

Testes para verificar a igualdade de parâmetros e a identidade de modelos de regressão não-linear em dados de experimento com delineamento em blocos casualizados

Adair José Regazzi¹, Carlos Henrique Osório Silva²

RESUMO

Considerou-se o ajustamento de equações de regressão não-linear e o teste da razão de verossimilhança, com aproximações pelas estatísticas qui-quadrado e F, para testar as hipóteses de igualdade de qualquer subconjunto de parâmetros e de identidade dos modelos para dados com repetições provenientes de experimento com delineamento em blocos completos casualizados. Concluiu-se que as duas aproximações podem ser utilizadas, mas a aproximação pela estatística F deve ser preferida, principalmente para pequenas amostras.

Palavras-chave: Comparação de equações não-lineares, razão de verossimilhança, teste de hipótese.

ABSTRACT

Tests for model identity and parameter equality with nonlinear regression models in data from randomized complete block design

We considered the adjustments of a set of nonlinear regression equations and the likelihood ratio test with approximations given by the χ^2 and also by the F statistics, in order to test model identity and equality of any subset of parameters, using data with repetitions from a randomized complete block design. We concluded that both approximations can be used but the F should be preferred for small sample sizes.

Key words: Hypothesis test, likelihood ratio, nonlinear equation comparisons.

Recebido para publicação em novembro de 2008 e aprovado em março de 2010

¹ Engenheiro-Agrônomo, Doutor. Departamento de Estatística, Universidade Federal de Viçosa, Av. Peter Henry Rolfs, s/n, 36570-000 Viçosa, MG, Brasil. adairreg@ufv.br.

² Engenheiro-Agrônomo, Ph. Doctor. Departamento de Estatística, Universidade Federal de Viçosa, Av. Peter Henry Rolfs, s/n, 36570-000 Viçosa, MG, Brasil. chos@ufv.br

INTRODUÇÃO

Em análise de regressão, com muita frequência deseja-se saber se um conjunto de equações ajustadas deve ser idêntico, ou seja, se a relação de variáveis em estudo pode ser representada por uma única equação. Vários autores apresentaram métodos para testar hipóteses relativas à identidade de modelos lineares, como, Graybill (1976), Neter *et al.* (1996), Regazzi (1993, 1999), entre outros. Bates & Watts (1988) apresentaram um teste assintótico baseado na razão de verossimilhança para comparar modelos de regressão não-linear como o utilizado para modelos lineares, com aproximação pela estatística F. Regazzi (2003) mostrou em detalhes a aplicação do teste da razão de verossimilhança (Rao, 1973) em modelos de regressão não-linear, com aproximação pela estatística qui-quadrado considerando a situação de somente um valor observado de Y para cada um de X. Regazzi & Silva (2004) utilizaram o teste da razão de verossimilhança com aproximações pelas estatísticas qui-quadrado e F para testar a hipótese de igualdade de qualquer subconjunto de parâmetros e, também, a identidade de modelos de regressão não-linear para dados com repetições de experimento com delineamento inteiramente casualizado. Concluíram que para um número total de observações suficientemente grande ($N > 120$) as duas aproximações são praticamente equivalentes. Para amostras menores, a aproximação pela estatística F deve ser preferida, uma vez que a taxa de erro tipo I foi sempre menor, independentemente do valor de N.

O principal objetivo do presente trabalho é mostrar como testar hipóteses sobre a igualdade de qualquer subconjunto de parâmetros e, também, a identidade de modelos de regressão não-linear a partir de dados de experimento com delineamento em blocos completos casualizados. Objetivou-se também fazer algumas considerações sobre o coeficiente de determinação, com maior ênfase em modelos de regressão não-linear.

MATERIAIS E MÉTODOS

Considerou-se um experimento com **I** níveis de um fator quantitativo e **r** repetições com delineamento em blocos completos casualizados, repetido em **H** diferentes condições experimentais. Para cada uma das **H** condições experimentais (**H** grupos) foi ajustado o modelo estatístico Michaelis-Menten a seguir:

$$y_{ij} = \frac{a_i x_{ij}}{b_i + x_{ij}} + \varepsilon_{ij}, j = 1, \dots, n \text{ e } i = 1, \dots, H \quad (1)$$

em que Y_{ij} é o valor observado na j-ésima unidade experimental do i-ésimo grupo; X_{ij} é o valor da variável independente associado a Y_{ij} ; a_i e b_i são parâmetros ($p = 2$); ε_{ij} é o erro aleatório com as pressuposições $\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$; $n = \mathbf{I}r$ é o número de observações de cada grupo

suposto constante; e, $n\mathbf{H} = \mathbf{N}$ é o número total de observações.

