UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUISTA FILHO"
FACULDADE DE CIÊNCIAS AGRONÔMICAS
CAMPUS DE BOTUCATU

COMPARAÇÃO ENTRE OS ESTIMADORES DE MÍNIMOS QUADRADOS ORDINÁRIOS E MÍNIMOS DESVIOS ABSOLUTOS EM MODELOS DE REGRESSÃO LINEAR SIMPLES – UMA APLICAÇÃO NA ENERGIA NA AGRICULTURA

MÁRCIA APARECIDA ZANOLI MEIRA E SILVA

Tese apresentada à Faculdade de Ciências Agronômicas da UNESP - Campus de Botucatu, para obtenção do título de Doutor em Agronomia - Área de Concentração: Energia na Agricultura.

BOTUCATU - SP
Janeiro – 2002

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUISTA FILHO"
FACULDADE DE CIÊNCIAS AGRONÔMICAS
CAMPUS DE BOTUCATU

COMPARAÇÃO ENTRE OS ESTIMADORES DE MÍNIMOS QUADRADOS ORDINÁRIOS E MÍNIMOS DESVIOS ABSOLUTOS EM MODELOS DE REGRESSÃO LINEAR SIMPLES – UMA APLICAÇÃO NA ENERGIA NA AGRICULTURA

MÁRCIA APARECIDA ZANOLI MEIRA E SILVA

Orientador: Prof. Dr. José Raimundo de Souza Passos

Tese apresentada à Faculdade de Ciências Agronômicas da UNESP - Campus de Botucatu, para obtenção do título de Doutor em Agronomia - Área de Concentração: Energia na Agricultura.

BOTUCATU - SP
Janeiro – 2002

Aos meus pais, **Fausto** e **Otília**.

Ao meu marido, **Júlio**,

e aos meus filhos, **Amanda** e **Thiago**.

AGRADECIMENTOS

À Deus, por permitir a conclusão desse projeto.

Ao Prof. Dr. José Raimundo de Souza Passos, pela amizade, estímulo, ensinamentos, orientação e principalmente por acreditar na realização desse trabalho.

À amiga Profa. Dra. Andréa Carla Gonçalves Vianna, pela amizade, incentivo constante e importante colaboração na parte computacional, o meu muito obrigada.

À minha família, pelo apoio e incentivo.

À minha filha que, apesar da pouca idade, soube entender os momentos em que estive ausente.

Ao meu filho que, através das batalhas que enfrentamos com sua chegada, deu-me força e coragem para remover as pedras encontradas no meu caminho, mostrando-me que elas existem para valorizar as nossas conquistas.

Em especial ao meu marido, pela paciência, carinho, compreensão e apoio constante.

SUMÁRIO

Página

LISTA DE FIGURAS VI
LISTA DE QUADROS VIII
1 RESUMO 01
2 SUMMARY 03
3 INTRODUÇÃO 05
4 REVISÃO DE LITERATURA 10
4.1 Regressão linear simples 10
4.1.1 Estimação dos parâmetros 13
4.2 Regressão L_1 14
4.3 O modelo Gama de distribuição de probabilidade 20
4.3.1 A distribuição Gama 20
4.3.2 Propriedades da distribuição Gama 22
4.3.3 Assimetria e curtose 22
4.3.4 A distribuição Gama padronizada 24
5 MATERIAL E MÉTODOS 25
5.1 Material 25
5.2 Metodologia utilizada 26
5.2.1 Simulação dos dados 26
5.2.2 Estimação dos parâmetros 28
5.2.3 Comparação dos estimadores 29
5.3 Material e método utilizado na aplicação 30
6 RESULTADOS E DISCUSSÃO 32
6.1 Estimadores de regressão L_1 32
6.2 Comparação dos estimadores 33
6.2.1 Análise comparativa dos métodos de mínimos quadrados e mínimos desvios absolutos com base no REQM – Razão dos Erros Quadráticos Médios..... 33
6.2.1.1 Efeito do parâmetro de escala (λ) do modelo Gama na REQM 33

6.2.1.2 Efeito das variações incluídas na simulação na REQM 34
6.2.2 Análise comparativa dos métodos de mínimos quadrados e mínimos desvios absolutos com base na DEQM – Diferença dos Erros Quadráticos Médios 38
6.2.2.1 Efeito do parâmetro de escala (λ) do modelo Gama na DEQM 38
6.2.2.2 Efeito das variações incluídas na simulação na DEQM 39
6.2.3 Análise comparativa dos métodos de mínimos quadrados e mínimos desvios absolutos com base em suas variância residuais médias 42
6.3 Aplicação 44
7 CONCLUSÕES 50
8 BIBLIOGRAFIA CONSULTADA 51
APÊNDICE 1: Momentos 55
APÊNDICE 2: Programa fonte em linguagem SAS para a simulação e a função RANGAM. ...58
APÊNDICE 3: Algoritmo de regressão L_1 60
APÊNDICE 4: Histogramas do teor de umidade do milho segundo cultivar safra 63

LISTA DE FIGURAS

Figura		Página
1	Efeito do parâmetro de escala (λ) do modelo Gama de distribuição de probabilidade na Razão dos Erros Quadráticos Médios (REQM), obtidos por simulação (n=5400), considerando tamanho amostral, coeficiente linear e angular do modelo de regressão linear, parâmetro de escala e de forma do modelo Gama e o coeficiente de assimetria (1000 repetições para cada combinação).....	34
2	Efeito dos coeficientes linear e angular do modelo de regressão (β_0) e (β_1), respectivamente, na Razão dos Erros Quadráticos Médios (REQM), obtidos por simulação (1000 repetições), segundo valores de do parâmetro de escala λ do modelo Gama.....	35
3	Efeito do tamanho amostral e do coeficiente de assimetria na Razão dos Erros Quadráticos Médios (REQM), obtidos por simulação (1000 repetições), segundo valores do parâmetro de escala λ do modelo Gama.....	37
4	Efeito do parâmetro de escala (λ) do modelo Gama de distribuição de probabilidade na Diferença dos Erros Quadráticos Médios (DEQM), obtidos por simulação (n=5400), considerando tamanho amostral, coeficiente linear e angular do modelo de regressão linear, parâmetro de escala e de forma do modelo Gama e o coeficiente de assimetria (1000 repetições para cada combinação).....	39
5	Efeito dos coeficientes linear e angular do modelo de regressão (β_0) e (β_1) respectivamente, na Diferença dos Erros Quadráticos Médios (DEQM), obtidos por simulação (1000 repetições), segundo valores de do parâmetro de escala λ do modelo Gama.....	40
6	Efeito do tamanho amostral e do coeficiente de assimetria na Diferença dos Erros Quadráticos Médios (DEQM), obtidos por simulação (1000 repetições), segundo valores do parâmetro de escala λ do modelo Gama.....	42

7 Variâncias residuais médias dos ajustes dos modelos de regressão linear simples utilizando os métodos de mínimos quadrados e mínimos desvios absolutos, obtidos por simulação (n=5400), considerando tamanho amostral, coeficiente linear e angular do modelo de regressão linear, parâmetro de escala e de forma do modelo Gama e o coeficiente de assimetria (1000 repetições para cada combinação)..... 43

8 Ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, cultivar1-safra2 (Guiscem, 2001)..... 46

9 Ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, cultivar7-safra1 (Guiscem, 2001)..... 46

10 Ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, cultivar7-safra3 (Guiscem, 2001)..... 46

11 Representação de alguns modelos probabilísticos no plano assimetria curtose ($(\alpha_3)^2\alpha_4$), pontos sob a curva do modelo Gama usado para a simulação de Monte Carlo e as amostras usadas na aplicação..... 49

12 Histogramas do teor de umidade de milho segundo cultivar1-safra2, cultivar7-safra1 e cultivar7-safra3 .....

LISTA DE QUADROS

Quadro		Página
1	Valores obtidos para a assimetria α_3 através da relação $\alpha_3 = 2/\sqrt{\eta}$ variando η de 0,25 a 9,75 com incremento de 0,50.....	28
2	Análise de variância referente ao ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, segundo cultivar safra (Guiscem, 2001).....	45
3	Estimativa dos parâmetros referente ao ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade do milho, segundo cultivar safra (Guiscem, 2001), obtidos por regressão L_1 ($\hat{\beta}_{0q}$, $\hat{\beta}_{1q}$) e por mínimos quadrados ($\hat{\beta}_{0mq}$, $\hat{\beta}_{1mq}$).....	47
4	Razão entre o erro quadrático médio da regressão L_1 e o erro quadrático médio dos mínimos quadrados para o conjunto de dados de teor de umidade do milho, segundo cultivar safra (Guiscem, 2001).....	48

1 RESUMO

Em muitas situações práticas em Energia na Agricultura pode-se utilizar modelos de regressão linear simples com o objetivo de compreender determinados fenômenos de interesse. No entanto, apesar de sua aparente simplicidade, esses modelos possuem certas pressuposições que devem ser observadas pelo pesquisador, como por exemplo, a normalidade dos erros, cuja violação traz sérios problemas com relação à qualidade dos estimadores de mínimos quadrados obtidos, podendo comprometer as conclusões do estudo. Desse modo, os modelos de regressão L_1, que tem como base a minimização da soma dos desvios absolutos, surgem como uma alternativa viável, pois fornecem estimadores robustos com relação à normalidade. Neste estudo, inicialmente, foram feitas comparações empíricas (simulação de Monte Carlo) entre os estimadores de mínimos quadrados e mínimos desvios absolutos de modelos de regressão linear simples com distribuição Gama padronizada, considerando a variação do parâmetro de escala entre 0,2 e 2,2 com incremento de 0,4 e o parâmetro de forma variando de 0,25 a 9,75 com incremento de 0,5. Foram, também, considerados os tamanhos de amostra variando de 20 a 100 com incremento de 20, com 1000 replicações. Nestas comparações, observou-se que: a razão e a diferença de erros quadráticos médios além de poderem ser usados com critérios para comparação da qualidade de estimadores, não diferindo entre si, produzem resultados diferentes do critério usual da variância residual; o parâmetro (λ) de escala do modelo Gama de probabilidade é responsável por diferenciar a qualidade dos

estimadores: $\lambda \leq 1$ o estimador de mínimos quadrados produz menor erro quadrático médio, caso contrário, o melhor estimador é o de mínimos desvios absolutos. Posteriormente, como aplicação prática desta metodologia, foram ajustados modelos de regressão linear simples a um conjunto de dados de teor de umidade em grãos de milho (*Zea mays* L.), cuja análise dos resultados obtidos, confirmaram a eficiência da regressão L_1.

