Análise das pressuposições e adequação dos resíduos em modelo de regressão linear para valores individuais, ponderados e não ponderados, utilizando procedimentos do SAS®

Janete Pereira Amador[1], Sidinei José Lopes[2], João Eduardo Pereira[1], Adriano Mendonça Souza[1], Marcos Toebe[3]

[1]Departamento de Estatística/CCNE - UFSM; Santa Maria, RS
[2]Departamento de Fitotecnia/CCR - UFSM; Santa Maria, RS
[3]PPGAGRO/CCR - UFSM; Santa Maria, RS
e-mail: janeteamador@hotmail.com

Resumo

Quando se quer estabelecer relações que possibilitem predizer uma ou mais variáveis em função de outras, a análise de regressão é a técnica apropriada. Existindo medidas repetidas da variável independente X, para diferentes medidas da variável dependente Y, o modelo de regressão pode ser ajustado de três maneiras diferentes: utilizando os valores individuais de X e Y (considerando todos os dados); com as médias de Y para os níveis de X (tratamento); e, ainda, utilizando as médias ponderadas de Y pelo número de repetições de cada nível de X. O objetivo deste trabalho é ajustar um modelo de regressão linear simples, através de valores individuais, com as médias ponderadas e não ponderadas dos tratamentos, a fim de testar os pressupostos para adequação do modelo, bem como, realizar a análise de variância, decompondo a soma de quadrados do erro em seus componentes, avaliando-se a falta de ajuste. Todas as técnicas foram realizadas através do suporte computacional *SAS*. Ob-

serva-se que os modelos ajustados para dados individuais e médias ponderadas apresentam os mesmos coeficientes. O teste para falta de ajuste só é possível de ser realizado com os dados individuais. A escolha da melhor estratégia adotada para analisar os dados deve ser decidida pelo pesquisador, mas sugere-se que, na disponibilidade dos dados individuais, a melhor estratégia seria estimar o modelo com estes, visto que apresentam informações mais precisas em relação à variabilidade do conjunto de dados, em relação ao uso das médias das variáveis.
Palavras-chave: ajuste de modelos de regressão, decomposição dos resíduos, teste de pressupostos.

Abstract

It is appropriate to use regression analysis establish relations that allow to predict tone or more variables in terms of others. When there are repeated measurements for independent variable X for different measurements for dependent variable Y, the regression model may be adjusted in three different ways: using individual values of X and Y (considering all data); with means of Y for levels of X (treatments) and, using weighted means of Y by the number of repetitions of each level of X (treatment). The objective of this study is to adjust a linear regression model by individual values with weighted and not weighted means of the treatments in order to test the presuppositions for the adequacy of the model and to analyze the variance decomposing the sum of squares of error in its components, thus evaluating the Lack of Fit. The adjustments of the models and its presuppositions were done in *SAS*. Thus, it was observed that the adjusted models for individual data and weighted means present the same coefficients. The test for Lack of Fit is only possible with individual data. The choice of best strategy to analyze the data should be decided by the researcher but it is suggested that, when all data of the research are accessible, the best strategy would be to estimate the model using individualized data since it presents more precise information regarding the variability of the data set which does not happen when working with means of variables.
Keywords: regression models adjustment, decomposition of residue, test of presuppositions.

Introdução

Nos diversos ramos da ciência, surge a necessidade de se estabelecer relações quantitativas entre o fenômeno observado e algumas variáveis independentes. Ou seja, ajustar um modelo matemático que seja capaz de explicar o fenômeno observado e que também seja capaz de proporcionar previsões dentro e, se possível, fora dos limites investigados. Para tanto, utiliza-se a técnica de análise de regressão.

Ao estabelecer um modelo de regressão, é necessário seguir alguns pressupostos que, de acordo com LEVINE *et al.* (2005), destacam-se os seguintes: homocedasticidade, normalidade dos resíduos, independência dos erros e linearidade. Depois de estabelecido o modelo de regressão, torna-se necessário verificar a qualidade do ajuste. Conforme COSTA *et al.* (2006), o método empregado para se avaliar quantitativamente a qualidade do ajuste de um modelo é a análise de variância (ANOVA).