As hipóteses que serão consideradas são as seguintes:

1 - $H_0^{(1)} : b_1 = \dots = b_H (= b)$ vs. $H_a^{(1)} : \text{nem todos } b_i \text{ são iguais.}$

2 - $H_0^{(2)} : a_1 = \dots = a_H (= a)$ vs. $H_a^{(2)} : \text{nem todos } a_i \text{ são iguais.}$

3 - $H_0^{(3)} : b_1 = \dots = b_H (= b) \text{ e } a_1 = \dots = a_H (= a)$ vs. $H_a^{(3)} : \text{Não } HO_{(3)} \text{ e.}$

Para $u = 1, \dots, H$, sejam as variáveis indicadoras:

$$d_u = \begin{cases} 1 & \text{se a observação } y_{ij} \text{ pertence ao grupo } u, \\ 0 & \text{em caso contrário.} \end{cases}$$

Então, a equação (1) pode ser reescrita como:

$$y_{ij} = \sum_{u=1}^H d_u \left[\frac{a_u x_{ij}}{b_u + x_{ij}} \right] + \varepsilon_{ij}, j = 1, \dots, n \text{ e } i = 1, \dots, H$$

O problema é comparar as equações ajustadas individualmente para os grupos. O método para testar as hipóteses formuladas apresentadas a seguir baseia-se no teste da razão de verossimilhança, com aproximações pelas estatísticas χ^2 e F.

Inicialmente, realizou-se uma análise para cada grupo, segundo o esquema apresentado na Tabela 1.

A análise de regressão conjunta dos **H** grupos pressupõe a homocedasticidade dos resíduos das regressões individuais para os grupos. A verificação dessa pressuposição poderá ser feita mediante um teste de homogeneidade de variância, como o teste de Bartlett, citado por Li (1964). Conforme Pimentel-Gomes (2000), quando uma relação de variâncias (no caso máx. QMRR/min. QMRR, QMRR = quadrado médio do resíduo da regressão) é menor que sete, é possível combinar as variâncias residuais e adotar uma estimativa comum. Quando esse quociente for superior a sete, convirá considerar separadamente subgrupos de modelos, de modo que dentro de cada subgrupo se tenha a homogeneidade de variância.

Tabela 1 - Esquema da análise de variância e regressão para cada grupo

Causa de Variação	Graus de Liberdade
"Média"	1
Blocos	r - 1
Tratamentos	I - 1
Tratamentos não-corrigido	I
Parâmetros da regressão	p
Falta de ajustamento	I - p
Resíduo=Erro	(I-1)(r-1)
Resíduo da regressão	(I-p) + (I-1)(r-1) = n - p - (r-1)

Considerem-se as seguintes definições:

$SQR_{h(DBC)}$ = soma de quadrados do resíduo para o h-ésimo grupo, para o delineamento em blocos completos casualizados (DBC). Esse é o erro usual segundo o delineamento experimental utilizado;

$SQRR_{h(DIC)}$ = soma de quadrados do resíduo da regressão (falta de ajustamento mais o resíduo) para o h-ésimo grupo, como se houvesse sido adotado o delineamento inteiramente casualizado (DIC);

$SQRR_{h(DBC)}$ = soma de quadrados do resíduo da regressão (falta de ajustamento mais o resíduo) para o h-ésimo grupo, para o delineamento em blocos completos casualizados;

$SQBlocos_h$ = soma de quadrados de blocos para o h-ésimo grupo;

$SQPR_h$ = soma de quadrados de parâmetros da regressão para o h-ésimo grupo;

$SQTrat_h$ = soma de quadrados para tratamentos do h-ésimo grupo;

G_h = total geral para o h-ésimo grupo; e

$C_h = G_h^2/n$ é a correção para o h-ésimo grupo.