COMPARISONS BETWEEN ORDINARY LEAST SQUARE AND LEAST ABSOLUTE ERRORS ESTIMATORS IN THE LINEAR REGRESSION – AN AGRICULTURE APPLICATION. Botucatu, 2002. 65. Tese (Doutorado em Agronomia/Energia na Agricultura) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista.

Author: Márcia Aparecida Zanoli Meira e Silva

Adviser: José Raimundo de Souza Passos

2 SUMMARY

Linear regression models are widely used to understand many phenomena in agriculture. In the order hand, although apparently simple, these models have some assumptions that must be considered, like non-normal error. The violations of this assumption may mislead the conclusions of the results of the research. L_1 regression models, which are based in the least absolute error, are very attractive as an alternative technique, since they have robust estimators with respected to normality. In this study, firstly, empirical comparisons was done (Monte Carlo simulation) between the least square estimator and the least absolute error (L_1 regression) of the linear regression model with p.d.f. standard gamma distribution. The gamma parameters used in the simulations were: scale parameter varied from 0,2 to 2,2 by 0,4 and shape parameter varied from 0,25 to 9,75 by 0,5. The sampling size varied from 20 to 100 by 20, with 1.000 replications. In this comparisons, we see that the ratio and the difference between mean square error can be use to compare the quality estimators instead the residual variance estimated from the model. The shaper parameter of the Gamma model is responsible to choice between the two criteria, i.e., least square and least absolute error methods: When the shaper parameter is greater than 1,0 we choice the least absolute error methods, in the other hand, we choice the least square method. Further linear regression models were fitted using the two techniques, least square and the least absolute error, to the data of humidity in *Zea*

mays L. in time, for several cultivars and crops. It was shown that, in majority cases, the L_1 regression had estimators with mean squared error smaller than the least square estimator.

Keywords: Least square method, least absolute error, Mont Carlo method, mean square error, standard gamma distribution.

3 INTRODUÇÃO

Em muitas situações práticas em Energia na Agricultura – como, por exemplo, a variação da produção de uma determinada cultura em função da adubação utilizada e o crescimento de uma determinada cultura em função do tempo - o pesquisador observa que o modelo mais apropriado para o estudo de determinado fenômeno é um modelo de regressão linear simples (3.1):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \text{para } i = 1, 2, ..., m \qquad (3.1)$$

em que, Y_i é a variável resposta associada a *i-ésima* observação; β_0 e β_1 são os componentes do vetor de parâmetros b com β_0 e β_1 sendo, respectivamente, o intercepto e a inclinação da reta. O vetor de parâmetros é considerado como sendo uma constante populacional, de valor desconhecido; os X_i´s são considerados como constantes pré-fixadas, ou seja, não são variáveis aleatórias; e os ε_i´s são os componentes aleatórios chamados de erros, sendo estes responsáveis pelas flutuações de Y.

As pressuposições básicas desse modelo são:

1. Os valores de X são fixos, não existe variação aleatória em X;
2. Para cada valor de X, existe uma sub população de Y, e Y é normalmente distribuída $Y \sim N(\mu,\sigma^2)$;
3. A variância de Y não muda com X, é constante;
4. A esperança de Y dado X tem uma relação linear com X:
   $E(Y/X) = \beta_0+\beta_1 X$;
5. Os valores de ε_i´s, para $i = 1, ..., m$, são independentes e identicamente distribuídos (i.i.d.) - todos os erros possuem o mesmo modelo probabilístico, no caso o modelo normal - com média μ e variância σ^2, constante: $\varepsilon_i \sim N(\mu;\sigma^2)$.

A violação de algumas dessas pressuposições pode ter conseqüências tanto na estimativa pontual como no aspecto inferencial de b. Bolfarine et al. (1992) abordaram as conseqüências da violação do item 1, para o caso em que os X_i´s são variáveis aleatórias. Nos modelos de regressão linear simples, quando essas pressuposições são violadas, o uso dos estimadores de mínimos quadrados para a inclinação da reta, por exemplo, fica subestimado, o mesmo ocorrendo para o estimador da variância e o coeficiente de determinação R^2.

Quando encontramos situações que violam a pressuposição 5 com relação a não normalidade, não podemos, por exemplo, construir o intervalo de confiança para b com base na distribuição "t" de Student, ou proceder a Análise de Variância entre b, para o caso em que cada modelo de regressão corresponde à um tratamento.

A partir dessas pressuposições, pode-se demonstrar que os estimadores obtidos pelo método de mínimos quadrados são também estimadores obtidos pelo método de máxima verossimilhança. Estes são não tendenciosos, e dentro da classe dos estimadores lineares não tendenciosos, possuem a propriedade de variância mínima.

Para estimar os parâmetros, usou-se o método dos mínimos quadrados, minimizando-se a soma de quadrados dos erros ($Z(\beta_0, \beta_1)$), cujas soluções são obtidas tomando-se as derivadas parciais da função Z em relação à β_0 e β_1 e igualando-as a zero.

Assim obtém-se os estimadores:

$$\hat{\beta}_1 = \frac{m\sum_{i=1}^{m} X_i Y_i - \left(\sum_{i=1}^{m} X_i\right)\left(\sum_{i=1}^{m} Y_i\right)}{m\sum_{i=1}^{m} X_i^2 - \left(\sum_{i=1}^{m} X_i\right)^2} \qquad (3.2)$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Uma outra maneira de solucionar o problema apresentado anteriormente, é trabalhar com a minimização da soma dos desvios absolutos, isto é, determinar b tal que:

$$\min \sum_{i=1}^{m} \left|\varepsilon_i\right|$$

$$\text{sujeito a: } Xb + e = \mathbf{Y}, \qquad (3.3)$$

Como neste trabalho $n = 2$, tem-se

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \quad b = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad e = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \qquad (3.4)$$

tal que

$X \in \Re^{mx2}$, é a matriz obtida por um conjunto de *m* medidas de observações em 1 variável independente;

$b \in \Re^2$, é o vetor de parâmetros do modelo;

$\mathbf{Y} \in \Re^m$, é o vetor de medidas na variável dependente;

$e \in \Re^m$, denota o vetor de erros aleatórios (desvios ou ruídos).

Este problema é chamado de Problema de Aproximação Linear no L_1, pois minimizar a soma dos desvios absolutos é o mesmo que minimizar os resíduos na norma 1, já que $\|\varepsilon\|_1 = \sum_{i=1}^{m} |\varepsilon_i|$. Este problema também é conhecido simplesmente por Regressão L_1.

A regressão L_1 foi proposta meio século antes do desenvolvimento da regressão de mínimos quadrados. Entretanto, devido a dificuldades computacionais não existentes na regressão de mínimos quadrados e o extenso desenvolvimento da regressão de mínimos quadrados em termos da teoria da probabilidade, a regressão L_1 não recebeu a devida atenção.

Nas últimas quatro décadas o interesse sobre a regressão L_1 cresceu muito, devido a sua reformulação como um problema de programação linear, surgindo a seguir diversos algoritmos eficientes para resolver o problema.

A regressão de erros absolutos (L_1) é atualmente considerada, entre os procedimentos robustos para regressão, uma técnica viável para a análise de dados. Ela é resistente a valores aberrantes e não requer uma "constante de ajuste" como os demais procedimentos robustos. O seu uso é recomendado quando os erros têm uma distribuição simétrica (Narula & Stangenhaus, 1988).

A regressão de erros absolutos foi estudada em diversos contextos sob uma variedade de nomes, tais como, mínima soma dos erros absolutos, mínimos desvios, erros ou valores absolutos.

Silva (1994) estudou o Problema de Regressão L_1 considerando-o como um problema de programação linear e explorou o fato da função ser linear por partes. Esta abordagem propiciou uma teoria específica, tendo como conseqüência a elaboração de um novo algoritmo que, para os conjuntos de dados testados, se mostrou mais eficiente que o tradicional método dos mínimos quadrados nos casos de regressão linear.

Os objetivos desse trabalhos são:

a) comparar a eficiência dos estimadores de mínimos quadrados e os de mínimos desvios absolutos – Regressão L_1 (Silva, 1994), utilizando-se a razão e a diferença entre os erros quadráticos médios para modelos de regressão linear simples com erro Gama através de simulação;
b) comparar os dois métodos através de suas variâncias residuais médias;
c) verificar o efeito da assimetria, do tamanho amostral e dos coeficientes linear e angular do modelo de regressão teórico na eficiência desses estimadores;
d) proceder uma aplicação de modelos de regressão linear em Energia na Agricultura.

A contribuição deste trabalho deve-se ao fato de que o uso da razão e da diferença entre os erros quadráticos médios para realizar as comparações entre os dois métodos consideradbs, embora simples, não é uma metodologia encontrada na literatura. Além disso, este trabalho faz uso de um novo método de regressão L_1 (Silva, 1994), o que aumenta sua contribuição.

4 REVISÃO DE LITERATURA

4.1 Regressão Linear Simples

Nesta seção, alguns resultados importantes de regressão linear simples, mencionados por Hoffmann & Vieira (1987), serão apresentados a seguir.

Uma das preocupações na elaboração de modelos estatísticos é que estes sejam fiéis à realidade, ou seja, deve-se, sempre que possível, construir modelos que contenham em sua estrutura bases e fundamentos teórico-científicos, sejam eles Biológicos, Físicos, Químicos, Agronômicos ou Econômicos. Assim, uma estratégia conveniente de análise é supor que cada observação é formada por duas partes: uma previsível e outra aleatória. Desse modo, cada observação poderia ser representada por:

observação = previsível + aleatória

A primeira componente, a parte previsível, incorpora o conhecimento que o pesquisador tem sobre o fenômeno e é usualmente expressa por uma função matemática com parâmetros desconhecidos.

Para a segunda componente, devido ao seu caráter aleatório, impõe-se que os mesmos obedeçam a algum modelo de probabilidade.

Com essas suposições, o trabalho estatístico passa a ser aquele de produzir estimativas para os parâmetros desconhecidos, baseando-se em amostras observadas.

Como motivação, suponha que temos *m* pares de observações (X_i, Y_i), *i = 1, 2,..., m* Podemos traçar esses pontos e tentar ajustar a eles uma função, dentre uma família de funções pré estabelecidas (por exemplo, funções lineares, quadráticas, etc.), tal que os pontos estejam "o mais próximo possível" da função. Mesmo que as variáveis X e Y tivessem uma relação exata, os pontos traçados não satisfariam esta relação, devido aos erros de medidas.