Os principais objetivos da análise de variância são: verificar se há falta de ajuste no modelo (lack of fit); obter a estimativa correta para a variância do modelo de regressão($\hat{s}^2$); e estimar o grau de ajuste e significância do modelo (GAUDIO & ZONDONADE, 2001). Quando o modelo proposto é correto, a média dos quadrados dos resíduos ($\hat{s}^2$)é um estimador sem viés da verdadeira variância (σ^2). Entretanto, quando o modelo não é adequado, ($\hat{s}^2$)estará estimando algo maior do que (σ^2), pois na soma dos quadrados estarão incluídos os vieses, devido à inadequação do modelo.

Nesse sentido, GAUDIO & ZONDONADE (2001) e SOUZA (1998) argumentam que o desvio padrão do modelo é um critério de ajuste do modelo ($\hat{s}^2$). No entanto, só é possível saber se ($\hat{s}^2$) é a estimativa correta de se não houver falta de ajuste no modelo. Sendo assim, para romper este ciclo, verifica-se, em primeiro lugar, a falta de ajuste do modelo proposto, através dos resíduos da regressão. Ainda, conforme os autores acima, os resíduos de um modelo de regressão contêm toda a informação necessária à compreensão dos motivos que fazem com que ele não consiga explicar 100% da variabilidade dos dados observados de Y. Existem basicamente dois motivos para que isso ocorra, sendo estes: presença de erros aleatórios relativos à determinação dos valores de Y e a especificação imprópria do modelo (falta de ajuste).

A análise de um modelo de regressão linear tem uma relação muito

forte com a qualidade de ajuste obtida, bem como, com a confiabilidade dos testes estatísticos sobre os parâmetros do modelo (CHARNET *et al.*, 1999). Assim, a análise dos resíduos tem uma importância fundamental na verificação da qualidade dos ajustes. Basicamente, essa análise fornece evidências sobre possíveis violações nas suposições do modelo e, quando for o caso, ainda fornece indícios da falta de ajuste do modelo.

LEVINE *et al.* (2005) define o resíduo como sendo o valor observado de Y_i menos o valor previsto de $\hat{Y}_i$, isto é, $e_i = (Y_i - \hat{Y}_i)$. Existem duas situações que devem ser bem caracterizadas em relação à verificação da falta de ajuste do modelo. A primeira é quando cada valor Y_i presente no conjunto de dados foi determinado uma única vez, ou seja, quando cada valor de Y_i for o resultado de uma medida de ponto único.

Nesse caso, a verificação da falta de ajuste pode ser feita qualitativamente, através da análise da distribuição dos resíduos do modelo. Se o modelo ajustado for apropriado para os dados, não haverá padrão aparente de resíduos em relação a X_i. No entanto, se o modelo ajustado não for apropriado, existirá uma relação entre os valores de X_i e e_i (GAUDIO & ZANDONADE, 2001; SUBRAMANIAN *et al.*, 2007). A segunda situação é quando os valores de Y_i, presentes no conjunto de dados, forem determinados em réplica (duplicata, triplicata, entre outras). Nesse caso, as repetições das medidas de Y_i podem ser utilizadas para obter a estimativa da variância do modelo. Tal estimativa representa o chamado erro puro, pois, se o conjunto de valores $X_{i1}, X_{i2}, ...X_{ik}$, é o mesmo para duas ou mais observações, somente erros aleatórios podem influenciar nos valores de Y_i e gerar diferenças entre eles (DRAPER & SMITH, 1981).

Para estimar o erro puro e a falta de ajuste, deve-se fazer uma decomposição algébrica dos desvios das respostas observadas em relação à resposta média global. O desvio de uma resposta em relação à média de todas as respostas observadas $Y_i - \bar{Y}$ pode ser dividido em duas parcelas (SEARLE, 1971; DRAPER & SMITH, 1981 e CHARNET *et al.* 1999):

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \qquad (1)$$

A primeira parcela ($\hat{Y}_i - \bar{Y}$) representa o desvio da previsão feita pelo modelo para o ponto em questão, $\hat{Y}_i$ em relação à média global $\bar{Y}$.

