MODELOS DE REGRESSÃO LINEAR MÚLTIPLA UTILIZANDO OS SOFTWARES R E STATISTICA: UMA APLICAÇÃO A DADOS DE CONSERVAÇÃO DE FRUTAS

Cecília P. Sassi, Felipe G. Perez, Letícia Myazato, Xiao Ye, Paulo H. Ferreira-Silva e Francisco Louzada

ICMC – USP – CP668 – CEP 13.566-590, São Carlos, SP, Brasil

louzada@icmc.usp.br

Resumo

Neste artigo são apresentados ajustes de modelos de regressão linear múltipla usando os *softwares* R e STATISTICA. Para isso, utilizou-se um conjunto de dados proveniente de uma pesquisa sobre conservação de frutas. Observou-se que as variáveis *tempo de contato*, *temperatura do processo* e *concentração da solução osmótica* têm impacto de sinal positivo sobre a variável resposta, a perda de peso (em %) da fruta em questão (abacaxi).

Palavras-chave: planejamento de experimentos, regressão linear múltipla, tabela ANOVA, *software* R, STATISTICA, conservação de frutas, análise de resíduos.

1. Introdução

Este artigo tem como foco a aplicação da regressão linear múltipla em dados relacionados ao planejamento de experimentos (dados de conservação de frutas) e a utilização de dois *softwares*, o R e o STATISTICA, para as devidas análises estatísticas.

O objetivo aqui é, então, ajustar modelos de regressão linear múltipla, mostrando em paralelo a utilização de dois *softwares* e explicando passo-a-passo cada função e comandos empregados.

O relatório é organizado da seguinte maneira. Na Seção 2 é apresentado um breve histórico dos dois *softwares* utilizados, o R e o STATISTICA. Na Seção 3 é descrita a metodologia empregada na análise (regressão linear múltipla em planejamento de

experimentos). Na Seção 4 são apresentados os resultados obtidos quando da aplicação da metodologia em questão a um conjunto de dados reais (dados de conservação de frutas). Comentários finais e conclusões, na Seção 5, finalizam o relatório.

2. Histórico

Nesta seção é apresentado um breve histórico dos dois *softwares* utilizados neste relatório, nominalmente, o R (seção 2.1) e o STATISTICA (seção 2.2).

2.1. O *software* livre R

O *software* R foi desenvolvido por Robert Gentleman e Ross Ihaka, ambos do Departamento de Estatística da Universidade de Auckland, da Nova Zelândia, e o nome R dado ao programa foi originado das iniciais dos autores, conhecidos por "R & R".

O objetivo inicial de "R & R", em 1991, era produzir um *software* para as suas aulas de laboratório baseado na linguagem S, utilizada pelo *software* comercial S-Plus, criado por Jonh M. Chambers da AT&T, que atualmente ajuda muito no desenvolvimento e aperfeiçoamento das ferramentas do programa.

Em 1993, algumas cópias foram disponibilizadas no StatLib, um sistema de distribuição de *softwares* estatísticos. Incentivado por um dos primeiros usuários deste programa, Martin Mächler do ETH Zürich (Instituto Federal de Tecnologia de Zurique, da Suíça), "R & R" lançaram em 1995 o código fonte do R, um *software* que podia ser utilizado pela Internet.

Desde 1997, o *software* R é atualizado e melhorado constantemente por um grupo de profissionais que de certa forma dominam a linguagem do programa. Hoje em dia, o R tem inúmeros usuários, devido ao fato de ser gratuito, porém exigindo dos usuários a lógica de programação.

Através do site http://cran.r-project.org, é possível fazer o *download* do programa, sendo compatível com quase todos os sistemas operacionais. Para usá-lo é necessário que o indivíduo conheça e digite comandos, além de poder criar suas próprias funções, caracterizando-se, assim, como um programa flexível, com inúmeros recursos. O *software* possui um conjunto de pacotes com atualizações que acrescentam potencialidades à versão base do R e com uma grande aplicação em assuntos relacionados à estatística, como a manipulação, avaliação e interpretação de dados.

Além disso, o *site* citado acima apresenta várias informações sobre como usar o R e uma central de correspondências, na qual profissionais de várias áreas buscam permanentemente mudanças no programa no intuito de facilitar ao máximo a sua utilização pelo usuário.

2.2. O *software* STATISTICA

Em 1984, uma parceria de um grupo de professores universitários e cientistas criou a empresa StatSoft com o objetivo de desenvolver procedimentos estatísticos que não estavam disponíveis no mercado ou disponíveis apenas no comando-*driven*.

Seus primeiros produtos foram o PsychoStat-2 e o PsychoStat-3, incluído menu-*driven* bibliotecas de procedimentos estatísticos flexível, integrado, com gerenciamento de dados. Os programas estavam disponíveis em versões para todas as plataformas de microcomputador e rapidamente ganharam popularidade.

Em 1985, a StatSoft começou a crescer rapidamente, lançando seu primeiro produto no mercado geral, o Suplemento de Estatística para o Lotus 1-2-3. No mesmo ano, a empresa lançou a StatFast, que foi o primeiro pacote de estatísticas divulgadas para o recém-introduzido Apple Macintosh. Em 1986, após o sucesso do Suplemento Estatístico de Lotus 1-2-3 e dos programas StatFast, duas novas unidades (o grupo de gráficos analíticos e de otimização numérica do grupo) foram adicionadas à Pesquisa StatSoft e Departamento de Desenvolvimento, criando também o CSS (sistema de estatísticas completo), com diversos módulos na área da estatística.

Entre 1988 e 1990, a organização lançou uma versão do CSS com uma integração única entre resultados numéricos e gráficos, em que toda a saída numérica poderia ser interativamente visualizada em uma variedade de maneiras. Foi investido também na parte gráfica do sistema e um resultado importante dessa decisão foi a nova tecnologia gráfica introduzida em 1991 na linha de *software* STATISTICA.

Uma das muitas características únicas do CSS é a abrangência e a qualidade dos procedimentos estatísticos, que foram desenvolvidos por especialistas nos respectivos campos de investigação ou prática. Em 1988, a empresa também lançou um programa mais poderoso para o Macintosh, o MacSS (Macintosh Sistema Estatístico). Esses anos marcaram o aumento constante na popularidade e reconhecimento da linha de *software* CSS.

A versão DOS do primeiro STATISTICA (a nova linha gráfica do *software* STATISTICA) foi lançada em Março de 1991. O *software* STATISTICA/DOS ofereceu a

maior seleção de gráficos em um único sistema disponível no mercado e uma série de vantagens tecnológicas. Algumas de suas características únicas e soluções de *interface* de usuário se tornaram padrões estatísticos e analíticos para *software* gráfico.

Em 1993, surge uma versão para Windows do STATISTICA. Entre 1994 e 1998, a StatSoft lançou o *software* STATISTICA 4.0, 4.5, 5.0, 5.1, 97 Edition e 98 Edition, e eles continuaram a definir o novo desempenho, capacidade e padrões de abrangência para as estatísticas. Na visão dos usuários, esses lançamentos só melhoravam a qualidade do *software*.

Entre 2008 e 2009, o STATISTICA 9 atualizou toda a linha de produtos StatSoft para computação de 64 bits.

Atualmente, a StatSoft é a maior empresa do mundo que oferece sistemas de controle de qualidade para empresas, apoiada por um serviço completo de escritórios em 20 locais no exterior, em todos os continentes. É hoje uma das maiores fornecedoras mundiais de *softwares* em todo o mundo analítico, e o seu principal produto, o *software* STATISTICA, disponível em vários idiomas, é usado em muitas universidades, institutos de pesquisa, empresas e unidades em mais de 60 países.

3. Metodologia

Nesta seção são apresentados conceitos sobre a regressão linear múltipla, que se referem a uma situação em que a reta ajustada não descreve bem o conjunto de dados e, com isso, podem ser levadas em consideração outras variáveis independentes que possivelmente influenciam no valor de *Y*, a variável dependente. Ou seja, a regressão múltipla pode ser usada no intuito de melhorar o modelo desenvolvido para explicar o comportamento das variáveis do banco de dados que estão sendo estudadas.