Genericamente, tais relações funcionais podem ser representadas por

$$Y = f(X_1, X_2, ..., X_k)$$

em que Y representa a variável dependente e os X_i's, *i = 1, 2, ..., k*, representam as variáveis independentes.

São exemplos de relações funcionais entre variáveis:

a) crescimento da população de um país (Y) em função dos anos (X);
b) variação da produção (Y) obtida numa cultura conforme a quantidade de nitrogênio (X_1), fósforo (X_2) e potássio (X_3) utilizada na adubação;
c) variação do preço (Y) de um produto no mercado em função da quantidade oferecida (X).

Suponhamos que a relação entre as duas variáveis X e Y seja linear

$$Y_i = f(X_i) + \varepsilon_i.$$

Tal relação é conhecida como uma regressão linear simples de Y em X, onde e é o erro de medida e, seu modelo estatístico é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

em que β_0 e β_1 são parâmetros desconhecidos e se quer estimá-los.

O coeficiente angular da reta (β_1) é também denominado coeficiente de regressão e, o coeficiente linear da reta (β_0) é também conhecido como termo constante da equação de regressão ou intercepto.

Ao estabelecer o modelo de regressão linear simples, pressupõe-se que:

i) A relação entre X e Y é linear.

ii) Os valores de X são fixos, isto é, X não é uma variável aleatória.

iii) O valor esperado do erro é nulo, $E(\varepsilon_i) = 0$.

iv) Para um dado valor de X, a variância do erro e é sempre σ^2, denominada variância residual, isto é, $E(\varepsilon_i^2) = \sigma^2$

v) O erro de uma observação é não correlacionado com o erro em outra observação, isto é, $E(\varepsilon_i \varepsilon_j) = 0$ para $i \neq j$

vi) Os erros têm distribuição normal.

A pressuposição vi) é necessária para que se possa utilizar a distribuição *t* de Student e a distribuição *F* para testar as hipóteses relacionadas aos valores dos parâmetros ou construir intervalos de confiança. Sendo assim, ela é uma pressuposição adicional.

Combinando-se as pressuposições iii), iv) e vi) tem-se , como já mencionado anteriormente, que os erros ε_i's são variáveis aleatórias independentes com média zero, variância σ^2 e distribuição normal, ou seja, $\varepsilon_i \sim N(0,\sigma^2)$.

O erro ε_i do modelo de uma regressão linear pode ser devido à influência de todas as variáveis que afetam a variável dependente e que não foram incluídas no modelo. Uma vez que tais variáveis não foram consideradas, elas devem ser as variáveis de menor interesse para o pesquisador e seus efeitos devem ser todos relativamente pequenos. Considerando que o número de fatores que podem afetar certa variável dependente é bastante

grande, e desde que seus efeitos sejam aditivos e independentes, pode-se concluir que o erro residual tem distribuição aproximadamente normal.

4.1.1 Estimação dos parâmetros

O primeiro passo, na análise de regressão, é obter os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros β_0 e β_1 da regressão. Os valores desses estimadores serão obtidos a partir de uma amostra de *m* pares de valores (X_i, Y_i), *i = 1, 2, ..., m*, que correspondem a *m* pontos num gráfico.

Considerando

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

sendo ε_i´s os desvios da regressão, o método dos mínimos quadrados consiste em adotar como estimativas dos parâmetros os valores que minimizam a soma dos quadrados dos desvios

$$\text{minimizar } Z(\beta_0, \beta_1) = \sum_{i=1}^{m} [Y_i - b_0 - b_1 X_i)]^2$$

A função Z terá mínimo quando suas derivadas parciais de primeira ordem em relação à β_0 e β_1 forem nulas.

$$\frac{\partial Z}{\partial \beta_0} = -2 \sum_{i=1}^{m} [Y_i - \beta_0 - \beta_1 X_i] = 0 \qquad (4.1)$$

$$\frac{\partial Z}{\partial \beta_1} = 2 \sum_{i=1}^{m} [Y_i - \beta_0 - \beta_1 X_i](-X_i) = 0 \qquad (4.2)$$

Simplificando, chega-se ao sistema de equações normais

$$\begin{cases} m\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{m} X_i = \sum_{i=1}^{m} Y_i & (4.3) \\ \hat{\beta}_0 \sum_{i=1}^{m} X_i + \hat{\beta}_1 \sum_{i=1}^{m} X_i^2 = \sum_{i=1}^{m} X_i Y_i & (4.4) \end{cases}$$

Resolvendo o sistema tem-se:

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^{m} X^2\right)\left(\sum_{i=1}^{m} Y_i\right) - \left(\sum_{i=1}^{m} X_i\right)\left(\sum_{i=1}^{m} X_i Y_i\right)}{m\sum_{i=1}^{m} X_i^2 - \left(\sum_{i=1}^{m} X_i\right)^2}$$

$$\hat{\beta}_1 = \frac{m\sum_{i=1}^{m} X_i Y_i - \left(\sum_{i=1}^{m} X_i\right)\left(\sum_{i=1}^{m} Y_i\right)}{m\sum_{i=1}^{m} X_i^2 - \left(\sum_{i=1}^{m} X_i\right)^2}$$

Na prática determina-se $\hat{\beta}_1$ em primeiro lugar e da equação (4.3) tem-se

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{m} Y_i}{m} - \hat{\beta}_1 \frac{\sum_{i=1}^{m} X_i}{m} = \overline{Y} - \hat{\beta}_1 \overline{X}$$

4.2 Regressão L_1

Considere *m* o número de observações e *(n-1)* o número de variáveis independentes. Assumindo, por hipótese, que $n < m$ e posto($\mathbf{X}$) = n, a Regressão L_1, isto é, a regressão que minimiza a soma dos erros absolutos, ou simplesmente regressão de erros absolutos, consiste em determinar o vetor b, com componentes β_0 e β_1, tal que

$$\min \|\varepsilon\|_1 = \min \left(\sum_{i=1}^{m} |y_i - \beta_0 - \beta_1 X_i| \right)$$

$$\text{sujeito a: } Xb + e = \mathbf{Y}, \qquad (4.5)$$

sendo X, b, e, $\mathbf{Y}$ dados conforme (3.4).

A regressão L_1 é considerada uma alternativa robusta à regressão de mínimos quadrados por diversos autores. Esta regressão, segundo Narula & Stangenhaus (1988), não necessita de uma "constante de ajuste" como outros métodos robustos de regressão.

O fato da regressão de erros absolutos ser menos sensível à valores aberrantes do que a regressão de mínimos quadrados, pode ser parcialmente explicado pelo fato de que observações inconsistentes são tratadas de forma diferente pelos dois métodos. Se uma observação está afastada do restante das observações, o método dos mínimos quadrados é mais influenciado por esse dado do que pelos demais, o que é indesejável. A regressão de erros absolutos, por outro lado, assegura um esforço igual tanto com relação aos dados aberrantes quanto ao restante do conjunto de dados. Portanto, podemos esperar que os dados aberrantes irão destacar-se mais quando se usa a regressão de erros absolutos. Deve-se ressaltar que a regressão de mínimos quadrados está para a regressão de erros absolutos assim como a média amostral está para a mediana da amostra.

A literatura apresenta diversos casos em que a regressão de mínimos quadrados foi considerada inadequada. Por exemplo, em alguns modelos econômicos o erro absoluto é uma medida mais adequada de perda do que a função de perda quadrática, implícita na regressão de mínimos quadrados, sendo que aqui a perda representa a gravidade do erro de predição não nulo para o pesquisador. O uso da regressão L_1 foi recomendado para estudos em economia em que os erros com variância não-finita são mais representativos do que os erros com variância finita. Como exemplo podemos citar: estimação de funções de investimento e previsão de investimentos; detecção de erros em conjuntos de dados, a fim de tentar descobrir formações na subsuperfície, onde recursos minerais podem existir; processamento de dados

orbitais em objetos espaciais; astronomia; modelagem de diversos dados geofísicos; estimação de funções de custo; análise de dados sísmicos; estimação de parâmetros farmacocinéticos; e obtenção de estimativas de estados em sistemas de potência. E mais, em diversos estudos comparativos o desempenho da regressão de erros absolutos foi pelo menos igual, quando não foi melhor do que a regressão de mínimos quadrados (Narula & Stangenhaus, 1988).

Com relação aos seus estimadores, pode-se afirmar que o estimador de mínima soma dos erros absolutos não se altera quando ocorrem mudanças nos valores da variável resposta associados a resíduos não nulos, desde que a observação permaneça do mesmo lado do hiperplano de regressão L_1 (continua com resíduo positivo ou negativo). Além disso, os estimadores dos parâmetros da regressão L_1 são estimadores de máxima verossimilhança, portanto, são assintoticamente não-tendenciosos e eficientes, quando os erros tem distribuição de Laplace. Estes estimadores têm uma elipsóide de concentração estritamente menor do que os estimadores de mínimos quadrados quando a distribuição dos erros é tal que a mediana amostral é um estimador de posição mais eficiente do que a média amostral. Baseado em estudos de Monte Carlo, o uso da regressão L_1 tem sido recomendado nos casos em que a distribuição dos erros é Laplace, Cauchy e distribuições normais contaminadas.

Assim, a regressão L_1 tem melhor desempenho do que o método dos mínimos quadrados quando os erros têm distribuição simétrica para a qual a mediana amostral é um estimador de posição mais eficiente do que a média amostral, quando os erros têm distribuição com caudas alongadas ou quando a função de perda baseada nos erros absolutos é mais apropriada do que a função perda quadrática. A regressão L_1 é menos sensível a valores aberrantes do que o método dos mínimos quadrados e fornece uma boa solução inicial para vários procedimentos robustos de regressão.

Bloomfield & Steiger (1980) relatam que o Problema da Aproximação Linear no L_1 (regressão L_1) foi sugerido por Boscovitch na metade do século XVIII antes da introdução do método dos mínimos quadrados, porém não foi muito usado devido ao desconhecimento de algoritmos eficientes para problemas com vários parâmetros. Os autores

relatam ainda que, mais de 100 anos depois, Edgeworth propôs o ajuste de curva e de modelos mais complicados pela minimização irrestrita de (4.5). Entretanto, os cálculos são mais complexos que a solução das equações lineares que resultam do método dos mínimos quadrados. Além disso, relatam que em 1930 um novo método foi introduzido por Rhodes e discutido por Singleton em 1940. O desenvolvimento da programação linear e a observação de Harris (1950), citada pelos autores, de que o problema de aproximação linear no L_1 pode ser transformado em um problema de programação linear, foi um grande avanço. Assim, o conceito simplista de estimar desvios absolutos mínimos, associado ao desenvolvimento de técnicas com custo computacional competitivo tornam o problema consideravelmente importante.