A segunda parcela é a diferença entre o valor observado e o valor predito. Elevando-se a expressão 1 ao quadrado e, fazendo-se o somatório de todos os pontos, teremos do lado esquerdo a soma quadrática total, SQ_t,

$$\sum (Y_i - \bar{Y})^2 = \sum \left[\left(\hat{Y}_i - \bar{Y} \right) + \left(Y_i - \hat{Y}_i \right) \right]^2 \qquad (2)$$

$$\sum (Y_i - \bar{Y})^2 = \sum \left(\hat{Y}_i - \bar{Y} \right)^2 + 2\sum \left(\hat{Y}_i - \bar{Y} \right)\left(Y_i - \hat{Y}_i \right) + \sum \left(Y_i - \hat{Y}_i \right)^2 \qquad (3)$$

Do lado direto, obtêm-se as somas quadráticas da regressão e dos resíduos, pois o somatório dos termos cruzados se anula. Pode-se escrever, então:

$$\sum (Y_i - \bar{Y})^2 = \sum \left(\hat{Y}_i - \bar{Y} \right)^2 + \sum \left(Y_i - \hat{Y}_i \right)^2 \qquad (4)$$

Os resíduos, por sua vez, são decompostos em dois componentes, a falta de ajuste e o erro puro.

O teste para falta de ajuste baseia-se na suposição de que, para um conjunto de dados em que há medidas repetidas da variável independente X, para diferentes medidas da variável dependente Y, é possível particionar a soma de quadrados do resíduo em dois termos: um é o chamado Erro Puro e o outro, Lack of Fit do modelo (SEARLE, 1971). É obrigatório se trabalhar com réplicas autênticas dos tratamentos para permitir o cálculo dos termos resultantes do desdobramento dos resíduos, considerando-se que para cada valor de X_i tenham sido determinados n_i respostas obtidas em repetições autênticas. Utiliza-se um segundo índice, j, para identificar a repetição Y_{ij}. Para cada nível *i*, teremos n_i resíduos deixados pelo modelo, um para cada resposta repetida. Somando-se os quadrados de todos eles, em todas as repetições e em todos os níveis, obtemos a soma quadrática residual. Admitindo-se que hajam níveis diferentes da variável *X*, podem-se escrever as expressões: Soma quadrática dos resíduos no nível i $(SQ_r)_i$:

$$(SQ_r)_i = \sum_{j}^{ni} (Y_{ij} - \hat{Y}_i)^2 \quad (n_i = \text{número de medições no nível i}). \qquad (5)$$

Soma quadrática residual:

$$SQ_r = \sum_{i}^{m} (SQ_r)_i = \sum_{i}^{m} \sum_{j}^{ni} (Y_{ij} - \hat{Y}_i)^2 \qquad (6)$$

(*m*=número de níveis distintos da variável independente).

Cada resíduo individual pode ser decomposto na diferença de dois termos:

$$(Y_{ij} - \hat{Y}_i) = (Y_{ij} - \overline{Y}_i) - (\hat{Y}_i - \overline{Y}_i) \qquad (7)$$

em que: $\overline{Y}_i$ é a média das respostas observadas no nível *i*. Elevando-se ao quadrado a equação (7) e, somando todas as observações, teremos do lado esquerdo a soma quadrática residual, SQ_r. Do lado direto, obtêm-se as somas quadráticas das duas parcelas, pois o somatório dos termos cruzados se anula. Pode-se escrever, então:

$$\sum_{i}^{m} \sum_{j}^{ni} (Y_{ij} - \hat{Y}_i)^2 = \sum_{i}^{m} \sum_{j}^{ni} (Y_{ij} - \overline{Y}_i)^2 + \sum_{i}^{m} \sum_{j}^{ni} (\hat{Y}_i - \overline{Y}_i)^2 \qquad (8)$$

O primeiro somatório do lado direito reflete a dispersão do sinal (resposta) Y_{ij}, em torno de suas médias, $\overline{Y}_i$ oferecendo uma medida do erro aleatório e, sendo, portanto, denominado de soma quadrática devida ao erro puro, SQ_{ep}.

O segundo somatório decorre do modelo e sua magnitude depende do afastamento da estimativa $\hat{Y}_i$ da respectiva média $\overline{Y}_i$. Esse termo fornece uma medida da falta de ajuste do modelo às respostas observadas, sendo chamado, por isso, de soma quadrática, devido à falta de ajuste, SQ_{faj}. Assim, com a decomposição da soma quadrática, obtém-se a tabela de análise de variância. A média quadrática é obtida pela divisão da soma quadrática pelo respectivo número de graus de liberdade.