Em regressão múltipla, a variável determinada é aquela que tenha correlação significativa com a variável a ser prevista. A variável está no centro das análises e deve ser identificado o seu impacto coletivo, assim como a contribuição de cada variável separada para o efeito geral da variável preditora.

A regressão linear múltipla é uma técnica multivariada cuja finalidade principal é obter uma relação matemática entre uma das variáveis estudadas (variável dependente ou resposta) e o restante das variáveis que descrevem o sistema (variáveis independentes ou explicativas), e reduzir um grande número de variáveis para poucas dimensões com o mínimo de perda de informação, permitindo a detecção dos principais padrões de similaridade, associação e

correlação entre as variáveis. Sua principal aplicação, após encontrar a relação matemática, é produzir valores para a variável dependente quando se têm as variáveis independentes (cálculo dos valores preditos). Ou seja, ela pode ser usada na predição de resultados, por meio da regra estatística dos mínimos quadrados.

A inclusão de novas variáveis na equação de ajuste pode ser feita para aumentar o grau de correlação entre os dados teóricos e reais. Tal modelo apresenta a seguinte equação:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + e, \tag{1}$$

sendo X_p a *p*-ésima variável observada, β_p o coeficiente associado à *p*-ésima variável e $e = Y - \hat{Y} = Y - \beta_0 - \beta_1 X_1 - ... - \beta_p X_p$ o erro que apresenta distribuição normal com média zero e variância σ^2.

O resultado do modelo (1) é um único valor que representa uma combinação do conjunto inteiro de variáveis que melhor atinge o objetivo da análise multivariada específica. A solução de ajuste de reta pode ser generalizada a um conjunto de pontos $(x_1, y_1),...,(x_n, y_n)$, pelo método de mínimos quadrados (MMQ), para que modelos de regressão múltipla possam ser utilizados. Em notação matricial, tem-se que:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \text{M} \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & x_{2p} \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & x_{np} \end{bmatrix} \text{ e } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \text{M} \\ \beta_p \end{bmatrix}, \tag{2}$$

e, assim, pode-se escrever:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}, \tag{3}$$

sendo $\boldsymbol{y}$ um vetor *n* x 1, *X* uma matriz *n* x (*p*+1), $\boldsymbol{\beta}$ um vetor *p* x 1 e $\boldsymbol{e}$ um vetor *n* x 1.

Para obter a solução de ajuste da função linear em $\boldsymbol{\beta}$ a um conjunto de pontos $(y_1, x_{11}, x_{12}, x_{13}, ..., x_{1p}),...,(y_n, x_{n1}, x_{n2}, ..., x_{np})$, pelo MMQ, deve-se minimizar a expressão $\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip})]^2 = \left(\boldsymbol{y} - X\boldsymbol{\beta}\right)' \left(\boldsymbol{y} - X\boldsymbol{\beta}\right)$. Derivando simultaneamente em termos de $\boldsymbol{\beta}$, tem-se:

$$\frac{\partial \boldsymbol{y}'\boldsymbol{y}}{\partial \boldsymbol{\beta}} - 2\frac{\partial \boldsymbol{y}' X\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} + \frac{\partial \boldsymbol{\beta}' X' X\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 0 - 2(\boldsymbol{y}' X)' + 2X' X\boldsymbol{\beta}, \tag{4}$$

no qual o vetor de dimensão *p* + 1 resulta na expressão $\left(X'X\right)^{-1} X'\boldsymbol{y} = \hat{\boldsymbol{\beta}}$, em que $\hat{\boldsymbol{\beta}}$ é o estimador não-viciado do modelo, desde que $\left(X'X\right)^{-1}$ exista.

Portanto, o modelo de regressão ajustado e o vetor de resíduos são dados, respectivamente, por:

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} \quad \text{e} \quad \boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}. \tag{5}$$

O estimador de σ^2 é dado matricialmente por:

$$\text{SQMRes} = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}' X'\boldsymbol{y}}{n - p - 1} \tag{6}$$

e a matriz de covariância de $\hat{\boldsymbol{\beta}}$ é dada pela fórmula:

$$Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(X'X\right)^{-1}. \tag{7}$$

Um aspecto importante para a validação de um ajuste de regressão linear múltipla é a análise de resíduos, que mostra a significância do modelo e avalia as contribuições das variáveis regressoras.

O resíduo, definido em (5), tem esperança e variância dadas, respectivamente, por $E[\boldsymbol{e}] = \boldsymbol{0}$ e $Cov[\boldsymbol{e}] = \sigma^2 \left[I - X\left(X'X\right)^{-1} X'\right]$.

A matriz $X\left(X'X\right)^{-1} X'$ é chamada "matriz *H*" (matriz *hat*) e os valores da diagonal principal são denominados "valores *h*", em que:

$$h = h_{ii}, \text{ com } 0 < h_{ii} < 1 \text{ e } h_{ii} = \sum_{j=1}^{n} h_{ij}^2, \; i = 1,...,n. \tag{8}$$

Assim, pode-se padronizar os resíduos da seguinte forma:

$$Z_i = \frac{e_i}{\sqrt{\sigma^2 \left(1 - h_{ii}\right)}}, \; i = 1,...,n. \tag{9}$$

Quando se trabalha com análise de regressão, deve-se realizar a análise de variância com o objetivo de comparar os modelos e avaliar a significância da regressão. Considerando o modelo de regressão linear múltipla, definido em (3), pode-se construir a tabela ANOVA (tabela de análise de variância, corrigida pela média), dada por:

Tabela 1 – Tabela ANOVA (corrigida pela média).

Fonte de variação	Soma de Quadrados	Graus de liberdade	Soma de Quadrados Médios	$F_{calculado}$
Regressão	SQReg	*p*	SQMReg	SQMRes/SQMRes
Erro	SQRes	*n*-*p*-1	SQMRes	
Total	SQT	*n*-1		

em que:

$$\text{SQReg} = \boldsymbol{y}'H\boldsymbol{y} - n\bar{y}^2$$

$$\text{SQRes} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'X'\boldsymbol{y}$$

$$\text{SQT} = \boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2$$

$$\text{SQMReg} = \text{SQRes}/p \tag{10}$$

$$\text{SQMRes} = \text{SQRes}/(n\text{-}p\text{-}1)$$

Na Tabela 1, as hipóteses em teste são:

$$H_0 : \beta_0 = \beta_1 = ... = \beta_p = 0 \ \textit{versus} \ H_1 : \text{Ao menos um } \beta_i \neq 0. \tag{11}$$

Assim, se a hipótese H_0 for a verdadeira, o modelo não está bem ajustado, pois os coeficientes são estatisticamente iguais a zero.

Pode-se também calcular o coeficiente de determinação e o coeficiente de determinação ajustado, dados respectivamente por:

$$R^2 = \text{SQReg/SQT} \ \text{ e } \ R^2_{ajustado} = 1\text{-}\frac{\text{SQRes}/(n\text{-}p\text{-}1)}{\text{SQT}/(n\text{-}1)}. \tag{12}$$

Com esta breve explicação sobre a análise de regressão múltipla, pode-se agora inseri-la no contexto de planejamento de experimentos.

Considere, por exemplo, um banco de dados com n observações, dois fatores (temperatura e pressão) e uma variável resposta (viscosidade). Neste caso, a matriz X e o vetor $\boldsymbol{y}$ são dados, respectivamente, por:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \text{M} & \text{M} & \text{M} \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \ \text{ e } \ \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \text{M} \\ y_n \end{bmatrix}. \tag{13}$$

Um modelo de regressão linear múltipla para tal experimento seria $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, com os β's representando os coeficientes do modelo, os X's as variáveis do modelo (fatores do planejamento) e e o erro.