Esta linha foi seguida por Wagner (1959), Barrodale & Roberts (1973) e Abdelmalek (1975), entre outros. Arenales (1991) cita também Narula & Wellington (1977), e Roberts & Ben-Israel (1970) e, afirma que embora com motivações diferentes, estes métodos são equivalentes.

Desta forma, o problema (4.5) é um problema típico de Programação Linear, em que problemas com estruturas particulares ocorrem com freqüência. Porém, a simples aplicação de um método geral de Programação Linear normalmente é muito dispendiosa. Assim, com o objetivo de obter grandes economias, tanto de memória como de tempo de processamento (pois se evita armazenar e operar com zeros), deve-se explorar a estrutura particular que cada problema possui.

Silva (1994) estudou o Problema de Regressão L_1 considerando-o como um problema de programação linear explorando o fato da função ser linear por partes, com o intuito de determinar sua solução numérica. Esta abordagem consistiu no desenvolvimento de todos os passos do Método Primal Simplex da Programação Linear para esse problema, gerando uma especialização do método simplex com um embasamento teórico que, como conseqüência, conduziu à elaboração de um novo algoritmo. A aplicação deste algoritmo para os conjuntos de dados testados mostrou-se mais eficiente que o tradicional método dos mínimos quadrados nos casos de regressão linear. A autora desenvolveu sua teoria e, conseqüentemente seu algoritmo, não apenas para regressão linear simples, mas para a

regressão linear múltipla e também fez uma generalização de seu método para os problemas de regressões quantílicas (Silva, 1994; Silva, 1997), uma vez que a regressão no L_1 é um caso particular das regressões quantílicas. Desenvolveu também uma análise pós-otimização, diferente das que constam na literatura, de forma bem simples e eficiente (Silva, 1994; Silva, 2000).

As Regressões Quantílicas são apresentadas em Koenker & Bassett (1978) como quantidades de interesse no desenvolvimento de procedimentos de estimativas robustas. Elas são definidas como soluções $\beta(\theta) \in \Re^n$ do problema:

$$\min \left(\sum_{i/\varepsilon_i < 0} (1-\theta)\varepsilon_i + \sum_{i/\varepsilon_i \geq 0} \theta\varepsilon_i \right)$$
$$\text{s.a. } Xb + e = \mathbf{Y}$$

tal que $\theta \in [0, 1]$.

Este problema também pode ser escrito como

$$\min \left(\sum_{i=1}^{m} \rho_\theta \, \varepsilon_i \right)$$
$$\text{s.a. } Xb + e = \mathbf{Y}$$

com

$$\rho_\theta = \begin{cases} \theta & \text{se } \varepsilon_i \geq 0 \\ (\theta - 1) & \text{se } \varepsilon_i < 0 \end{cases} \qquad \text{e} \qquad \theta \in [0, 1]$$

ou ainda

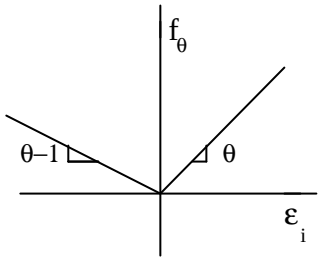
$$\min_b f_q(b, e) = \min \left(\sum_{i=1}^{m} f_q(e_i) \right)$$
$$\text{s.a. } Xb + e = \mathbf{Y} \tag{4.6}$$

tal que

$X \in \Re^{mxn}$, com posto(X) = n

$\theta \in [0, 1]$

f_θ é uma função linear por partes ilustrada a seguir:



Observe que para θ = 0,5 temos a regressão L_1, pois

$$f_{0,5}(\beta, \varepsilon) = \sum_{i=1}^{m} f_{0,5}(\varepsilon_i) = \frac{1}{2}\sum_{i=1}^{m} |\varepsilon_i|$$

e o fator 0,5 não interfere na minimização.

Algumas aplicações da regressão quantílica e seus principais fundamentos teóricos podem ser encontrados em Koenker & Portnoy (1997).

Stangenhaus & Narula (1991) utilizaram o método de Monte Carlo com o objetivo de determinar o tamanho amostral, no qual são válidas as propriedades assintóticas dos estimadores do modelo de regressão linear simples, tanto no caso de mínimos quadrados como na minimização da soma dos desvios absolutos, a partir da distribuição normal contaminada, Cauchy e Laplace, para o componente aleatório ε_i e consequentemente para a variável resposta Y.

4.3 O Modelo Gama de Distribuição de Probabilidade

4.3.1 A Distribuição Gama

A distribuição Gama é um importante modelo de descrição probabilística para variáveis contínuas que assumem somente valores positivos. Esta distribuição possui como caso particular a distribuição exponencial, utilizada em estudos de confiabilidade, e a distribuição Qui-quadrado, de amplo uso na inferência estatística.

A distribuição Gama é o modelo apropriado para descrever o tempo necessário até a ocorrência de exatamente η eventos independentes, se os eventos ocorrem a uma taxa constante λ. Esta distribuição tem inúmeras aplicações, principalmente em tempo de calibração de instrumentos, teoria dos jogos, teoria Bayesiana, análise de sobrevivência e análise de confiabilidade (Hahn & Shapiro, 1967).

Muitos fenômenos, como por exemplo, o estudo do tempo de falha de capacitores (Hahn & Shapiro, 1967) e o tempo de vida de equipamentos elétricos, entretanto, possuem variáveis aleatórias que não podem ser justificados teoricamente como tendo uma distribuição Gama, apesar de empiricamente resultarem em bons ajustes. Através desta distribuição pode-se conhecer mais sobre o funcionamento dos sistemas biológicos e agrícolas, principalmente os mais complexos, tendo como balizadores algumas distribuições probabilísticas empíricas.

A *função Gama* (Γ) é definida por

$$\Gamma(\eta) = \int_0^\infty x^{\eta-1} e^{-x}\, dx, \qquad \eta > 0$$

e $\Gamma(\eta) = (\eta - 1)!$ em que η é um inteiro positivo.

Seja X uma variável aleatória contínua, com valores não negativos. Diz-se que X tem *distribuição de probabilidade Gama* de parâmetros η e λ, se sua *função densidade de probabilidade* (f.d.p.) é dada por

$$f(x;\eta,\lambda) = \begin{cases} \dfrac{\lambda^{\eta}}{\Gamma(\eta)} x^{\eta-1} e^{-\lambda x}, & x>0,\ \eta>0,\ \lambda>0 \\ 0, & \text{caso contrário} \end{cases} \tag{4.7}$$

Quando λ é fixo e η varia, obtém-se uma grande quantidade de formas; por outro lado, se η é fixo e λ varia, a forma da distribuição não muda, mas apenas sua escala. Por conseguinte, η e λ são parâmetros, respectivamente, de forma e escala (Hahn & Shapiro, 1967).

A *distribuição Gama acumulada* é dada por

$$F(x;\eta,\lambda) = \begin{cases} \dfrac{\lambda^{\eta}}{\Gamma(\eta)} \int_0^x t^{\eta-1} e^{-\lambda t}\, dt, & x \geq 0, \\ 0, & x<0. \end{cases} .$$

Fazendo $y = \lambda t$ e $dt = \frac{1}{\lambda} dy$ tem-se:

$$F(x;\eta,\lambda) = \frac{1}{\Gamma(\eta)} \int_0^x y^{\eta-1} e^{-y}\, dy, \quad x \geq 0$$

que fornece a probabilidade de X assumir valores menores ou iguais a x e, se encontra tabelada para os vários valores de η. A função

$$h(x) = \int_0^x y^{\eta-1} e^{-y}\, dy$$

é conhecida como *função Gama incompleta*.

4.3.2 Propriedades da distribuição Gama

a) Se $\eta = 1$, $f(x;\ \eta,\ \lambda)$ se transforma em $f(x;\ \eta,\ \lambda) = \lambda e^{-\lambda x}$, $x \geq 0$. Portanto, a *distribuição exponencial* é um caso particular da Gama.

b) Quando o parâmetro η é um inteiro positivo a função de distribuição acumulada da distribuição Gama pode ser expressa em termos da função de distribuição acumulada da distribuição de Poisson, como mostra Meyer (1969).

c) Se X tem uma distribuição Gama de parâmetros η e λ, dada por (4.7), tem-se

$$E(X) = \frac{\eta}{\lambda} \qquad \text{e} \qquad V(X) = \frac{\eta}{\lambda^2} \tag{4.8}$$

cuja demonstração pode ser vista em Silveira Júnior et al. (1992).

d) A distribuição Gama se aproxima de uma distribuição normal quando η aumenta (Hahn & Shapiro, 1967).

4.3.3 Assimetria e Curtose

Os momentos são quantidades que auxiliam na descrição de uma distribuição de probabilidade. Os momentos centrados na média de ordem dois, três e quatro (Apêndice 1), são usados para determinação do grau de assimetria e de curtose.

Por *assimetria* deve-se entender o grau de afastamento de uma distribuição do seu eixo de simetria e é utilizada para comparar a simetria de duas distribuições cujas escalas de medidas diferem. Uma distribuição poderá ser assimétrica positiva ($\alpha_3 > 0$) quando há uma freqüência maior de observações à esquerda ou menores do que a média; assimétrica negativa ($\alpha_3 < 0$) quando os valores de X são mais freqüentes à direita da média; ou simétrica ($\alpha_3 = 0$) quando a média separa os valores de X em duas partes iguais, uma metade à direita e a outra à esquerda.

Uma medida objetiva do grau de assimetria de uma distribuição é dada pelo coeficiente de assimetria α_3, ou seja:

$$\alpha_3 = \frac{\mu_3}{(\mu_2)^{3/2}}$$

em que μ_2 e μ_3 são, respectivamente, o segundo e o terceiro momentos centrados na média (Apêndice 1).

A *curtose* pode ser conceituada como o grau de achatamento de uma distribuição. Uma distribuição é chamada de leptocúrtica se $\alpha_4 > 3$ e de platicúrtica se $\alpha_4 < 3$. A distribuição é denominada mesocúrtica quando $\alpha_4 = 3$.