Quando existirem medidas repetidas da variável independente X para diferentes medidas da variável dependente Y, o modelo de regressão pode ser ajustado de três maneiras diferentes: utilizando os valores individuais de X e Y (considera todos os dados); com as médias de Y para os níveis de X (tratamentos) e, ainda, utilizando as médias ponderadas de Y pelo número de repetições de cada nível de X (tratamento).

Nesse sentido, conforme Draper & Smith (1981), quando houver o mesmo número de repetições para os níveis de X, os três ajustes fornecem igual estimativa para os parâmetros do modelo. Quando o número de repetições é diferente, a regressão realizada com todos os dados e a regressão com as médias ponderadas continuam fornecendo estimativas iguais dos parâmetros, no entanto, a regressão com as médias não ponderadas apresenta estimativa diferente.

Em vista do exposto, este trabalho tem como objetivo ajustar um modelo de regressão linear simples, através de três estratégias: valores individuais, com as médias ponderadas e não ponderadas dos tratamentos, testar os pressupostos para adequação do modelo, bem como realizar a análise de variância decompondo a soma de quadrados do erro em seus componentes. Além disso, apresentar todos os procedimentos para realizar as análises, através do sistema computacional *SAS v. 8.0 (Statististical Analysis System*).

Material e métodos

Os dados utilizados para o ajuste do modelo foram oriundos de um experimento realizado em área experimental do Departamento de Fitotecnia da Universidade Federal de Santa Maria - RS. Neste experimento, foi estudado o efeito de três densidades de plantas (tratamentos, X) sobre a produção de fitomassa seca da parte aérea (MS, Y) de mamona (kg). As densidades foram de 1,0, 1,2 e 1,4 m entre plantas, mantendo-se constante o espaçamento entre linhas, de 1,0 m. As programações, realizadas no *SAS v. 8.0*, para aplicação da técnica proposta, encontram-se no Apêndice 1.

Ajustaram-se os modelos de regressão, utilizando-se três estratégias: os valores individuais de X e Y, as médias ponderadas de Y pelo número de repetições dos níveis X e as médias não ponderadas. Depois de estabelecido o modelo de regressão, verificou-se a qualidade do ajuste.

O método empregado para se avaliar numericamente a qualidade do ajuste de um modelo foi a análise de variância (ANOVA) (Costa *et al.*, 2006). Para realizar a validação dos modelos, utilizaram-se procedimentos do SAS. As hipóteses testadas, em termos gerais, para a validação dos modelos foram: H_0: o modelo segue determinado pressuposto,

contra H_1: o modelo não segue determinado pressuposto. Uma forma de testar a normalidade é através de testes de aderência, como o teste de Shapiro-Wilk; pelo teste de White, verificou-se o pressuposto de homogeneidade de variâncias; a independência dos resíduos, através da estatística Durbin Watson; e a linearidade foi verificada através do teste de F da análise de variância, todos fornecidos pelo programa computacional *SAS*.

Resultados

Observa-se na, Tabela 1, que na análise dos dados individuais e das médias ponderadas, as estimativas dos parâmetros do modelo são as mesmas. Porém, na análise das médias não ponderadas, há estimativas diferentes dos outros dois critérios, pelo fato de apresentar números desiguais de repetições para os tratamentos (níveis de X). Quanto aos coeficientes de determinação, verifica-se que estes são sensivelmente melhores para os modelos obtidos através dos valores médios. Uma regressão com valores médios sugere maior capacidade preditiva do que uma regressão sobre os dados individuais, uma vez que os valores médios apresentam menor variabilidade que os valores individuais. Outro fato que indica um melhor ajuste para os modelos com valores médios é o desvio padrão $(\hat{s})$, que aparece com valores menores do que com os dados individuais.

O desvio padrão do modelo é uma medida de variabilidade da distribuição condicional de Y para valores fixos de X. Utilizam-se todos os resíduos da reta ajustada de regressão para calcular o desvio padrão do modelo, pois se supõe que todas as distribuições condicionais tenham a mesma variância. Dessa forma, o desvio padrão do modelo serve como referência para a escolha do melhor modelo, isto é, aquele que tem o menor desvio padrão (Hill *et al.*, 1999).