Outro exemplo consiste em um planejamento fatorial 2^3, ou seja, três fatores (temperatura, pressão e concentração), uma variável resposta (rendimento do processo) e as variáveis codificadas, com dois níveis cada e, finalmente, com quatro pontos centrais. Em termos matriciais, tem-se:

$$X = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{e} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix}. \tag{14}$$

De maneira análoga à anterior, tem-se o modelo de regressão $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$.

Com a matriz X e o vetor $\boldsymbol{y}$ devidamente definidos, os próximos passos, para cada um dos exemplos apresentados, seriam: a obtenção das estimativas dos β's, a construção da tabela ANOVA e, por fim, a análise de resíduos. Todos esses passos são apresentados em detalhes na próxima seção.

4. Aplicação e resultados

Nesta seção é feita uma análise de regressão aplicada em planejamento de experimentos, comparando as saídas dos *softwares* R e STATISTICA, em cada passo da análise.

Os dados utilizados são de uma pesquisa referente à conservação de frutas (retirados do livro de Barros Neto; Scarminio e Bruns, 2007), em que o objetivo é o desenvolvimento de produtos com longo prazo de validade, isto é, cujas propriedades sensórias e nutritivas se pareçam ao máximo com as da fruta *in natura*. Diante disso, um trabalho foi feito usando planejamento de experimentos (no caso, um planejamento composto central foi adotado), para estudar como a desidratação de pedaços de abacaxi dependia de três fatores: o tempo de contato (variável x_1), a temperatura do processo (variável x_2) e a concentração da solução osmótica (variável x_3), sendo que a perda de peso relativa (variável Y) ao final de cada ensaio foi tomada como medida do nível de desidratação. Assim, tem-se a seguinte matriz X (com os valores codificados dos três fatores) e o vetor $\boldsymbol{y}$ para a obtenção das estimativas dos parâmetros:

$$X = \begin{bmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \\ -1.682 & 0 & 0 \\ 1.682 & 0 & 0 \\ 0 & -1.682 & 0 \\ 0 & 1.682 & 0 \\ 0 & 0 & -1.682 \\ 0 & 0 & 1.682 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{e} \quad \boldsymbol{y} = \begin{bmatrix} 47.34 \\ 53.00 \\ 53.64 \\ 54.28 \\ 48.85 \\ 53.73 \\ 55.19 \\ 58.31 \\ 51.90 \\ 57.34 \\ 47.62 \\ 57.35 \\ 50.73 \\ 57.68 \\ 56.24 \\ 55.74 \\ 57.23 \\ 56.85 \\ 55.42 \end{bmatrix}. \tag{15}$$

Deste modo, procede-se com as estimativas dos parâmetros do modelo de regressão, a análise de variância e a análise de resíduos. No *software* R, foram usados os comandos apresentados no Quadro 1 a seguir.

Quadro 1 – Comandos em *R* para ajuste do modelo de regressão linear múltipla aos dados de frutas.

```
> x1 <- c(-1,1,-1,1,-1,1,-1,1,-1.682,1.682,0,0,0,0,0,0,0,0,0)
> x2 <- c(-1,-1,1,1,-1,-1,1,1,0,0,-1.682,1.682,0,0,0,0,0,0,0)
> x3 <- c(-1,-1,-1,-1,1,1,1,1,0,0,0,0,-1.682,1.682,0,0,0,0,0)
> y <- c (47.37, 53, 53.64, 54.28 ,48.85, 53.73, 55.19, 58.31, 51.9, 57.34, 47.62, 57.35, 50.73,

57.68, 56.24, 55.74, 57.23, 56.85, 55.42)


> library(rsm)                  #instala o pacote rsm
> summary(rsm(y~SO(x1,x2,x3)))      #modelo ajustado
```

A partir dos comandos acima (Quadro 1), foram obtidas as saídas (resultados) referentes à tabela ANOVA (ver Quadro 2) e às estimativas dos β's da regressão (ver Quadro 3).

Quadro 2 – Tabela ANOVA.

```
Residual standard error: 1.042 on 9 degrees of freedom
Multiple R-squared: 0.9538,     Adjusted R-squared: 0.9076
F-statistic: 20.65 on 9 and 9 DF,  p-value: 5.657e-05

Analysis of Variance Table

Response: y
                 Df  Sum Sq Mean Sq F value    Pr(>F)
FO(x1, x2, x3)    3 156.792  52.264 48.0974 7.267e-06
TWI(x1, x2, x3)   3   7.489   2.496  2.2973  0.146257
PQ(x1, x2, x3)    3  37.716  12.572 11.5696  0.001924
Residuals         9   9.780   1.087
Lack of fit       5   7.521   1.504  2.6635  0.181915
Pure error        4   2.259   0.565
```

Quadro 3 – Estimativas dos parâmetros do modelo de regressão linear múltipla ajustado.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.3180     0.4656 120.948 9.17e-16 ***
x1            1.7147     0.2821   6.079 0.000184 ***
x2            2.5505     0.2821   9.042 8.22e-06 ***
x3            1.4262     0.2821   5.056 0.000684 ***
x1:x2        -0.8437     0.3685  -2.289 0.047823 *
x1:x3         0.2162     0.3685   0.587 0.571787
x2:x3         0.4212     0.3685   1.143 0.282533
x1^2         -0.7141     0.2821  -2.532 0.032151 *
x2^2         -1.4688     0.2821  -5.207 0.000559 ***
x3^2         -0.8608     0.2821  -3.052 0.013757 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para realizar a análise de regressão no *software* STATISTICA, é preciso, primeiramente, selecionar o tipo de planejamento de experimentos adotado (planejamento composto central), conforme as Figuras 1 e 2 a seguir.

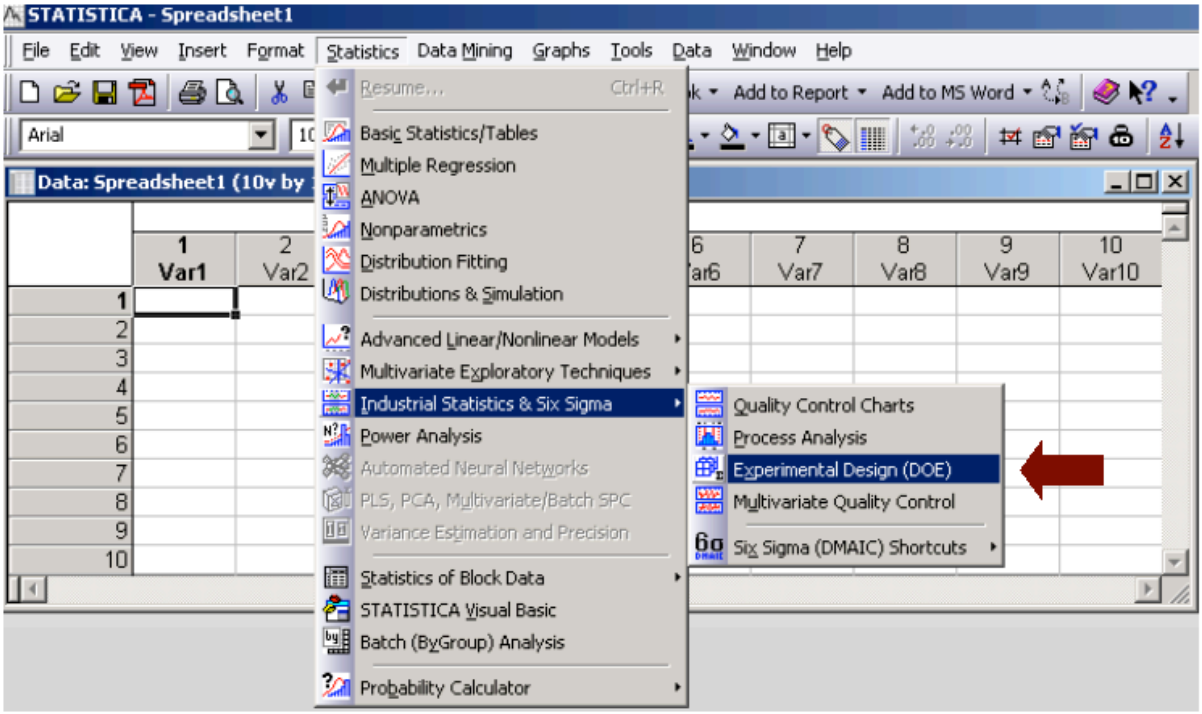


Figura 1 – Seleção do módulo de planejamento e análise de experimentos (DOE).