A relação

$$\alpha_4 = \frac{\mu_4}{\mu_2^2}$$

é uma medida relativa de curtose e μ_4 é o quarto momento centrado na média (Apêndice 1).

Considerando η o parâmetro de forma, a assimetria e a curtose da distribuição Gama são dadas, respectivamente por (Hahn & Shapiro, 1967):

$$\alpha_3 = \frac{2}{\sqrt{\eta}}$$

$$\alpha_4 = \frac{3(\eta+2)}{\eta} = \frac{3}{2}\,\alpha_3^2 + 3$$

4.3.4 A Distribuição Gama Padronizada

Conforme Cohen & Whitten (1988), a distribuição Gama padronizada pode ser obtida através da seguinte transformação:

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} \tag{4.9}$$

Assim, a função densidade de probabilidade da distribuição Gama padronizada resultante, que segue uma distribuição Gama (0, 1, α_3) sendo α_3 o parâmetro de forma, é dada por

$$g(z;\,0,\,1,\,\alpha_3) = \begin{cases} \left(\frac{2}{\alpha_3}\right)^{4/\alpha_3^2} \left(z+\frac{2}{\alpha_3}\right)^{4/\alpha_3^2-1} \exp\left[\frac{-2}{\alpha_3}\left(z+\frac{2}{\alpha_3}\right)\right], & \frac{-2}{\alpha_3} < z < \infty \\ 0 & \text{caso contrário} \end{cases}$$

Esta distribuição Gama padronizada é a distribuição utilizada neste trabalho.

5 MATERIAL E MÉTODOS

5.1 Material

Os procedimentos utilizados para o desenvolvimento desse trabalho foram os computacionais.

Tais procedimentos foram realizados utilizando o programa *Statistical Analysis System*, SAS, (SAS, 1996), em um computador Pentium III 500 MHz, com 128 MB de memória RAM e 20 GB de disco rígido, junto ao Departamento de Bioestatística do Instituto de Biociências da UNESP – Campus de Botucatu, para a realização das simulações necessárias e a utilização do método dos mínimos quadrados.

Foi utilizado também, um computador Celeron Intel 400 MHz, 64 MB de memória RAM e 20 GB de disco rígido, de uso particular, para executar o método de regressão L_1, implementado por Silva (1994), através do compilador *Borland Pascal V.7*.

5.2 Metodologia Utilizada

5.2.1 Simulação dos dados

A capital de Mônaco, Monte Carlo, é conhecida por seus empreendimentos arriscados onde as pessoas jogam com sua sorte na esperança de ganhar fortunas. O resultado desses jogos é determinado por fatores aleatórios. Analogamente, a técnica conhecida como simulação de Monte Carlo é baseada em um procedimento que produz um resultado fundamentado em jogos cuja conseqüência depende de um número de fatores aleatórios. Em um estudo de simulação, o "jogo" é uma representação matemática ou funcional de um sistema e, os fatores aleatórios são as variáveis aleatórias que são usadas para representar os elementos do sistema. Um estudo de simulação geralmente envolve os seguintes passos (Shapiro & Gross, 1981):

1. determine o modelo matemático e/ou funcional que representa o sistema de entrada;
2. associe a cada variável de entrada uma distribuição de probabilidade;
3. gere valores aleatórios para cada variável de entrada e utilize-os no modelo funcional para calcular os valores de saída;
4. repita os passos anteriores muitas vezes para obter um conjunto de dados que represente o sistema de entrada;
5. resuma o conjunto de dados ajustando uma distribuição empírica e calculando os momentos e percentis desejados.

Este trabalho utilizou-se desta técnica de Monte Carlo para gerar os conjuntos de dados analisados.

Inicialmente assumiu-se um modelo de regressão linear simples representado por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (5.1)$$

com $i = 1, 2, ..., n$, sendo n o tamanho da amostra.

Neste modelo, β_0 e β_1 são parâmetros da população conhecidos, e os erros (componentes aleatórios) ε_i seguem uma distribuição Gama padronizada com parâmetro α_3 (assimetria), média zero e desvio padrão 1, isto é, $\varepsilon_i \sim$ Gama $(0, 1, \alpha_3)$.

Como mencionado anteriormente, a função densidade de probabilidade da distribuição Gama padronizada é dada por:

$$g(z; 0, 1, \alpha_3) = \left(\frac{2}{\alpha_3} \right)^{4/\alpha_3^2} \left(z + \frac{2}{\alpha_3} \right)^{4/\alpha_3^2 - 1} \exp\left[\frac{-2}{\alpha_3} \left(z + \frac{2}{\alpha_3} \right) \right], \quad \frac{-2}{\alpha_3} < z < \infty$$

e, através das expressões (4.8) e (4.9) tem-se

$$z = \frac{\lambda X - \eta}{\sqrt{\eta}} \tag{5.2}$$

sendo η um parâmetro de forma e λ o parâmetro de escala.

Devido a relação existente entre a assimetria e o parâmetro de forma η, quando $\eta = 1$ tem-se $\alpha_3 = 2$ e, portanto, $\alpha_3^2 = 4$, obtendo-se o modelo exponencial; aumentando-se o valor de η a distribuição Gama se aproxima da normal ($\alpha_3^2 \cong 0$).

O Quadro 1 mostra os valores de α_3 para alguns valores de η em particular, os quais foram utilizados nas simulações.

Quadro 1 - Valores obtidos para a assimetria α_3 através da relação $\alpha_3 = 2/\sqrt{\eta}$ variando η de 0,25 a 9,75 com incremento de 0,50.

η	α_3	η	α_3
0,25	4,0000	5,25	0,8729
0,75	2,3094	5,75	0,8341
1,25	1,7889	6,25	0,8000
1,75	1,5119	6,75	0,7698
2,25	1,3333	7,25	0,7428
2,75	1,2060	7,75	0,7184
3,25	1,1094	8,25	0,6963
3,75	1,0328	8,75	0,6761
4,25	0,9701	9,25	0,6576
4,75	0,9177	9,75	0,6405

Considerando o modelo (5.1) e a relação entre α_3 e η, foram gerados vários conjuntos de observações através de simulações realizadas no SAS, utilizando-se da função RANGAM para gerar a distribuição Gama (Apêndice 2) e da expressão (5.2) para obter a distribuição Gama padronizada.

Tais conjuntos foram obtidos variando-se o tamanho da amostra de 20 a 100 com incremento de 20; λ variou de 0,2 a 2,2 com incremento de 0,4 e; η variou de 0,25 a 9,75 com incremento de 0,50. Além disso, com base nos valores obtidos na aplicação, considerou-se β_0 valendo 20, 30 e 40; bem como β_1 valendo –0,5, -0,4 e –0,3. Para cada combinação, β_0, β_1, tamanho de amostra, λ e η, com 1000 repetições cada, o modelo (5.1) foi gerado utilizando a distribuição Gama padronizada.

5.2.2 Estimação dos parâmetros

Na etapa seguinte determinou-se as estimativas dos parâmetros do modelo gerado anteriormente por dois métodos:

a) método dos mínimos quadrados, que é o método clássico utilizado em regressões lineares;

b) método desenvolvido por Silva (1994), que é uma especialização do método primal simplex da programação linear para o problema de regressão L_1 e explora o fato da função objetivo ser linear por partes (Apêndice 3).

5.2.3 Comparação dos estimadores

Concluída a etapa anterior, procedeu-se à comparação entre os estimadores obtidos por ambos os métodos, bem como as análises da eficiência do método de regressão L_1 em relação ao método dos mínimos quadrados.

Deve-se lembrar que, dado um vetor de parâmetros $q = (\beta_0, \beta_1)^t$ o erro quadrático médio (EQM) do estimador $\hat{q}$, considerando a distribuição Euclidiana, é dado por:

$$EQM(\hat{q}) = E\left\| q - \hat{q} \right\|^2 = \sqrt{\left(\beta_0 - \hat{\beta}_0\right)^2 + \left(\beta_1 - \hat{\beta}_1\right)^2}$$

Além disso, sabe-se da inferência estatística que, dados dois estimadores $\hat{q}_1$ e $\hat{q}_2$ de q, se $EQM(\hat{q}_1) < EQM(\hat{q}_2)$, então $\hat{q}_1$ é melhor que $\hat{q}_2$ (Mood et al., 1974).

A partir do exposto, definiu-se R_{qmq} como sendo a razão entre os erros quadráticos médios dos estimadores da regressão L_1 e os erros quadráticos médios dos estimadores de mínimos quadrados, ou seja:

$$R_{qmq} = \frac{EQM(\hat{\theta}_q)}{EQM(\hat{\theta}_{mq})} \qquad (5.3)$$

sendo $\hat{\theta}_q$ o vetor dos estimadores da regressão L_1 e $\hat{\theta}_{mq}$ o vetor dos estimadores de mínimos quadrados. Deve-se lembrar que EQM é sempre maior que zero e, assim esta razão pode ser criada.

A razão (5.3) foi utilizada como critério de comparação entre os dois métodos. Quando esta razão é menor que 1 temos que a regressão L_1 é melhor que o método dos mínimos quadrados e, portanto, é esta relação que nos interessa. As comparações foram efetuadas utilizando as seguintes etapas:

a) efeito do parâmetro de escala do modelo Gama na REQM
b) efeito dos coeficientes linear e angular do modelo de regressão na REQM;
c) efeito do tamanho amostral e do coeficiente de assimetria na REQM.

Estas mesmas etapas foram utilizadas para comparar os dois métodos utilizando a diferença entre os erros quadráticos médios.

O uso da razão dos erros quadráticos médios (5.3) para realizar as comparações entre os dois métodos considerados, embora simples, não é uma metodologia encontrada na literatura. Desta forma, este trabalho apresenta uma metodologia nova, simples e eficiente.

A comparação entre os dois métodos foi realizada também, utilizando-se as variâncias residuais médias desses métodos, visto que é este tipo de comparação que encontra-se na maioria das pesquisas (Narula & Wellington, s.n.t).

5.3 Material e Método Utilizado na Aplicação

Guiscem (2001) analisou a qualidade fisiológica das sementes de milho doce em função do teor de água na colheita e da temperatura de secagem em espiga, utilizando 3 cultivares de milho com 11 tratamentos, compostos pela combinação colheita e secagem.