$SQRR_{h(DBC)} = SQRR_{h(DIC)} - SQBlocos_h$

$SQFalta\ de\ ajustamento_{h(DBC)} = SQRR_{h(DBC)} - SQPR_{h(DBC)}$

$SQPR_h + SQFalta\ de\ ajustamento_{h(DBC)} = SQTrat_h + C_h$, com I graus de liberdade.

Ω = espaço paramétrico para o modelo completo;

ω = espaço paramétrico para o modelo reduzido sob H_0 ; e

t = número de parâmetros a serem testados (dependente da hipótese).

Em que:

$$SQRR(\Omega)_{DBC} = SQRR(\Omega)_{DIC} - \sum_{h=1}^H SQBlocos_h = \sum_{h=1}^H SQRR_{h(DBC)},$$

com $N - Hp - H(r-1)$ graus de liberdade.

$$SQRR(\omega)_{DBC} = SQRR(\omega)_{DIC} - \sum_{h=1}^H SQBlocos_h,$$

com $N - [t + H(p-t)] - H(r-1)$ graus de liberdade.

Considere-se o problema geral de testar a hipótese de nulidade: $H_0 : \theta \in \omega$ versus $H_a : \theta \in \omega^c$, em que ω é um subconjunto do espaço paramétrico Ω e ω^c é o complemento de ω . A estatística do teste da razão de verossimilhança é:

$$L = \left(\frac{\hat{\sigma}_\Omega^2}{\hat{\sigma}_\omega^2} \right)^{N/2},$$

sendo $\hat{\sigma}_\Omega^2$ estimativa de máxima verossimilhança de σ^2 sem restrição no espaço paramétrico e $\hat{\sigma}_\omega^2$ a estimativa

de máxima verossimilhança de σ^2 com as restrições impostas por H_0 . Para grandes amostras de tamanho N, conforme Rao (1973), tem-se:

$$-2 \ln L = -N \ln \left(\frac{\hat{\sigma}_\Omega^2}{\hat{\sigma}_\omega^2} \right) \xrightarrow[N \rightarrow \infty]{d} \chi_v^2.$$

Designa-se de completo o modelo ajustado na ausência de restrição em Ω e de reduzido o modelo ajustado na restrição definida por H_0 .

Uma aproximação do teste da razão de verossimilhança de H_0 é provida pela seguinte estatística qui-quadrado (Rao, 1973):

$$\chi_{Calculado}^2 = -N \ln \left(\hat{\sigma}_\Omega^2 / \hat{\sigma}_\omega^2 \right) = -N \ln \left(\frac{SQRR(\Omega)_{DBC}}{SQRR(\omega)_{DBC}} \right).$$

A regra de decisão consiste em rejeitar H_0 no nível de significância α se $\chi_{Calculado}^2 \geq \chi_{\alpha}^2(v)$, em que $v = p_\Omega - p_\omega$ é o número de graus de liberdade, sendo p_Ω e p_ω os números de parâmetros estimados nos modelos completo e reduzido, respectivamente.

Um teste aproximado da razão de verossimilhança de H_0 também é dado pela estatística F:

$$F(H_0) = \frac{[SQRR(\omega)_{DBC} - SQRR(\Omega)_{DBC}] / [t(H-1)]}{SQRR(\Omega)_{DBC} / [N - Hp - H(r-1)]}$$

sob H_0
 $\sim F\{t(H-1); [N - Hp - H(r-1)]\}$

Note-se que $t(H-1) = p_\Omega - p_\omega$ é igual à diferença entre os números de graus de liberdade associados aos resíduos da regressão dos modelos reduzido e completo. O numerador da F é exatamente igual ao numerador do teste para o DIC (Regazzi & Silva, 2004), pois:

$$SQRR(\omega)_{DBC} - SQRR(\Omega)_{DBC} = SQRR(\omega)_{DIC} - \sum_{h=1}^H SQBlocos_h - [SQRR(\Omega)_{DIC} - \sum_{h=1}^H SQBlocos_h] = SQRR(\omega)_{DIC} - SQRR(\Omega)_{DIC}$$

Para ilustrar a aplicação dos testes apresentados anteriormente, utilizaram-se os dados da Tabela 2, em que $H = 2$.