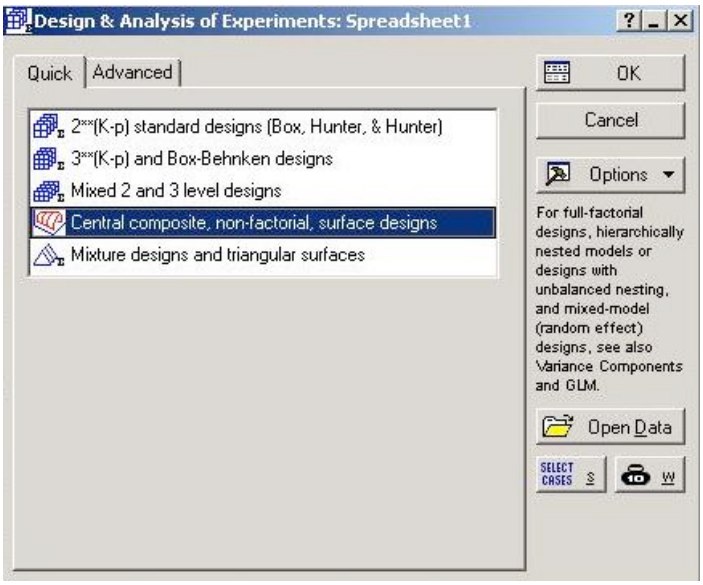


Figura 2 – Escolha do tipo de planejamento de experimentos adotado (composto central).

Após a escolha do planejamento composto central, definem-se as variáveis do modelo, conforme as Figuras 3 e 4.

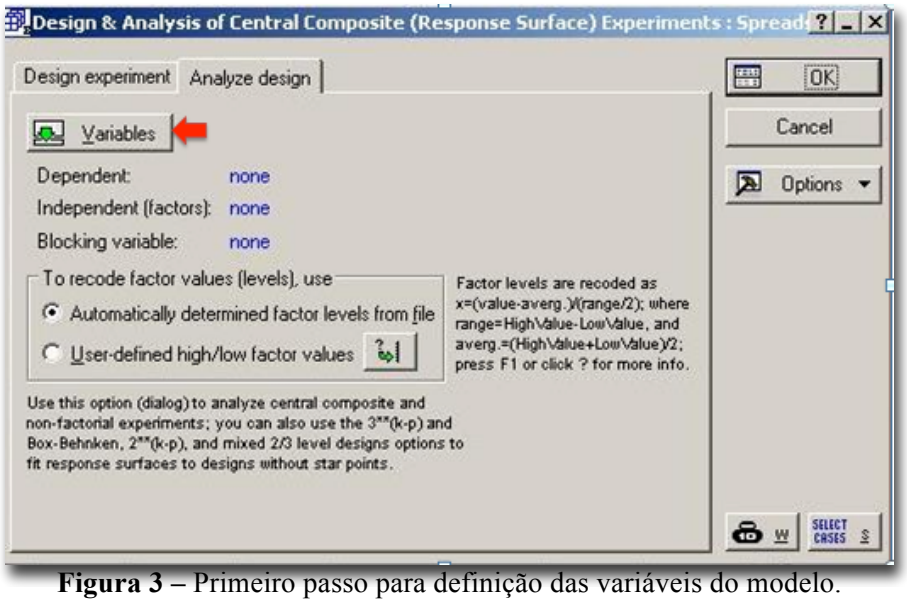


Figura 3 – Primeiro passo para definição das variáveis do modelo.

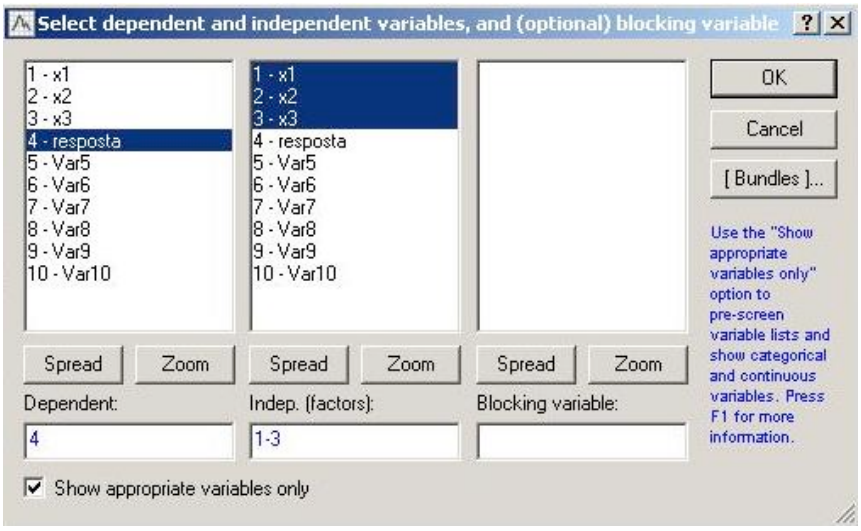


Figura 4 – Definição das variáveis.

Em seguida, na caixa *Include in model* da alça *Model* (ver Figura 5) definem-se as variáveis/efeitos (interação, termos lineares, termos quadráticos) presentes no modelo a ser ajustado.

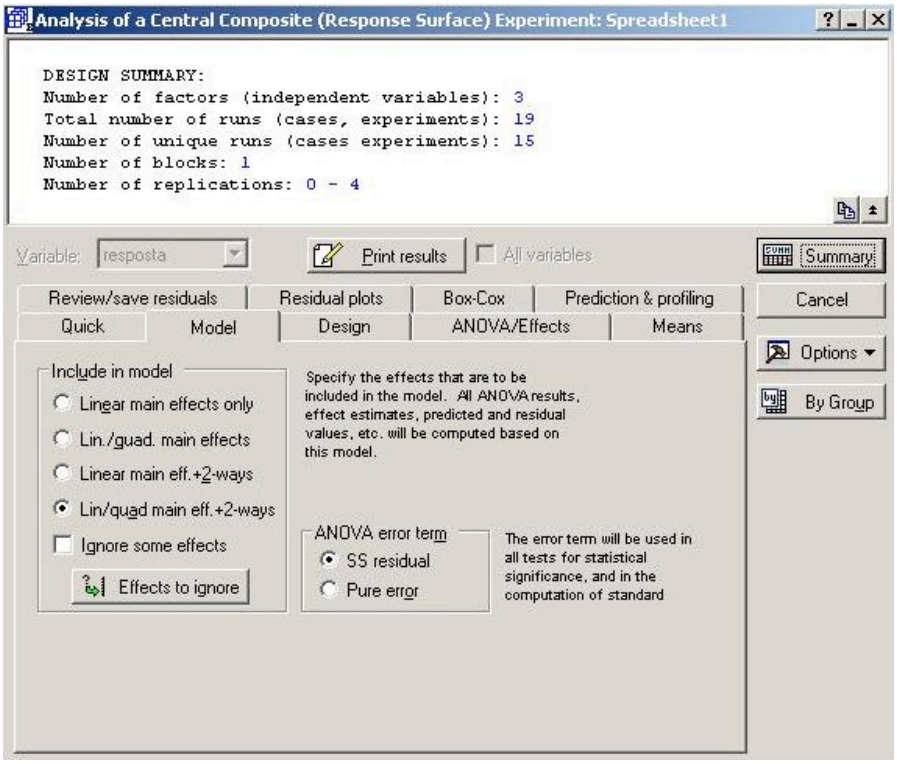


Figura 5 – Escolha do modelo a ser ajustado (efeitos).

Agora, é necessário fazer o ajuste do modelo, clicando em *ANOVA table* na alça *Quick* (ver Figura 6). Para a obtenção dos coeficientes estimados, basta clicar em *Summary: Effect estimates* nesta mesma alça.