As programações para os ajustes dos modelos podem ser verificadas no Apêndice 1.

A falta de ajuste testada, com 1 e 18 graus de liberdade, não foi significativa, sendo o nível de significância alfa de 0,772 (Tabela 2). Isso mostra que o modelo é adequado para descrever o comportamento da produção de fitomassa seca de mamona, em relação às diferentes densidades de

plantas. Além disso, sendo o modelo adequado, o QM= 1099,32 pode ser usado como estimador, não-tendencioso, da variância do modelo assim como do desvio padrão. O pressuposto de linearidade, verificado através do teste de F na análise de variância, indicou, para os modelos ajustados, conforme os três critérios, a aceitação da hipótese de linearidade. Ou seja, a relação entre X e Y (diferentes densidades de plantas e produção de fitomassa seca) pode ser descrita através de um modelo linear.

Quando o pressuposto da linearidade é violado, o pesquisador deve verificar a necessidade de utilizar mais variáveis para descrever o fenômeno em questão, ou estar ciente de que o modelo de regressão linear não é o melhor modelo explicativo para o estudo das variáveis envolvidas.

Os testes aplicados para validação dos modelos encontram-se na Tabela 3 e, no Apêndice 1, as programações para realização destes.

Verifica-se que, para os três critérios adotados, não houve rejeição da hipótese de normalidade dos resíduos em nível de 1% de erro. Usando o "*proc model*" e através do comando "fit", é possível realizar os três testes para validação do modelo. O *proc model* foi aplicado para os dados individuais e médias não ponderadas. No entanto, para utilizá-lo, é necessário definir os parâmetros do modelo através do subcomando "*parms*", sendo novamente estimados os parâmetros e realizada a ANOVA. Para as médias ponderadas, foi utilizado o "*proc univariate*". Primeiramente, cria-se a variável resíduo com o subcomando "*var*", sendo testada a variável resíduo pela opção "*normal*". Mais detalhes dessa opção encontram-se no Apêndice 1.

Tabela 1. Modelos de regressão ajustados para a relação densidade de plantas (X) de mamona e produção de massa seca da parte aérea (Y), conforme os valores individuais de X e Y, as médias ponderadas de Y pelo número de repetições dos níveis de X e as médias não ponderadas.

Modelo	Parâmetros				R2	$R^2_{ajustado}$	$\hat{s}$
	b0	Pr>t	b1	Pr>t			
Dados individuais	996,84	<0,000	570,1	<0,0001	0,8958	0,8953	33,15
Médias ponderadas	996,84	0,0106	570,1	0,015	0,9994	0,9989	9,99
Médias não ponderadas	998,21	0,0103	568,98	0,015	0,9994	0,9989	3,79

Tabela 2. Análise de variância para os modelos de regressão ajustados para a relação densidade de plantas (X) de mamona e produção de massa seca da parte aérea (Y), conforme os valores individuais de X e Y, as médias ponderadas de Y pelo número de repetições dos níveis de X e as médias não ponderadas.

Modelos de Regressão					
Dados Individuais					
Fontes de Variação	*GL*	*SQ*	*QM*	*F*	*Pr>F*
Regressão	1	179535	179535	163,31	0,0001
Resíduos	19	20887,22	1099,32		
Falta de ajuste	1	99,85	99,85	0,086461	0,7720
Erro puro	18	20787,37	1154,854		
Total	20	200422			
Médias Ponderadas					
Regressão	1	179535	179535	1797,88	0,0150
Resíduos	1	99,85	99,85		
Total	2	179634			
Médias não Ponderadas					
Regressão	1	25899	25899	1803,00	0,0150
Resíduos	1	14,365	14,365		
Total	2	25914			

GL=graus de liberdade, SQ= Soma de Quadrados, QM= Quadrado Médio, F= F de Snedecor, Pr>F= nível de significância.

Tabela 3 . Testes para validação dos modelos de regressão ajustados para a relação densidade de plantas (X) de mamona e produção de massa seca da parte aérea (Y), conforme os valores individuais de X e Y, as médias ponderadas de Y pelo número de repetições dos níveis de X e as médias não ponderadas.