A importância do estudo de Guiscem (2001) deve-se ao fato de que um dos fatores que mais afeta a qualidade do fruto está relacionado com o seu alto teor de água, assim

a secagem da planta pode melhorar a qualidade do produto. A secagem natural em campo na própria planta pode acarretar perdas físicas e qualitativas potencialmente prejudiciais. Por outro lado, a secagem artificial requer uma maior produção de milho para cobrir seu alto custo, devido ao grande gasto de energia para a secagem. Assim, o estudo do teor de umidade da planta a fim de se obter a menor taxa é de fundamental importância tanto na agronomia quanto em energia na agricultura.

Guiscem (2001) não realizou o ajuste do modelo de regressão linear em seus conjuntos de dados, não comprometendo assim suas conclusões. Se os conjuntos de dados satisfizessem todas as pressuposições do modelo de regressão linear ele poderia ser utilizado, pois o parâmetro estimado $\hat{\beta}_1$ representa a perda média do teor de umidade da planta. Por se adequarem aos nossos propósitos, este trabalho utiliza três dos conjuntos de dados de Guiscem (2001) como exemplo prático.

Assim, o método de regressão L_1 de Silva (1994) e o método dos mínimos quadrados foram aplicados nestes conjuntos de dados e os seus resultados analisados.

6 RESULTADOS E DISCUSSÃO

Neste capítulo apresentaremos os resultados e discussão do trabalho em forma de seções para uma melhor compreensão.

6.1 Estimadores de regressão L_1

Os estimadores de β_0 e β_1 da regressão L_1 foram obtidos, como já mencionado, pela aplicação do método de Silva (1994).

As simulações realizadas, considerando as variações do tamanho amostral, λ, η, β_0, β_1 e 1.000 repetições para cada caso, geraram 5.400.000 conjuntos de dados. Desse valor, ao aplicar regressão L_1, 1.567 conjuntos (0,0290%) não convergiram em 1.000 iterações. O problema de convergência ocorreu devido ao fato de Silva (1994) em seu programa não ter considerado a possibilidade de ocorrer base degenerada, isto é, no método simplex da programação linear, no qual o método de regressão L_1 de Silva (1994) se baseia, existem situações em que a solução encontrada se alterna entre dois valores e necessita de um tratamento especial (simplex lexicográfico) para escolher a melhor solução e atingir a convergência.

Esses 1.567 conjuntos de observações não foram considerados neste estudo.

6.2 Comparação dos Estimadores

As análises dos resultados obtidos estão apresentadas segundo três diferentes linhas. A primeira tem como base a razão dos erros quadráticos médios dos métodos estudados, a segunda é baseada na diferença dos erros quadráticos médios e, a terceira utiliza a comparação das variâncias residuais dos modelos ajustados pelos dois métodos. Esta última linha é encontrada na maior parte dos artigos científicos(Narula & Wellington, s.n.t).

6.2.1 Análise comparativa dos métodos de mínimos quadrados e mínimos desvios absolutos com base no REQM – Razão dos Erros Quadráticos Médios.

6.2.1.1 Efeito do parâmetro de escala (l) do modelo Gama na REQM

De todas as variações incluídas nas simulações: tamanho amostral, coeficiente linear e angular do modelo de regressão linear, parâmetro de escala e de forma do modelo Gama e o coeficiente de assimetria; o parâmetro de escala λ foi o único a afetar as REQM´s de duas maneiras opostas. Valores de λ inferiores a 1,0 produzem REQM maiores que 1,0 - favorecendo o método de mínimos quadrados; já valores de λ superiores a 1,0 o efeito é o oposto, isto é, produzem REQM menores que 1,0 - favorecendo assim o método de mínimos desvios absolutos (Figura 1).

Observa-se que cada grupo de λ (≤ 1 e >1), contém 50% das observações. Assim, nas situações estudadas (n=5.400) 50% dos casos sugerem o uso do método dos mínimos desvios absolutos (regressão L_1) e os outros 50% o método dos mínimos quadrados.

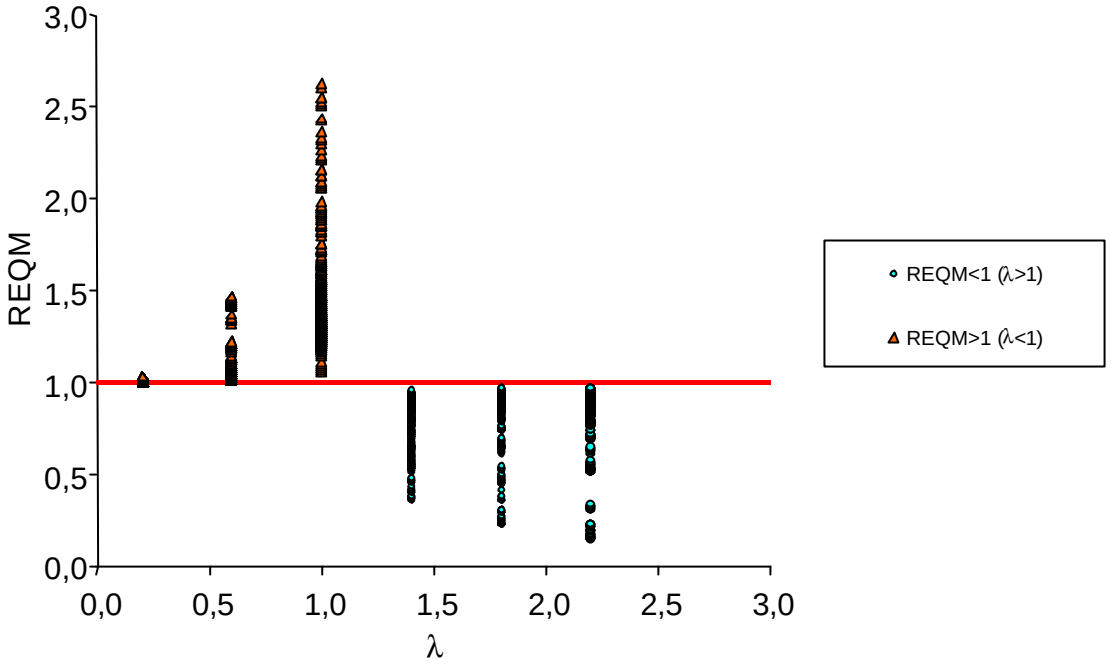


Figura 1 - Efeito do parâmetro de escala (λ) do modelo Gama de distribuição de probabilidade na Razão dos Erros Quadráticos Médios (REQM), obtidos por simulação (n=5400), considerando tamanho amostral, coeficiente linear e angular do modelo de regressão linear, parâmetro de escala e de forma do modelo Gama e o coeficiente de assimetria (1000 repetições para cada combinação).

6.2.1.2 Efeito das variações incluídas na simulação na REQM

Pela Figura 2, pode-se verificar que a variação de β_0 não influencia na razão dos erros quadráticos médios, tanto para $\lambda<1$ (caso em que o método dos mínimos quadrados é melhor), como para $\lambda>1$ (situação em que o método dos mínimos desvios absolutos é melhor). Observa-se também que, para $\lambda<1$, independente dos valores de β_0, há uma maior concentração de valores da razão dos erros quadráticos médios em sua parte inferior. Fato semelhante pode ser observando para $\lambda>1$, porém esta concentração é maior para valores da razão mais próximos de 1.

Resultado análogo pode ser observado para a variação de β_1, tanto em relação a λ quanto a concentração dos valores da razão dos erros quadráticos médios.

Assim, pode-se concluir que a variação dos coeficientes linear (β_0) e angular (β_1) do modelo de regressão não afetam a razão dos erros quadráticos médios para os casos analisados.

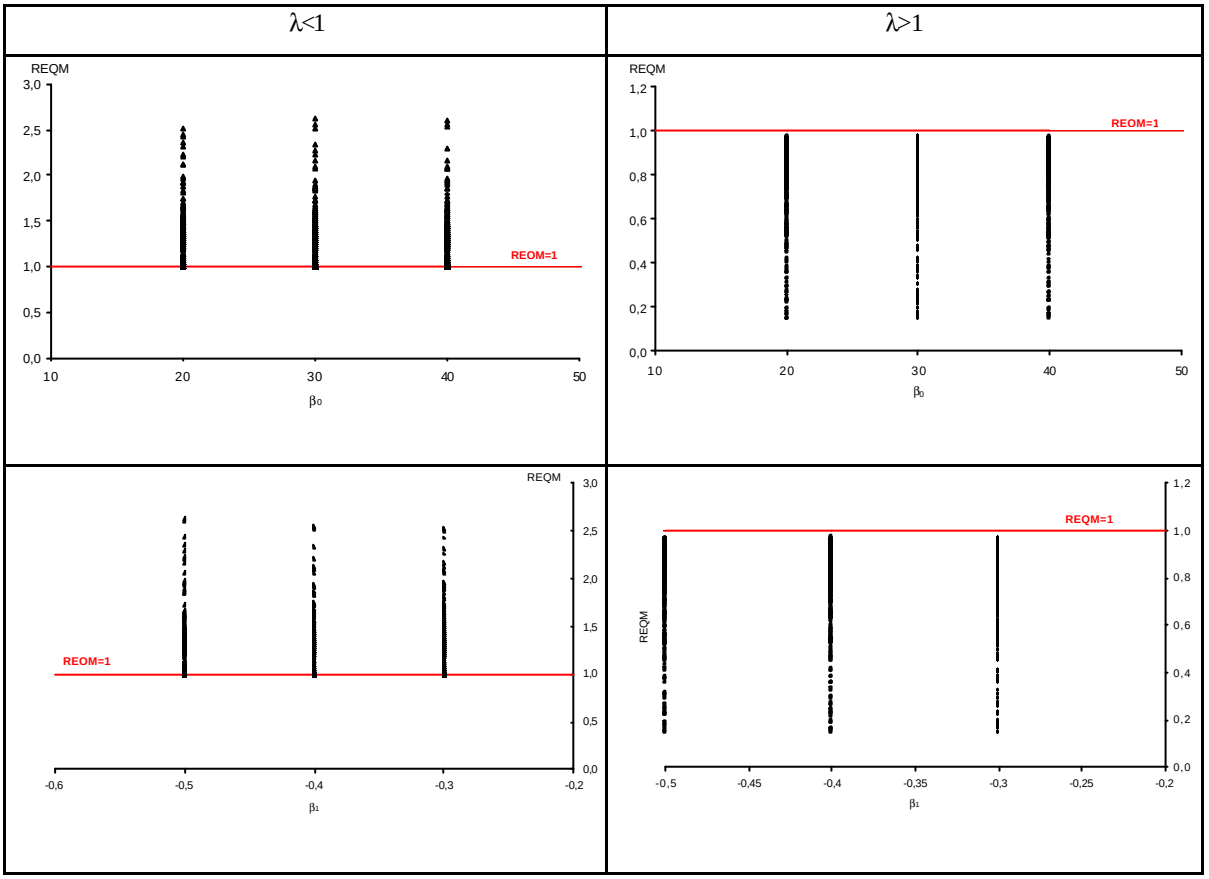


Figura 2 – Efeito dos coeficientes linear e angular do modelo de regressão (β_0) e (β_1), respectivamente, na Razão dos Erros Quadráticos Médios (REQM), obtidos por simulação (1000 repetições), segundo valores de do parâmetro de escala λ do modelo Gama.