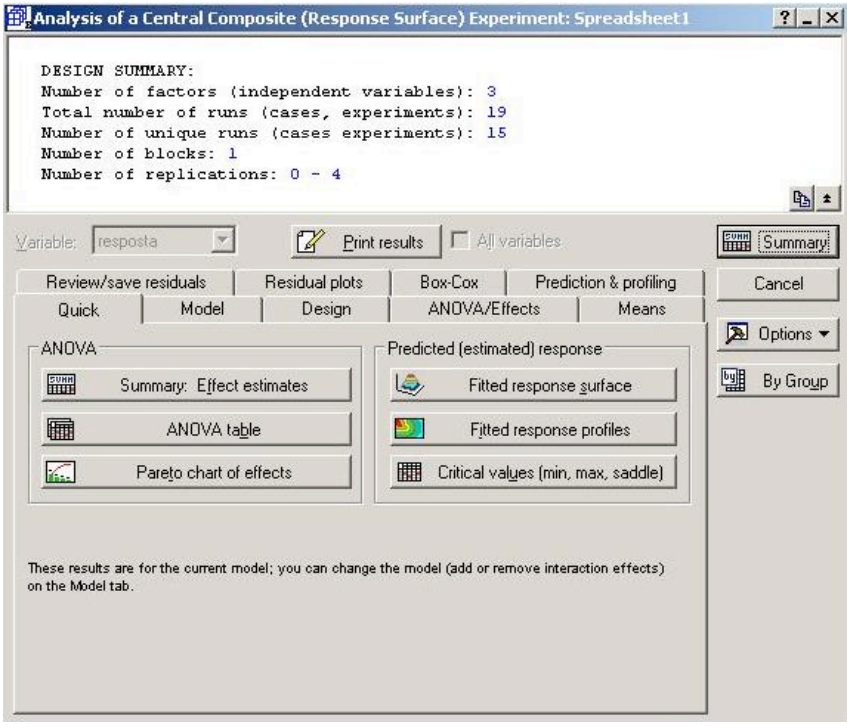
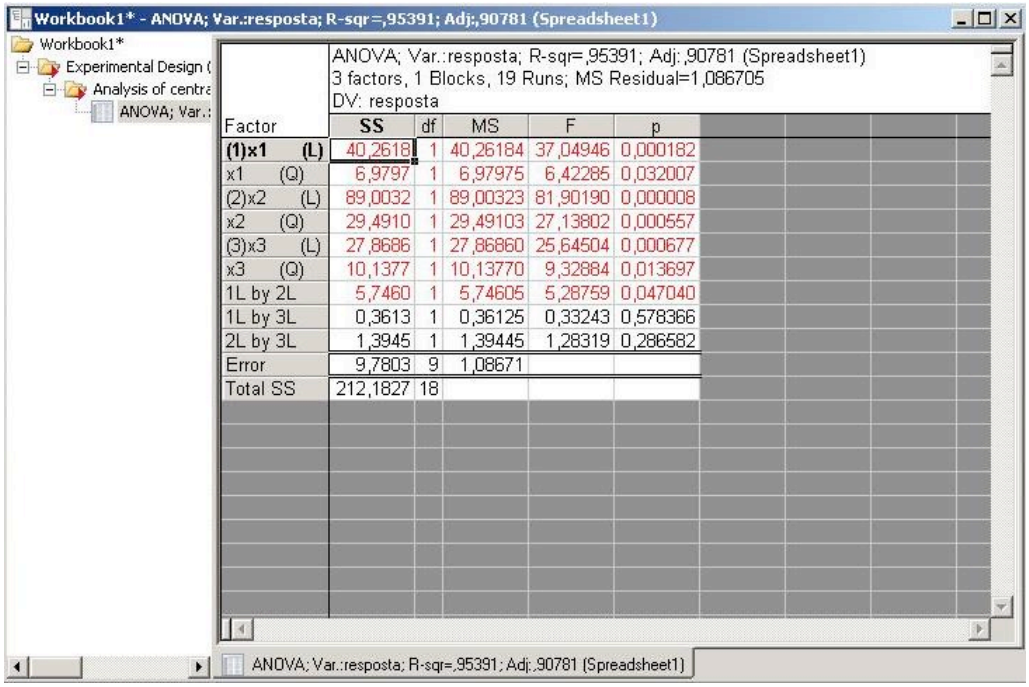


Figura 6 – Tabela de função do *software* (alça *Quick*).

Seguindo as instruções anteriores, são obtidas as seguintes tabelas (saídas do STATISTICA), apresentadas nas Figuras 7 e 8.



Workbook1* - ANOVA; Var.:resposta; R-sqr=,95391; Adj:,90781 (Spreadsheet1)

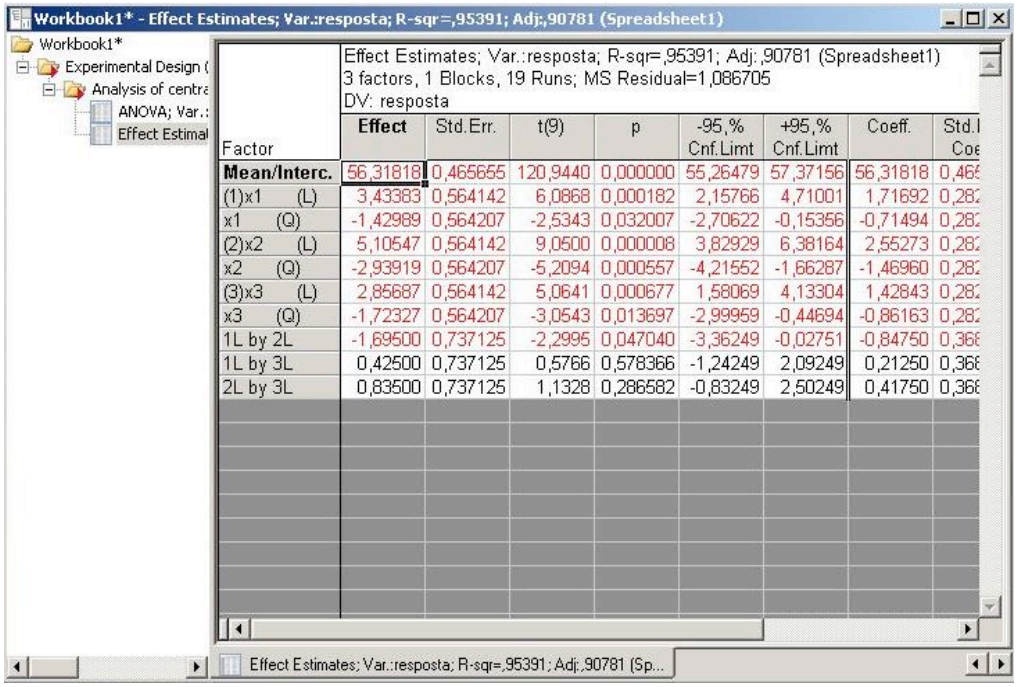
Workbook1*
Experimental Design (
Analysis of centra
ANOVA; Var.:

ANOVA; Var.:resposta; R-sqr=,95391; Adj:,90781 (Spreadsheet1)
3 factors, 1 Blocks, 19 Runs; MS Residual=1,086705
DV: resposta

Factor	SS	df	MS	F	p
(1)x1 (L)	40,2618	1	40,26184	37,04946	0,000182
x1 (Q)	6,9797	1	6,97975	6,42285	0,032007
(2)x2 (L)	89,0032	1	89,00323	81,90190	0,000008
x2 (Q)	29,4910	1	29,49103	27,13802	0,000557
(3)x3 (L)	27,8686	1	27,86860	25,64504	0,000677
x3 (Q)	10,1377	1	10,13770	9,32884	0,013697
1L by 2L	5,7460	1	5,74605	5,28759	0,047040
1L by 3L	0,3613	1	0,36125	0,33243	0,578366
2L by 3L	1,3945	1	1,39445	1,28319	0,286582
Error	9,7803	9	1,08671		
Total SS	212,1827	18			

ANOVA; Var.:resposta; R-sqr=,95391; Adj:,90781 (Spreadsheet1)

Figura 7 – Tabela ANOVA.