As análises foram realizadas utilizando-se os procedimentos GLM e NLIN do SAS (1990). Maiores detalhes podem ser vistos em Regazzi & Silva (2004).

RESULTADOS E DISCUSSÃO

Na Tabela 3 está apresentado o resumo das análises individuais para cada grupo. Note-se que o resíduo da regressão é dado pela Falta de Ajustamento mais o resíduo (erro livre dos efeitos de blocos). Na Tabela 4 estão apresentadas as estimativas dos parâmetros do modelo sem restrição no espaço paramétrico (Ω) e com as restri-

ções especificadas por ω_1 , ω_2 e ω_3 . Nas Tabelas 5 e 6 encontram-se os resultados dos testes para as três hipóteses formuladas com aproximações dadas pelas estatísticas χ^2 e F, respectivamente.

Como o teste de $H_0^{(3)}$ foi significativo ($P < 0,01$), pode-se concluir que as duas equações diferem significativamente. Assim, a equação comum, cujas estimativas estão apresentadas na Tabela 4 (modelo ω_3), não pode ser adotada para os dois grupos. Entretanto, a hipótese de que o valor do parâmetro b é igual para os dois grupos não foi rejeitada ($P > 0,05$). Assim, o modelo escolhido é o com três parâmetros (a_1 , a_2 e b). A equação ajustada é:

$$\hat{y} = d_1 \frac{208,4x}{0,0577 + x} + d_2 \frac{166,1x}{0,0577 + x}, \text{ grupo 1: } d_1=1 \text{ e } d_2=0 \text{ e}$$

grupo 2: $d_1=0$ e $d_2=1$.

Na Tabela 7 está apresentado o resultado do teste para a falta de ajustamento do modelo escolhido.

Uma vez que a falta de ajustamento foi não significativa ($p = 0,4074$), pode-se concluir que esse modelo é adequado para descrever os dados.

É importante ressaltar que os testes apresentados são aproximados, pois em modelos de regressão não-linear as estatísticas χ^2 e F apresentadas não possuem distribuição exata. Com base em Regazzi & Silva (2004), a aproximação F deve ser preferida, uma vez que proporciona uma taxa de erro tipo I menor, independentemente do tamanho da amostra. Mesmo em modelos lineares, segundo Anderson & Mclean (1974) os erros aleatórios introduzidos pelos erros de restrição impedem que uma correta análise de regressão seja realizada (a menos que se possam combinar esses erros), porque uma pressuposição básica em toda análise de regressão é que exista somente um erro aleatório e que ele seja oriundo da própria variável y. Por essa razão, quando os tratamentos forem níveis de um fator quanti-

Tabela 2 – Valores de velocidade da reação (contagem/min²) obtidos em seis níveis de concentração de substrato de um experimento com Puromycin¹

Grupo	Concentração de Substrato em ppm (x)					
	0,02	0,06	0,11	0,22	0,56	1,10
1	76 47	97 107	123 139	159 152	191 201	207 200
2	67 51	84 86	98 115	131 124	144 158	160 156

¹ Fonte: Adaptado de Bates & Watts (1988). Grupo1: enzima tratada com Puromycin; e Grupo 2: enzima não tratada.

Tabela 3 – Resumo das análises de variância e regressão para cada grupo

Causa de Variação	G. L.	Somadas de Quadrados	
		Grupo 1	Grupo 2
“Média”	1	240550,0833	157323,0000
Blocos	1	4,0833	3,0000
Tratamentos	5	30161,4167	14816,0000
Tratamentos não-corrigido	(6)	270711,5000	172139,0000
Parâmetros da regressão	2	270214	171680
Falta de ajustamento	4	497,9000	458,9000
Resíduo=Erro	5	693,4167	402,0000
Resíduo da regressão	(9)	1191,3167	860,9000

Tabela 4 – Estimativas dos parâmetros do modelo completo (Ω) e dos modelos reduzidos (ω_1 , ω_2 e ω_3) e respectivas somas de quadrados e números de graus de liberdade do resíduo da regressão (SQRR e GLRR)

Parâmetro	Modelo			
	Ω	ω_1	ω_2	ω_3
a_1	212,7	208,6	-	-
b_1	0,0641	-	0,0452	-
a_2	161,0	166,1	-	-
b_2	0,0484	-	0,0853	-
a	-	-	190,4	186,7
b	-	0,0577	-	0,0568
SQRR _{DBC}	2052,2167	2237,9167	5298,3167	7805,9167
GLRR _{DBC}	18	19	19	20

tativo é recomendável, sempre que apropriado, planejar os experimentos com delineamento inteiramente casualizado. Entretanto, em muitos experimentos a formação de blocos torna-se necessária para controle local.