Analisando o tamanho amostral e considerando λ<1, observa-se pela Figura 3 que, quanto maior o tamanho da amostra, maior o valor da razão dos erros quadráticos médios e, a concentração destes valores tende a diminuir.

Considerando $\lambda>1$, nota-se que o tamanho amostral parece não influenciar nos valores de REQM. Observa-se também que, a concentração dos valores da razão dos erros quadráticos médios tende a ser maior para os valores de REQM mais próximos de 1.

Pela Figura 3, pode-se observar o efeito que o coeficiente de assimetria (α_3) exerce sobre a razão dos erros quadráticos médios. À medida que o coeficiente de assimetria aumenta, a razão também aumenta, quando $\lambda<1$ e, o oposto ocorre no caso de $\lambda>1$, isto é, nesta situação quanto maior α_3 mais próximo de zero está a razão dos erros quadráticos médios.

Assim, conclui-se que a variação do tamanho amostral afeta a razão dos erros quadráticos médios apenas no caso de $\lambda<1$, isto é, no caso em que o método dos mínimos quadrados é melhor. Com relação ao coeficiente de assimetria, conclui-se que, para $\lambda<1$, quanto maior o coeficiente de assimetria, melhor é o método dos mínimos quadrados e, para $\lambda>1$, aumentando o valor de α_3 o método dos mínimos desvios absolutos torna-se mais eficiente.

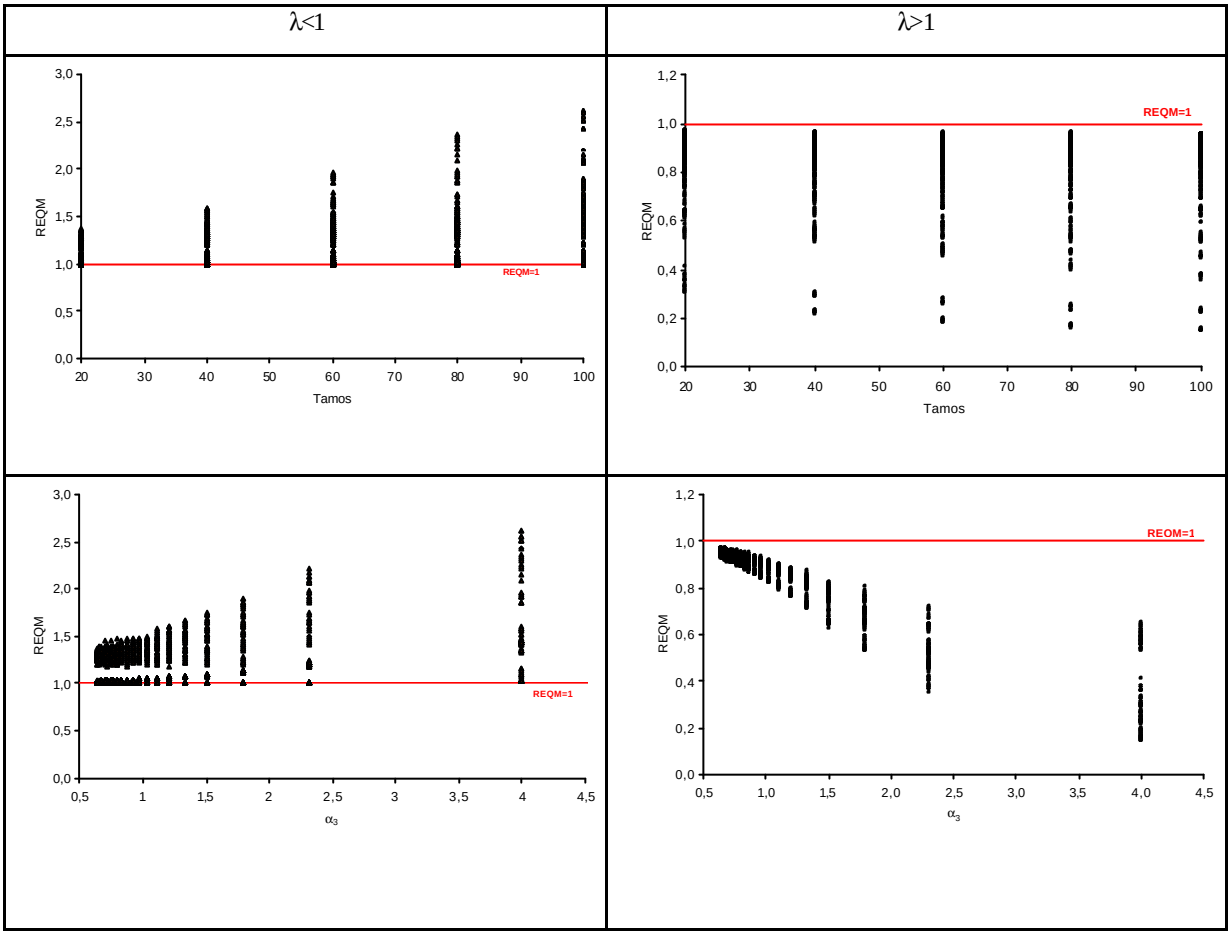


Figura 3 – Efeito do tamanho amostral e do coeficiente de assimetria na Razão dos Erros Quadráticos Médios (REQM), obtidos por simulação (1000 repetições), segundo valores do parâmetro de escala λ do modelo Gama.

6.2.2 Análise comparativa dos métodos de mínimos quadrados e mínimos desvios absolutos com base na DEQM – Diferença dos Erros Quadráticos Médios.

6.2.2.1 Efeito do parâmetro de escala (l) do modelo Gama na DEQM

Considerando as diferenças dos erros quadráticos médios

$$DEQM = EQM\_Q - EQM\_MQ$$

sendo EQM_Q o erro quadrático médio da regressão L_1 e EQM_MQ o erro quadrático médio do método de mínimos quadrados, tem-se que se DEQM<0 o método da regressão L_1 é melhor que o método de mínimos quadrados, conseqüentemente, o método de mínimos quadrados é melhor caso DEQM>0.

Assim como no caso da razão dos erros quadráticos médios, considerando todas as variações incluídas nas simulações apenas o parâmetro de escala λ afeta as DEQM´s. Valores de λ inferiores a 1,0 produzem DEQM maiores que zero - favorecendo o método de mínimos quadrados; já valores de λ superiores a 1,0 o efeito é o oposto, isto é, produzem DEQM menores que zero - favorecendo assim o método de regressão L_1 (Figura 4).

Observa-se pela Figura 4 que para o caso em que $\lambda>1$, situação em que a regressão L_1 é melhor, a concentração dos valores das diferenças é maior e mais próxima de zero quanto mais próximo de 1 estiver λ, conseqüentemente, a medida que λ aumenta, a diferença dos erros quadráticos médios também aumenta em módulo, isto é, melhor é o método da regressão L_1.

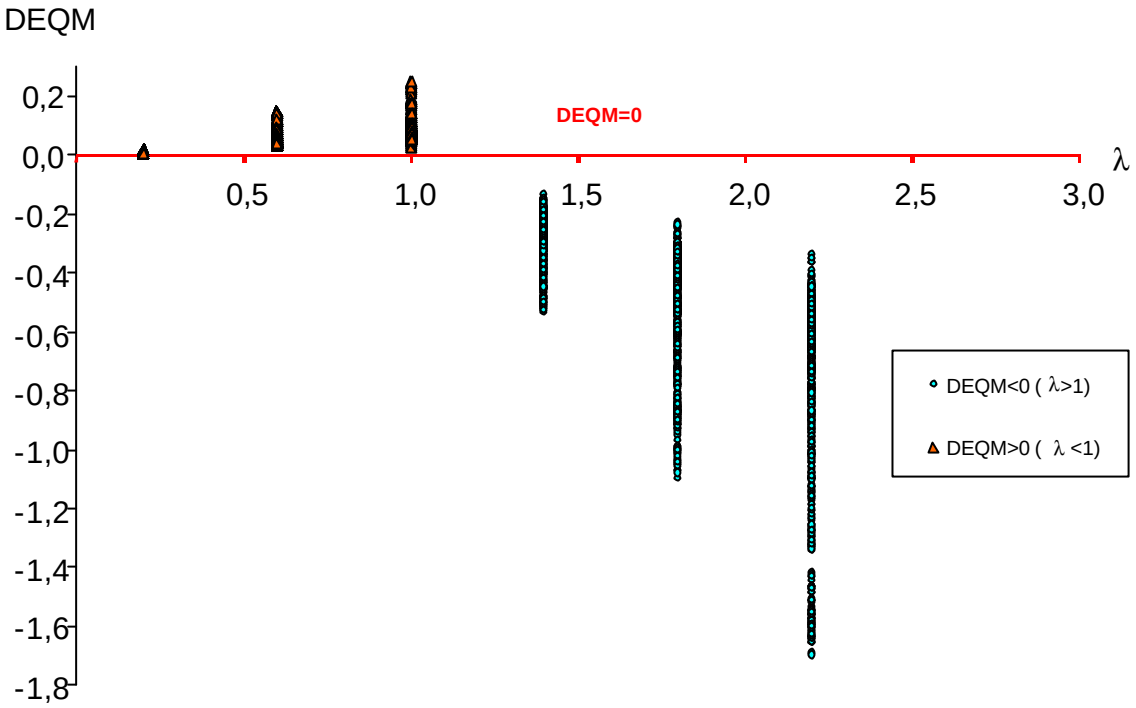


Figura 4 - Efeito do parâmetro de escala (λ) do modelo Gama de distribuição de probabilidade na Diferença dos Erros Quadráticos Médios (DEQM), obtidos por simulação (n=5400), considerando tamanho amostral, coeficiente linear e angular do modelo de regressão linear, parâmetro de escala e de forma do modelo Gama e o coeficiente de assimetria (1000 repetições para cada combinação).