Workbook1* - Effect Estimates; Var.:resposta; R-sqr=,95391; Adj:,90781 (Spreadsheet1)

Workbook1*
Experimental Design (
Analysis of centra
ANOVA; Var.:
Effect Estimat

Effect Estimates; Var.:resposta; R-sqr=,95391; Adj:,90781 (Spreadsheet1)
3 factors, 1 Blocks, 19 Runs; MS Residual=1,086705
DV: resposta

Factor	Effect	Std.Err.	t(9)	p	-95,% Cnf.Limt	+95,% Cnf.Limt	Coeff.	Std.I Coe
Mean/Interc.	56,31818	0,465655	120,9440	0,000000	55,26479	57,37156	56,31818	0,46
(1)x1 (L)	3,43383	0,564142	6,0868	0,000182	2,15766	4,71001	1,71692	0,28
x1 (Q)	-1,42989	0,564207	-2,5343	0,032007	-2,70622	-0,15356	-0,71494	0,28
(2)x2 (L)	5,10547	0,564142	9,0500	0,000008	3,82929	6,38164	2,55273	0,28
x2 (Q)	-2,93919	0,564207	-5,2094	0,000557	-4,21552	-1,66287	-1,46960	0,28
(3)x3 (L)	2,85687	0,564142	5,0641	0,000677	1,58069	4,13304	1,42843	0,28
x3 (Q)	-1,72327	0,564207	-3,0543	0,013697	-2,99959	-0,44694	-0,86163	0,28
1L by 2L	-1,69500	0,737125	-2,2995	0,047040	-3,36249	-0,02751	-0,84750	0,36
1L by 3L	0,42500	0,737125	0,5766	0,578366	-1,24249	2,09249	0,21250	0,36
2L by 3L	0,83500	0,737125	1,1328	0,286582	-0,83249	2,50249	0,41750	0,36

Effect Estimates; Var.:resposta; R-sqr=,95391; Adj:,90781 (Sp...

Figura 8 – Tabela com as estimativas (efeitos) dos parâmetros do modelo ajustado.

Embora as saídas da tabela ANOVA apresentadas pelos dois *softwares* (Quadro 2 e Figura 7) sejam ligeiramente diferentes, observa-que as mesmas variáveis (ver Quadro 3 e

Figura 8) são significativas no ajuste realizado: *o intercepto, o tempo de contato (termos linear e quadrático), a temperatura do processo (termos linear e quadrático), a concentração da solução osmótica (termos linear e quadrático), a interação entre tempo de contato e temperatura do processo.*

As estimativas dos parâmetros segundo os dois *softwares* (ver Quadro 3 e Figura 8) são parecidas e, de acordo com seus valores (os coeficientes dos termos lineares são todos positivos), pode-se afirmar que, aumentando o tempo de contato, a temperatura do processo e a concentração da solução osmótica, obtém-se um aumento, em %, na perda de peso da fruta (desidratações mais intensas).

O modelo de regressão linear múltipla ajustado é apresentado abaixo:

$$\hat{y} = 56.32 + 1.71x_1 + 2.55x_2 + 1.43x_3 - 0.71x_1^2 - 1.47x_2^2 - 0.86x_3^2 - 0.84x_1x_2 + 0.22x_1x_3 + 0.42x_2x_3.$$

Para avaliar a adequabilidade (qualidade) do ajuste realizado, é necessário verificar se as hipóteses de normalidade e variância constante (homoscedasticidade) dos resíduos são satisfeitas.

No R, isso pode ser feito a partir dos comandos apresentados no Quadro 4, que resultam na geração dos gráficos normal probabilístico, dos valores observados vs. valores preditos, e dos resíduos vs. valores preditos.

Quadro 4 – Comandos em *R* para a geração dos principais gráficos de análise de resíduos.

```
> modelo <- rsm(y~SO(x1, x2, x3))
> a <- fitted(modelo)              #preditos do modelo
> t <- resid(modelo)               #residuos do modelo
> plot(y ,a ,main="Valores preditos vs. observados ",ylab="Valores preditos",xlab="Perda de
peso,%",pch=1,lwd=3,cex.lab=1.5, cex.main=1.5)
> abline(0,1)          #obter o gráfico preditos vs. observados

> plot(modelo)         #clicar duas vezes na tela do gráfico. O segundo gráfico que aparece neste

comando é da reta de probabilidade normal. Para salvá-lo em outro documento, é só fazer ctrl c

e ctrl v


> plot(a ,t ,main = "Valores preditos vs. resíduos ", ylab = " Resíduos ", xlab = " Valores preditos
", pch = 1, lwd = 3, cex.lab = 1.5, cex.main = 1.5)
> abline(0,1)          #obter o gráfico resíduos vs. preditos
```

A partir de tais comandos (Quadro 4), são obtidos os gráficos apresentados nas Figuras 9, 10 e 11.

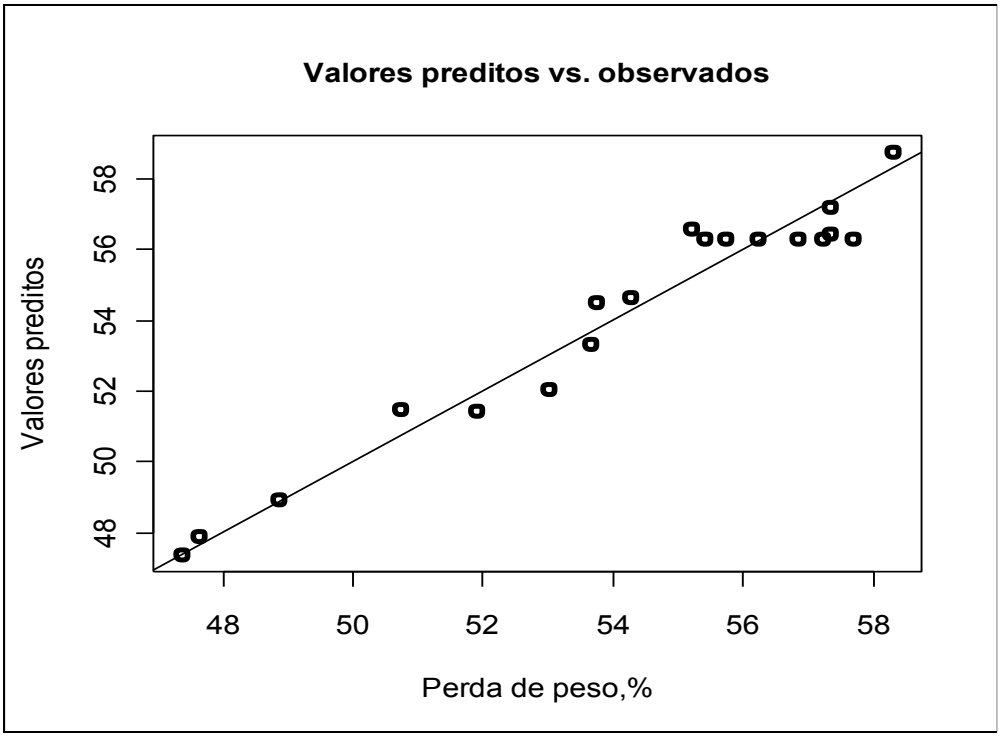


Figura 9 – Gráfico dos valores observados vs. valores preditos, gerado pelo R.