Julgou-se oportuno apresentar alguns comentários sobre o coeficiente de determinação em modelos de regressão não-linear, tendo em vista que na prática cálculos e interpretações são feitos de maneira inadequada. Em modelos de regressão linear que incluem o termo constante β_0 (intercepto), o coeficiente de determinação R^2 representa a proporção da variação explicada pelo modelo. Nesse caso, o quadrado do coeficiente de correlação entre os valores observados e preditos é exatamente o R^2 . Se o modelo é linear e o termo constante não está presente (sem intercepto), o R^2 é redefinido conforme Searle (1971), e muito cuidado deve ser tomado na sua interpretação, pois ele não é mais igual ao quadrado do coeficiente de correlação entre os valores observados e preditos. Pode ocorrer que o valor do coeficiente de determinação, na versão sem intercepto, domine em muito o valor correspondente ao caso com intercepto em modelos equivalentes. Um exemplo numérico desse fato pode ser visto em Souza (1998)

Segundo Souza (1998), no caso de modelo de regressão não-linear, a adequacidade do ajustamento pode ser medida pelo quadrado do coeficiente de correlação entre os valores observados e preditos. Essa medida pode ser calculada com a utilização da seguinte fórmula:

$$R^2 = \left(\frac{\sum_{i=1}^n y_i \hat{y}_i}{\sqrt{\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}}} \right)^2 = 1 - \frac{SQR}{SQTotal_c}$$

em que SQR é a soma de quadrados do resíduo e $SQTotal_c$ a soma de quadrados total corrigida pela média. Na realidade, a igualdade na fórmula anterior só vai funcionar se a matriz jacobiana do modelo linear aproximante tiver uma coluna de 1's ou qualquer valor constante $k \neq 0$. O que se observa, na prática, é que, em muitos trabalhos de pesquisa, no caso não-linear o cálculo do R^2 não é feito de uma única maneira. Alguns utilizam a fórmula $R^2 = 1 - \frac{SQR}{SQTotal_c}$ (Souza, 1998),

outros empregam a fórmula $R^2 = 1 - \frac{SQR}{SQTotal_{nc}}$, na qual

a $SQTotal_{nc}$ é a soma de quadrados total não corrigida pela média. Com esses cálculos, às vezes obtêm-se valores extremamente altos, por exemplo, $R^2 = 99\%$, mesmo havendo enorme discrepância entre os valores observados e preditos. Um engano comumente mantido é a crença de que o R^2 , a “proporção da variação explicada”, é usado para decidir se um modelo de regressão não-linear resulta num bom ajuste aos dados. Somente quando se

Tabela 5 – Resultados dos testes das hipóteses $H_0^{(1)}$, $H_0^{(2)}$ e $H_0^{(3)}$ pela estatística qui-quadrado

Hipótese	χ^2_{Calc}	G. L. (v)	$P(\chi^2_v > \chi^2_{Calc})$
$H_0^{(1)} : b_1 = b_2 = b$	2,079	1	0,1493
$H_0^{(2)} : a_1 = a_2 = a$	22,763	1	1,8326x10 ⁻⁶
$H_0^{(3)} : a_1 = a_2 = a$ e $b_1 = b_2 = b$	32,063	2	0,1090x10 ⁻⁶

Tabela 6 - Resultados dos testes das hipóteses $H_0^{(1)}$, $H_0^{(2)}$ e $H_0^{(3)}$ pela estatística F

Hipótese	F_{Calc}	G. L.	$P(F > F_{Calc})$
$H_0^{(1)} : b_1 = b_2 = b$	1,63	1 e 18	0,2179
$H_0^{(2)} : a_1 = a_2 = a$	28,47	1 e 18	45,2x10 ⁻⁶
$H_0^{(3)} : a_1 = a_2 = a$ e $b_1 = b_2 = b$	25,23	2 e 18	6,0051x10 ⁻⁶