Critérios	Testes						
	Shapiro-Wilk		White		Durbin Watson		
	Valor	P	Valor	P	Valor	P<WD	P>DW
Dados individuais	0,82	0,11	0,95	0,62	1,71	0,1788	0,8212
Médias ponderadas	0,81	0,15	2,00	0,57	2,99	*	*
Médias não ponderadas	0,75	0,10	3,00	0,22	3,00	*	*

**** não apresenta valor de P= probabilidade.***

Com o teste de White, testou-se a igualdade de variância dos erros aleatórios. Não houve a rejeição desse pressuposto para nenhum dos modelos. Para testá-lo, usa-se a opção "*fit*", seguido do nome do teste, no caso, 'White'. Para o modelo com as médias ponderadas, utilizou-se o "*proc reg*". Após, definiu-se o modelo com a opção "*model.../SPEC*". O comando SPEC gera a estatística do valor teórico da distribuição Qui-quadrado para o teste de White. Essa opção só é válida quando for definido o modelo no "*model*".

Utilizando a estatística de Durbin-Watson, observou-se que os resíduos para os dados individualizados não são autocorrelacionados. Com a opção "*DW*", acrescido de "*/DWPROB*", é gerada a estatística do teste com as probabilidades para avaliar a existência de correlações negativas e positivas dos resíduos. Quando o valor de $p <$ DW, indica que os resíduos são correlacionados positivamente e, quando $p >$DW, ocorre uma correlação negativa.

O SAS mostrou-se uma ferramenta estatística extremamente versátil para a realização dos procedimentos de análise, os quais podem ser realizados de duas formas diferentes. A primeira, através do "*proc model*", usando as opções "*parms*" e "*fit*". A segunda forma é por intermédio do "*proc reg*". Neste, os testes para validação são chamados separadamente, ou seja, para cada teste, é preciso usar um *proc reg*. Quanto à escolha da forma com que se queira montar a programação, fica a critério do pesquisador, já que as duas se mostraram eficazes.

Conforme LEVINE *et al.* (2005), quando os resíduos sucessivos são positivamente autocorrelacionados, o valor da estatística D irá se aproximar de zero. Se os resíduos não forem autocorrelacionados, o valor de D estará próximo de dois. Se existir autocorrelação negativa, o que é raro, D será maior que dois e poderia se aproximar do seu valor máximo, sendo igual a quatro. O valor obtido no teste D é comparado na tabela Durbin-Watson, o primeiro valor "d_l" representa o valor crítico inferior, quando não existe autocorrelação. Se D estiver abaixo de "d_l", existem evidências de autocorrelação entre os resíduos. O segundo valor "d_u" representa o valor crítico superior de D, acima do qual conclui-se que não há evidência de autocorrelação entre os resíduos. No entanto, se D estiver entre "d_l" e "d_u", não se pode tirar uma conclusão definitiva. Baseado nessas informações e de acordo com os resultados do teste Durbin-Watson, para os modelos com as médias ponderadas e não-ponderadas,

não se tem um estudo conclusivo sobre a evidência de autocorrelação dos resíduos para esses dois modelos. Quando se tem acesso a todos os dados da pesquisa, a melhor estratégia seria estimar o modelo através dos dados individualizados, já que este apresenta informações mais precisas em relação à variabilidade do conjunto original, o que não acontece quando se trabalha com as médias. A escolha final do tipo de modelagem depende entre outros fatores da finalidade a que se destina o estudo e do tipo de resposta que o pesquisador busca.

Conclusão

Por meio das três estratégias utilizadas para modelagem em análise de regressão, usando dados individuais, médias ponderadas e médias não ponderadas, as duas primeiras apresentaram os mesmos valores para os parâmetros do modelo ajustado. Na análise de variância para o modelo com os dados individuais, foi possível verificar a decomposição da soma de quadrados do resíduo em seus componentes, erro puro e falta de ajuste. Verificou-se que o teste só é possível se existirem repetições nos níveis ou tratamentos. Pelas análises dos pressupostos, o modelo ajustado com os dados individuais apresentou resultados mais conclusivos em relação aos outros dois modelos. O SAS apresenta-se como ferramenta versátil para a realização dos procedimentos de análises dos modelos de regressão.