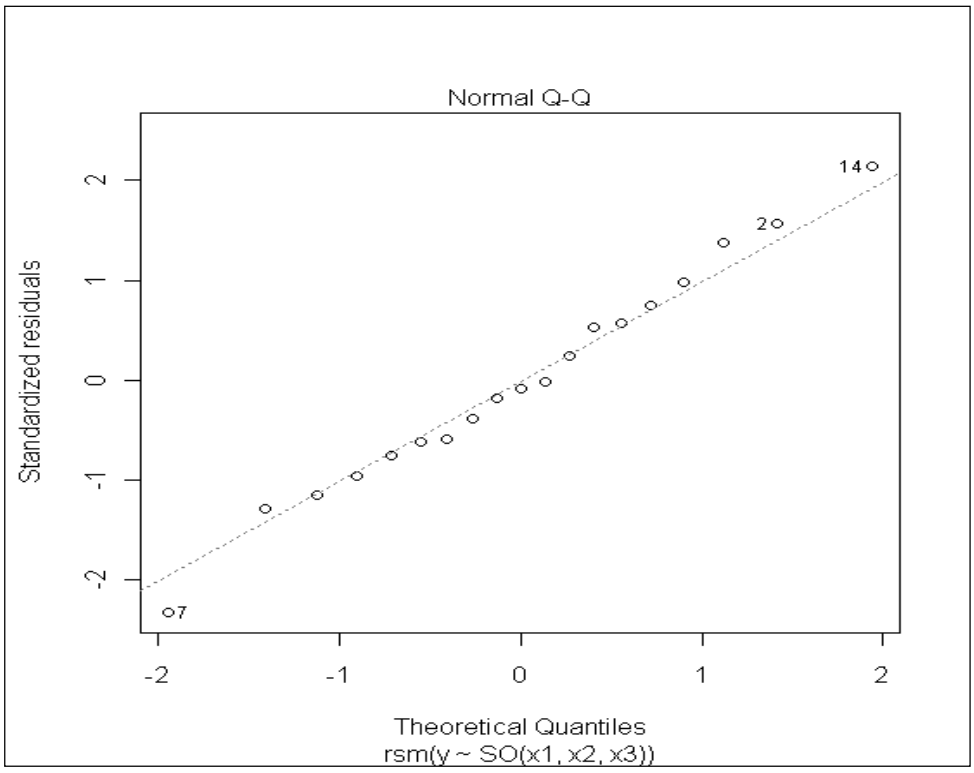


Figura 10 – Gráfico normal probabilístico dos resíduos, gerado pelo R.

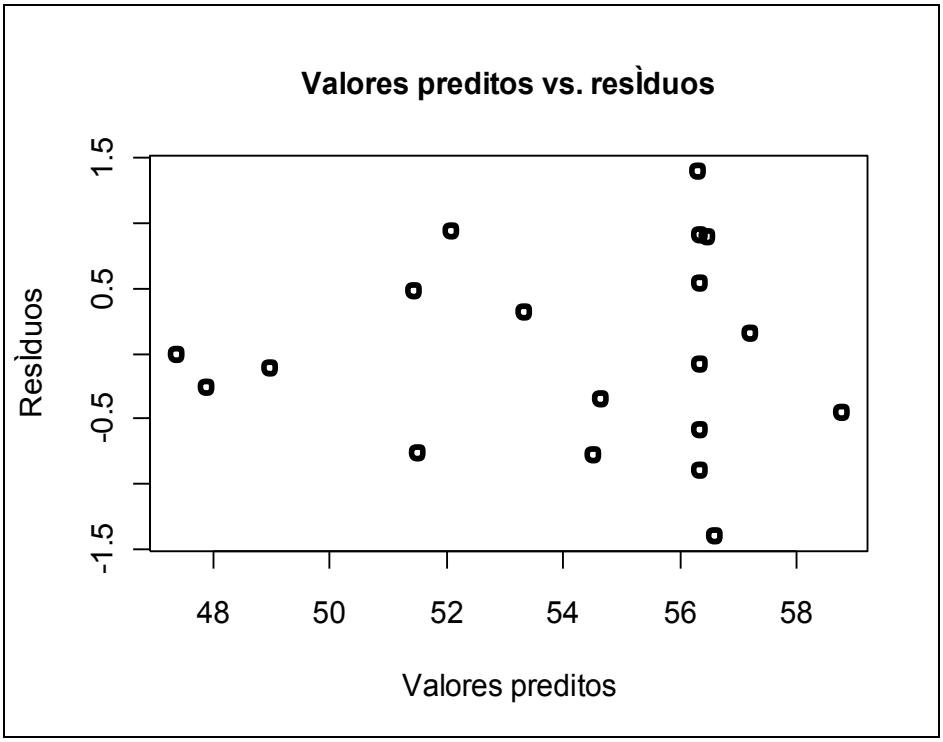


Figura 11 – Gráfico dos resíduos vs. valores preditos, gerado pelo R.

Para a geração de tais gráficos no STATISTICA, é preciso seguir os seguintes passos, de acordo com a Figura 12. Na alça *Residual plots*, clique em *'Observed vs. predicted values'*, *'Normal plot'* e *'Predicted vs. residual values'*, para obter os gráficos apresentados nas Figuras 13, 14 e 15, respectivamente.

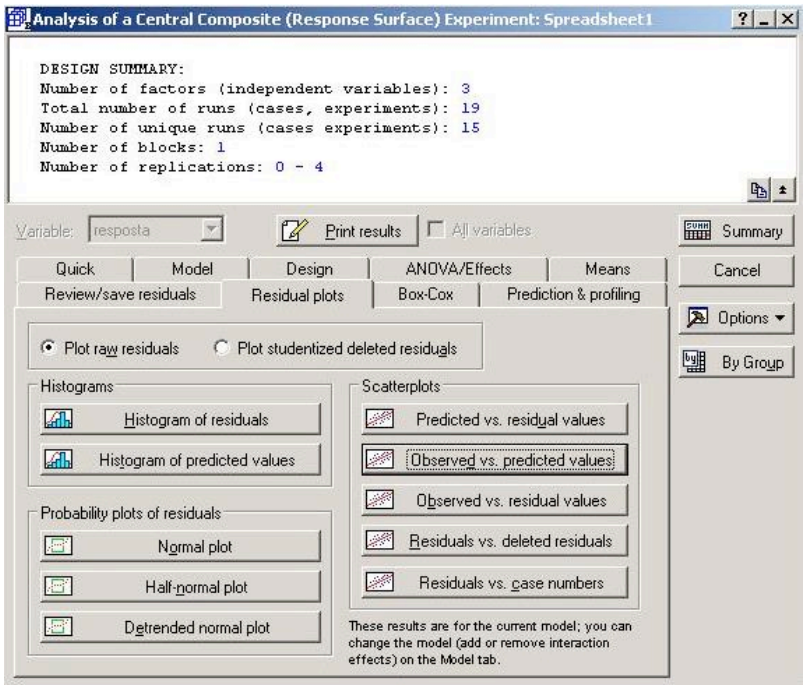


Figura 12 – Janela que contém os principais gráficos utilizados na análise de resíduos.

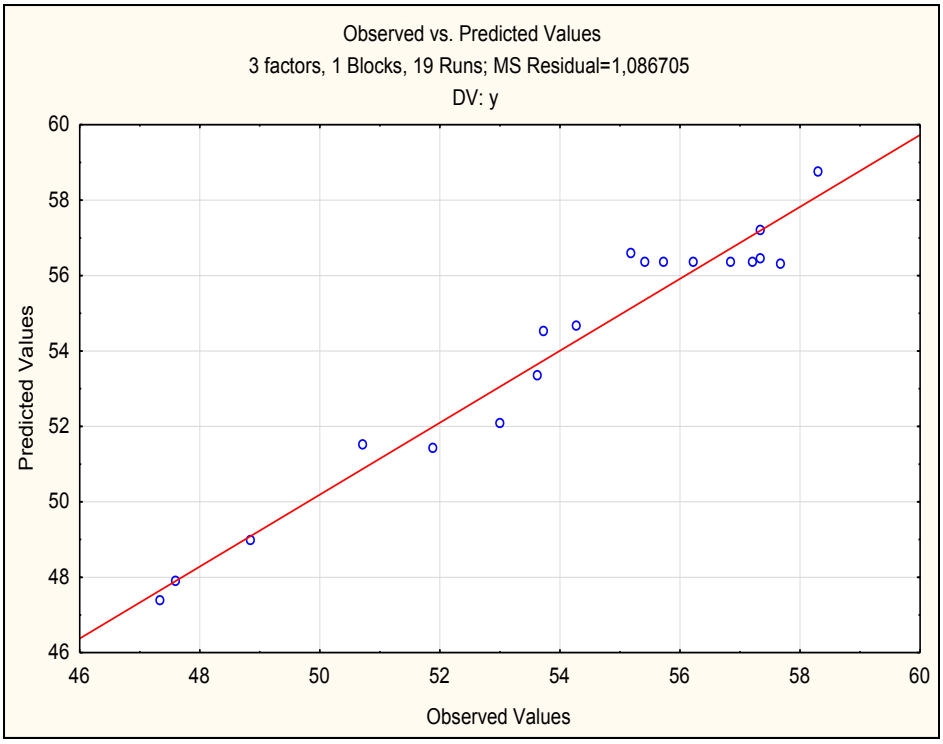


Figura 13 – Gráfico dos valores observados vs. valores preditos, gerado pelo STATISTICA.

