 1. Estimativas dos parâmetros para as retas estimadas para os dados de idade-pressão sangüínea sistólica

Grupo	$\hat{\beta}_0$	$\hat{\beta}_1$	$\overline{X}$	s_X^2	$s_{Y/X}^2$
Feminino	113,610	0,479	32,082	107,259	329,592
Masculino	100,863	0,520	32,433	115,204	265,777

Para identificar em que caso as retas estimadas se enquadram os testes para paralelismo e de igualdade de interceptos foram realizados.

a) Teste de Paralelismo: $H_0 : \beta_{1M} = \beta_{1F}$.

$S^2_{P,\,Y/X} = 312{,}454$ e $S^2_{\hat{\beta}_{1M} - \hat{\beta}_{1F}} = 0{,}00468$

A estatística de teste é T=0,599. Para esta estatística, o valor crítico bilateral é dado por p-valor= $2P(T \geq |0{,}599|) = 0{,}549$. Considerando o nível de significância α igual a 5%, observa-se que o p-valor > α. Logo a hipótese nula não será rejeitada, ou seja, há evidência amostral suficiente para que a hipótese de paralelismo não seja rejeitada.

b) Teste de igualdade de Interceptos: $H_0 : \beta_{0M} = \beta_{0F}$.

$S^2_{P,\,Y/X} = 312{,}454$ e $S^2_{\hat{\beta}_{0M} - \hat{\beta}_{0F}} = 5{,}422$

A estatística de teste é T=-5,474. Para esta estatística, o valor crítico bilateral é dado por p-valor $= 2P(T \geq |-5{,}08|) \cong 0$. Portanto, a hipótese nula é rejeitada para qualquer valor do nível de significância α. Há forte evidência amostral de que a hipótese de igualdade de interceptos não seja verdadeira.

Para aplicar o **Método II** foi ajustada uma equação de regressão de todo o conjunto de dados e, em seguida, esta foi separada em duas, uma para cada um dos sexos, através do uso de uma variável *dummy*:

$$Z = \begin{cases} 0, & \text{se o indivíduo é homem} \\ 1, & \text{se o indivíduo é mulher} \end{cases}.$$

Reta geral : $\hat{Y} = 100{,}443 + 0{,}531X + 13{,}167Z - 0{,}052XZ$

Retas ajustadas: $\hat{Y}_M = 100{,}443 + 0{,}531X$ (Z=0);

$\hat{Y}_F = (100{,}443 + 13{,}167) + (0{,}531 - 0{,}052)X$ (Z=1).

A análise de variância do ajuste das retas é apresentada na Tabela 2.

Tabela 2. ANOVA pelo método II para idade-pressão sangüínea sistólica.

Fonte de variação	GL	SQ	QM	F
Regressão (X)	1	77071,00	77071,00	228,00
Resíduo	3007	1016470,00	338,03	
Regressão (X, Z)	2	155686,00	77843,00	249,49
Resíduo	3006	936335,00	312,01	
Regressão (X, Z, XZ)	3	155863,00	51954,00	166,49
Resíduo	3005	936158,00	312,05	

a) Teste para Paralelismo: $H_0 : \beta_3 = 0$

A estatística de teste é F(XZ/X, Z)=0,57. O p-valor com 1 e 3005 graus de liberdade é igual a 0,45, logo não se rejeita H_0 para qualquer valor de α. Portanto, não há evidência amostral para que a hipótese de paralelismo seja rejeitada.

b) Comparando Interceptos; $H_0 : \beta_2 = 0$

A estatística de teste é F(Z/X, XZ) = 251,93. O p-valor para esta estatística com 1 e 3005 graus de liberdade é aproximadamente igual 0, logo, rejeita-se a hipótese nula para qualquer valor de α diferente de zero. Portanto, há evidência amostral de que a hipótese de igualdade de interceptos não é verdadeira.

c) Teste para coincidência: $H_0 : \beta_2 = \beta_3 = 0$.

A estatística de teste é F(Z/X) = 126,25. O p-valor com 2 e 3005 graus de liberdade é aproximadamente igual 0, logo rejeita-se a hipótese nula para qualquer valor de α diferente de zero. Portanto, não há evidência amostral de coincidência das retas estimadas para cada um dos sexos.

Para os dados de idade-pressão sangüínea sistólica os resultados obtidos da aplicação dos métodos I e II, revelaram que as retas estimadas para masculino e feminino não são coincidentes, são paralelas, tem interceptos diferentes e tem a forma: $Y = \beta_0 + \beta_1 X + E$.

Dos resultados obtidos pode-se observar que a aplicação do método I é equivalente a aplicação do método II.

Referências

KLEINBAUM, D. G. *et al*. *Applied regression analysis and other multivariable methods*. London: Duxbury Press, 1998.

KRZANOWSKI, W. *An introduction to statistical modelling*. London: Arnold, 1998.

NETER, J. *et al*. *Applied linear statistical models.* 4.ed. Chicago: McGraw-Hill Companies, Inc.1996.

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

Received on May 31, 2001.
Accepted on November 08, 2001.