Referências

CHARNET, R. *et al.* **Análise de modelos de regressão linear com aplicações**. Campinas, SP: Unicamp, 1999.

COSTA, T. M. *et al.* **Utilização de planilha eletrônica para calibração instrumental, análise da variância e testes de significância de um método espectrométrico.** Revista Analytica, n. 21, p. 46-51, 2006.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. JohnWiley&Sons:New York, 1981.

GAUDIO, A. C. ZANDONADE, E. **Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica**. Quim. Nova, v. 24, n.5, 658-671, 2001.

HILL, C. *et al*. **Econometria**. São Paulo: Saraiva, 1999.

LEVINE, D. M. *et al*. **Statística - teoria e aplicações usando o microsoft Excel em português**, Rio de Janeiro: LTC, 2005. 3 ed.

SAS Institute Inc. **SAS Software: Reference, Version 8**, Cary, NC: SAS Institute Inc., 1999.

SEARLE, S. R. **Linear models**. New York: John Wiley, 1971.

SOUZA, G. S. **Introdução aos modelos de regressão linear e não-linear**. Brasília: EMBRAPA – SPI, 1998.

SUBRAMANIAN, A.; COUTINHO, A. S.; da SILVA, L. B. **Aplicação de método e técnica multivariados para previsão de variáveis termoambientais e perceptivas.** Produção, v. 17, n. 1, p. 052-070, 2007.

Submetido em: 04/04/2010
Aceito em: 08/07/2011

APÊNDICE 1

Programação do SAS para execução das análises

```
dm 'output; clear; log; clear;';
options formdlim='*' pageno=001 ls=80;
DATA reg;
INPUT X Y;
CARDS;
1   1572.08
1   1650.25
1.2     1638.34
1.2     1668.75
.....................................
1.4     1775.38
1.4     1875.38;
;
PROC REG DATA=reg;
MODEL y= x;  /*SQr*/
TITLE 'Análise de regressão com as
observações individulisadas';
run;
PROC ANOVA
DATA=reg;
CLASS X; MODEL Y = X ; /* SQep */
TITLE 'Obtendo a soma de quadrados
relacionada ao erro puro';
run;
PROC MEANS N MEAN NWAY;
CLASS X; VAR Y; OUTPUT OUT=MEANS N=NUM
MEAN=MY;
run;
PROC REG DATA=MEANS; WEIGHT NUM; /*
Residual = falta de ajuste */
TITLE'Regressão com as médias ponderadas
dos tratamentos';
 MODEL MY = X;
 run;
PROC REG DATA=MEANS;
```

```
        TITLE3 'Regressão com as médias não
ponderadas';
        MODEL MY =X;
        proc model data=reg;/*testando os
pressupostos para a análise de regressão*/
        title 'testando os pressupostos para os
dados individualisados';
        parms A B;
        y=A + B*X;
        fit y/White normal DW DWPROB;/*DWPROB da a
probabilidade de Durbin Watson */
        run;
        proc model data=MEANS;/*testando os
pressupostos para a análise de regressão*/
        title 'testando os pressupostos para o
modelo com as médias não ponderadas';
        parms A B;
        MY=A + B*X;
        fit MY/White normal DW DWPROB;/*DWPROB da
a probabilidade de Durbin Watson */
        run;
        proc reg data=MEANS;WEIGHT NUM;/*outra
forma de testar os pressupostos para a análise de
regressão*/
        title 'testando os pressupostos de
homocedasticidade com modelo das médias
ponderadas';
        model MY=x/SPEC;
        run;
        proc reg data=MEANS;WEIGHT NUM;/*outra
forma de testar os pressupostos para a análise de
regressão*/
        title 'testando os pressupostos de
independência dos resíduos com modelo das médias
ponderadas';  model MY=x/DW DWPROB;
        run;
        proc reg; WEIGHT NUM;/*criando a variável
resíduo, que será guardada no arquivo virtual B
com o nome de resíduo */
        model MY=x;  /*para ser testado a sua
normalidade*/
        output out=B R= resíduo;
```

```
        title 'testando os pressupostos de
normalidade dos resíduos com modelo das médias
ponderadas';
        proc univariate data=B normal; var
resíduo;
        run;
        quit;
```