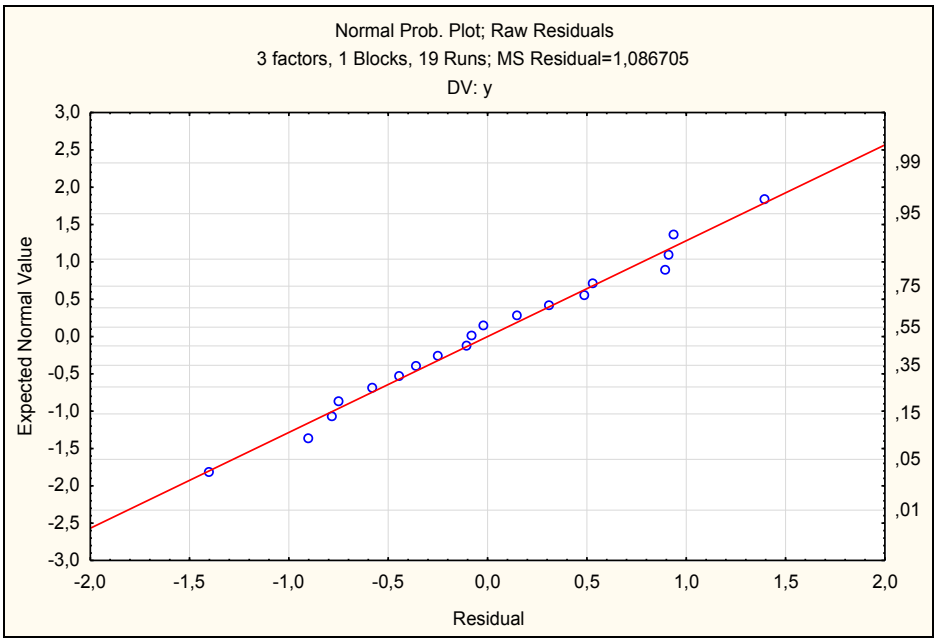


Figura 14 – Gráfico normal probabilístico dos resíduos, gerado pelo STATISTICA.

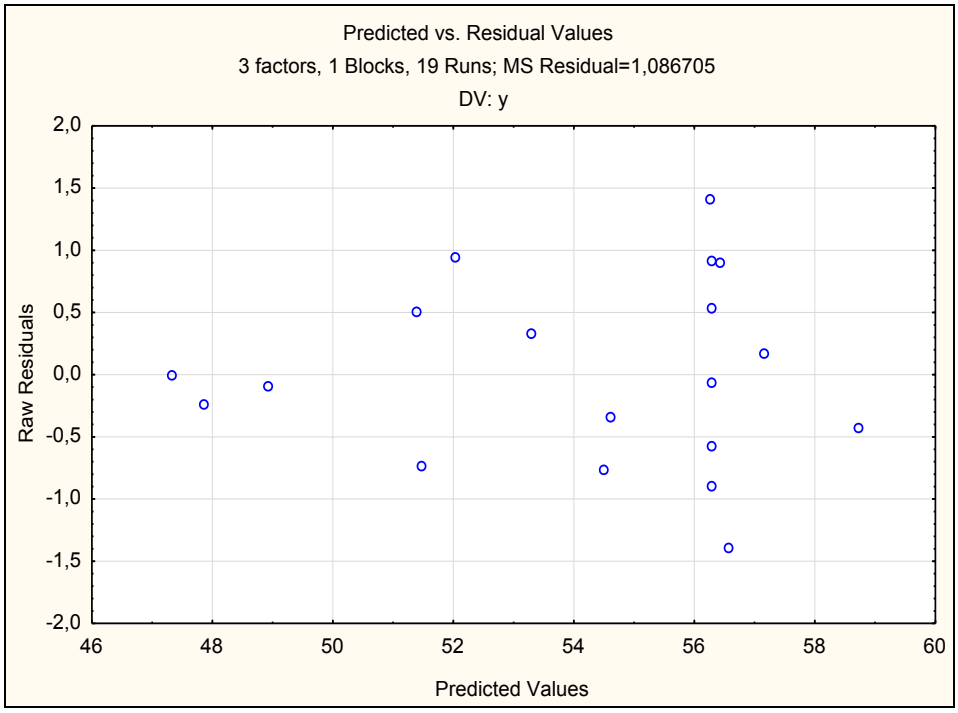


Figura 15 – Gráfico dos resíduos vs. valores preditos, gerado pelo STATISTICA.

Analisando os gráficos anteriores (Figuras 9, 10, 11, 13, 14 e 15), pode-se concluir que os resíduos do modelo ajustado seguem distribuição normal, pois, pelas Figuras 10 e 14, observa-se que os pontos se aproximam da reta identidade. E os resíduos atendem também aos pressupostos de variância constante (homoscedasticidade), como mostram as Figuras 11 e 15.

5. Conclusão

A quantidade de dados disponíveis atualmente para a análise é, em sua grande maioria, bastante extensa e requer a utilização de *softwares* habilitados para a leitura, geração de gráficos e tabelas indispensáveis para o direcionamento das informações, tornando-as úteis e fazendo com que a tomada de decisões seja correta.

Diante do que foi apresentado neste trabalho, é plausível dizer que pesquisadores podem considerar ambos os *softwares* R e STATISTICA para realizar as análises, desde que toda a teoria estatística seja seguida de forma correta. Particularmente, no presente artigo são apresentados ajustes de modelos de regressão linear múltipla usando ambos os *softwares*.

Esses dois *softwares* são constantemente melhorados e aperfeiçoados ao longo dos anos e, hoje em dia, há muitos usuários que fazem uso deles, seja em universidades ou empresas. E, na maioria das vezes, esse uso tem como objetivo otimizar o tempo gasto para fazer análises,

facilitar os cálculos e obter resultados precisos e confiáveis, seja em pesquisas científicas, trabalhos acadêmicos ou em trabalhos empresariais.

Bibliografia

- BARROS NETO, B.; SCARMINIO, I. S.; BRUNS, R. E. **Como fazer experimentos**. 3. ed. Campinas: Editora Unicamp , 2007.

- Estatística no programa R. Disponível em <http://www.estatisticanor.xpg.com.br/4.html>. Acessado em 21 de outubro de 2011.

- MONTGOMERY, D.C. **Design and Analysis of Experiments.** Hoboken, NJ : John Wiley, 2009.

- SOUZA, Emanuel F. M. de. PETERNELLI, Luiz A. MELLO, Márcio P de. **Software Livre R: aplicação estatística**. Disponível em: <http://www2.ufersa.edu.br/portal/view/uploads/setores/137/Apostilas%20e%20Tutoriais%20-%20R%20Project/Apostila%20R%20-%20GenMelhor.pdf>. Acessado em 8 setembro de 2011.

- *StatSoft Company History*. Disponível em: <http://www.statsoft.com/company/history/>. Acessado em 21 de outubro de 2011.