Tabela 7 – Resultado do teste da falta de ajustamento do modelo com três parâmetros (a_1 , a_2 e b)

Causa de Variação	GL.	S.Q.	Q.M.	F
Falta de ajustamento	9	1142,5000	126,9444	1,16 ns
Resíduo=Erro	10	1095,4167	109,5417	
Resíduo da regressão	19	2237,9167		

ns – Não significativo ($p = 0,4074$)

tem um modelo linear com o termo constante é que o R^2 representa a variação explicada pelo modelo. O fato é que, independentemente de haver ou não um termo constante no modelo, o R^2 não tem nenhum significado óbvio no caso de modelos de regressão não-linear, e segundo Ratkowsky (1990), ele nunca precisa ser calculado. Como nos trabalhos de pesquisa há tendência de querer apresentar o R^2 , recomenda-se que o cálculo seja efetuado via $R^2 = \left(\frac{\sum \hat{y}_i}{n} \right)^2$ e não interpretá-lo como no caso de regressão linear com intercepto. Pode-se utilizar este R^2 mais como uma estatística descritiva, tendo o devido cuidado na sua interpretação. É importante verificar se não há grandes discrepâncias entre os valores observados e preditos, olhar a magnitude da variância residual para decidir se é suficientemente pequena, utilizar análise de resíduos, dentre outras técnicas de diagnóstico e, assim, optar pelo modelo mais adequado.

CONCLUSÕES

A identidade de modelos de regressão não-linear e a igualdade de qualquer subconjunto de parâmetros podem ser verificadas por meio do teste da razão de verossimilhança utilizando as aproximações F ou qui-quadrado. A aproximação F deve ser preferida.

REFERÊNCIAS

- Anderson, VL & Mclean, RA (1974) Restriction error: another dimension in teaching experimental statistics. *The American Statistician*, 28:145-152.
- Bates, DM & Watts, DG (1988) *Nonlinear regression analysis and its applications*. New York, John Wiley. 365p.
- GRAYBILL, FA (1976) *Theory and application of the linear model*. Belmont, Duxbury Press. 704p.
- Li, JCR (1964) *Statistical inference*. Ann Arbor, Edwards Brother, v.1, 658p.
- Neter, J & Kutner, MH, Nachtsheim, CJ & Wasserman, W (1996) *Applied linear statistical models*. 4th ed. New York, McGraw-Hill, 1408p.
- Pimentel Gomes, F (2000) *Curso de estatística experimental*. Piracicaba, Degaspari, 477p.
- Rao, CR (1973) *Linear statistical inference and its applications*. New York, John Wiley. 560p.
- Ratkowsky, DA (1990) *Handbook of nonlinear regression models*. New York and Basel. Marcel Dekker, Inc. 241p.
- Regazzi, AJ (1993) Teste para verificar a identidade de modelos de regressão e a igualdade de alguns parâmetros num modelo polinomial ortogonal. *Revista Ceres*, 40: 176-195.
- Regazzi, AJ (1999) Teste para verificar a identidade de modelos de regressão e a igualdade de parâmetros no caso de dados de delineamentos experimentais. *Revista Ceres*, 46: 383-409.
- Regazzi, AJ (2003) Teste para verificar a igualdade de parâmetros e a identidade de modelos de regressão não-linear. *Revista Ceres*, 50:9-26.
- Regazzi, AJ & Silva, CHO (2004) Teste para verificar a igualdade de parâmetros e a identidade de modelos de regressão não-linear. I. dados no delineamento inteiramente casualizado. *Revista de Matemática e Estatística*, 22:33-45.
- SAS Institute Inc. (1990) *SAS/STAT User's guide*. Version 6, 4th ed., v.2, Cary, NC, SAS Institute.
- Searle, SR (1971) *Linear models*. New York, John Wiley & Sons. 532p.
- Souza, GS (1998). *Introdução aos modelos de regressão linear e não-linear*. Brasília, EMBRAPA-SPI/EMBRAPA-SEA. 505p.