6.2.2.2 Efeito das variações incluídas na simulação na DEQM

Pela Figura 5, observa-se que, tanto para o coeficiente linear (β_0) quanto para o coeficiente angular (β_1) do modelo de regressão, nas duas situações de λ, à medida que os valores de DEQM se afastam do eixo de referência (DEQM=0), a concentração destes valores diminui, com uma diminuição mais significativa no caso em que $\lambda<1$, isto é, caso onde o método de mínimos quadrados é melhor.

Além disso, como na razão dos erros quadráticos médios (REQM), observa-se que a variação dos coeficientes linear (β_0) e angular (β_1) do modelo de regressão não afetam a diferença dos erros quadráticos médios para os casos $\lambda<1$ e $\lambda>1$.

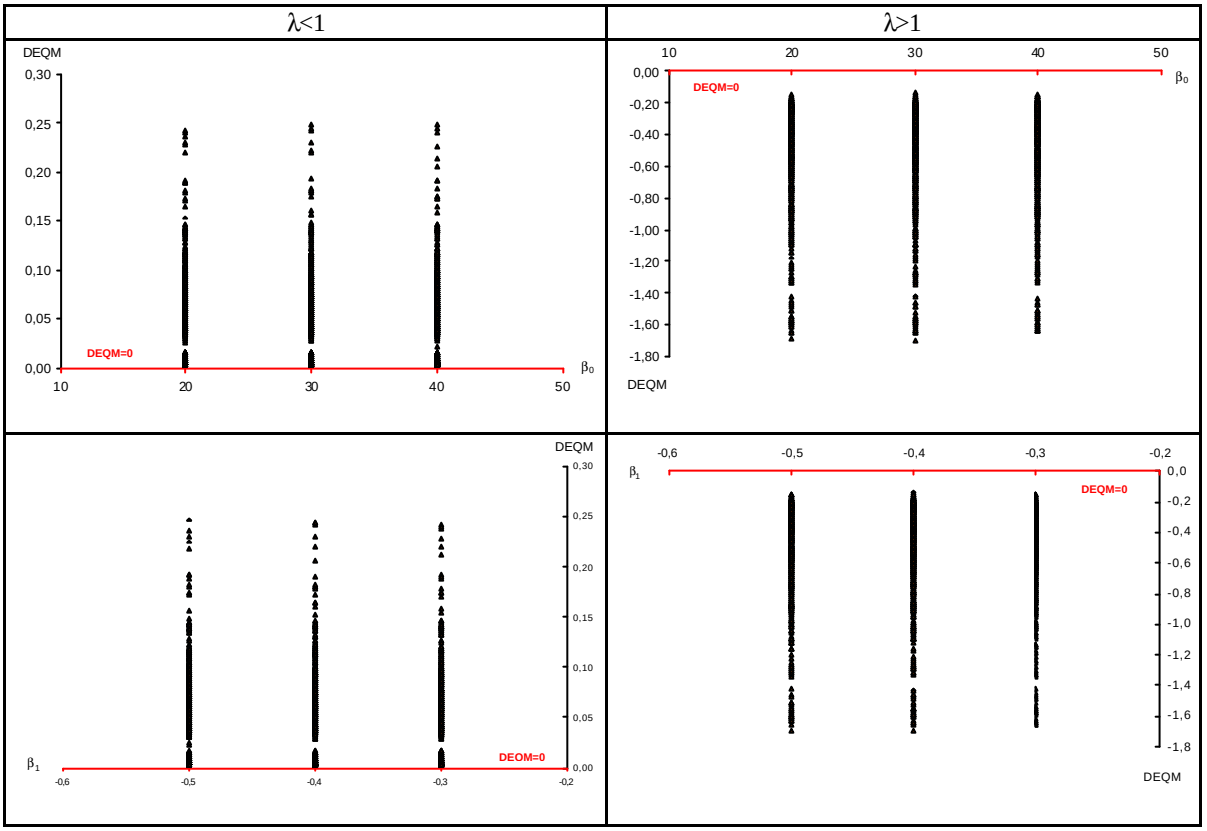


Figura 5 – Efeito dos coeficientes linear e angular do modelo de regressão (β_0) e (β_1) respectivamente, na Diferença dos Erros Quadráticos Médios (DEQM), obtidos por simulação (1000 repetições), segundo valores de do parâmetro de escala λ do modelo Gama.

Analisando o tamanho amostral e considerando $\lambda<1$, observa-se pela Figura 6 que, quanto maior o tamanho da amostra, maior o valor da diferença dos erros quadráticos médios, favorecendo o método de mínimos quadrados.

Considerando $\lambda>1$, nota-se que o tamanho amostral parece não influenciar nos valores de DEQM.

Pela Figura 6, pode-se observar também o efeito que o coeficiente de assimetria (α_3) exerce sobre a diferença dos erros quadráticos médios. Para $\lambda<1$, à medida que o coeficiente de assimetria aumenta, também aumenta a diferença DEQM, favorecendo o método de mínimos quadrados. Quando $\lambda>1$, quanto maior α_3 melhor é o método de mínimos desvios absolutos, pois a diferença se distancia cada vez mais do eixo referencial (DEQM=0).

Assim, conclui-se que a variação do tamanho amostral afeta a diferença dos erros quadráticos médios apenas no caso de $\lambda<1$, isto é, no caso em que o método de mínimos quadrados é melhor. Com relação ao coeficiente de assimetria, conclui-se que, para $\lambda<1$, quanto maior o coeficiente de assimetria, melhor é o método de mínimos quadrados e, para $\lambda>1$, o método favorecido é regressão L_1.

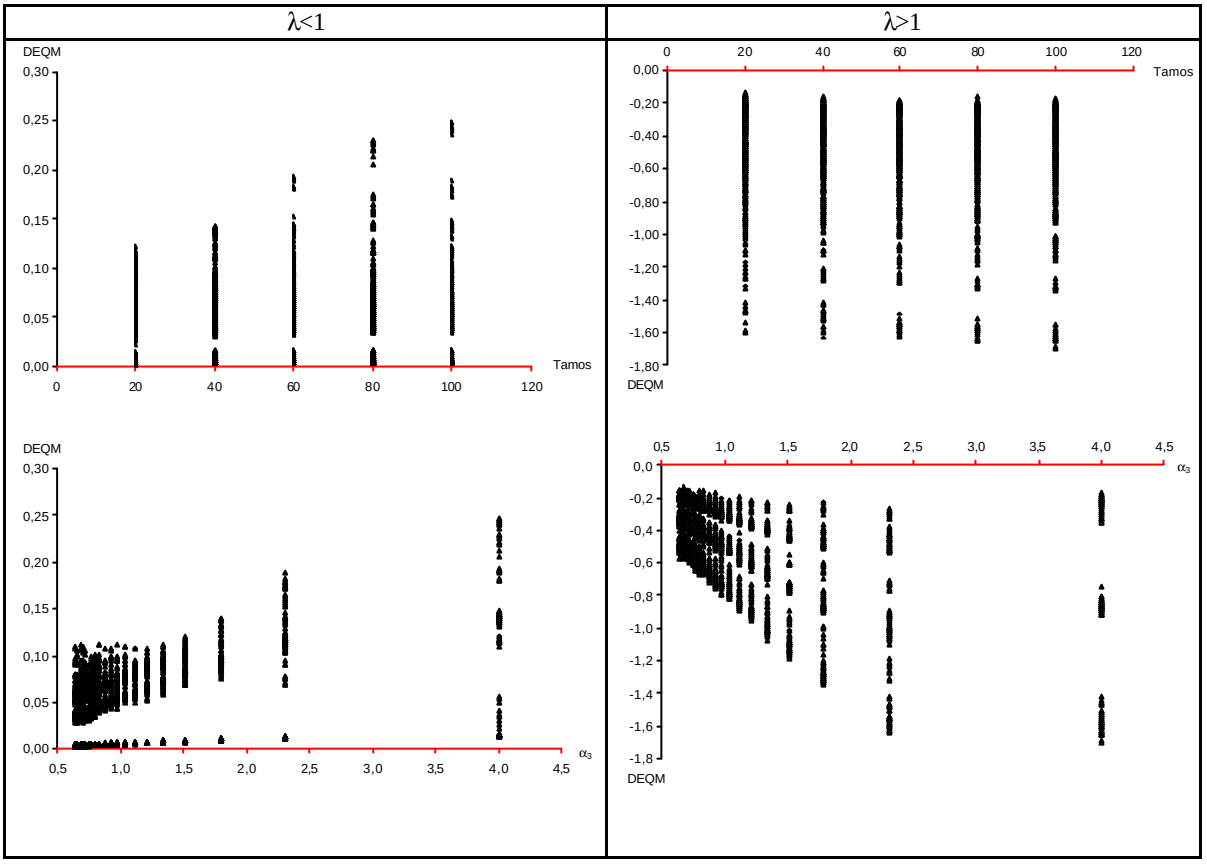


Figura 6 – Efeito do tamanho amostral e do coeficiente de assimetria na Diferença dos Erros Quadráticos Médios (DEQM), obtidos por simulação (1000 repetições), segundo valores do parâmetro de escala λ do modelo Gama.

6.2.3 Análise comparativa dos métodos de mínimos quadrados e mínimos desvios absolutos com base em suas variância residuais médias.

A Figura 7 mostra o comportamento dos dois métodos utilizados neste trabalho considerando suas variâncias residuais médias.

Com base nesta figura, pode-se constatar que, a variância residual média do método dos mínimos desvios absolutos é sempre menor que a dos mínimos quadrados, pois os dados apresentam-se abaixo da reta de referencia (VRES_MQ = VRES_MDA), com apenas alguns pontos sobre a reta.

Nota-se que, à medida que λ aumenta, a variância residual média dos dois métodos também aumenta, porém a variância residual média referente ao método de mínimos quadrados aumenta mais que a dos mínimos desvios absolutos, pois os dados se distanciam dada vez mais do eixo de referencia.

Assim, conclui-se que, com relação a variância residual média, quanto maior o valor de λ, mais eficiente é o método de mínimos desvios absolutos e, além disso, este método é sempre melhor ou igual ao método de mínimos quadrados, conforme consta na literatura existente.

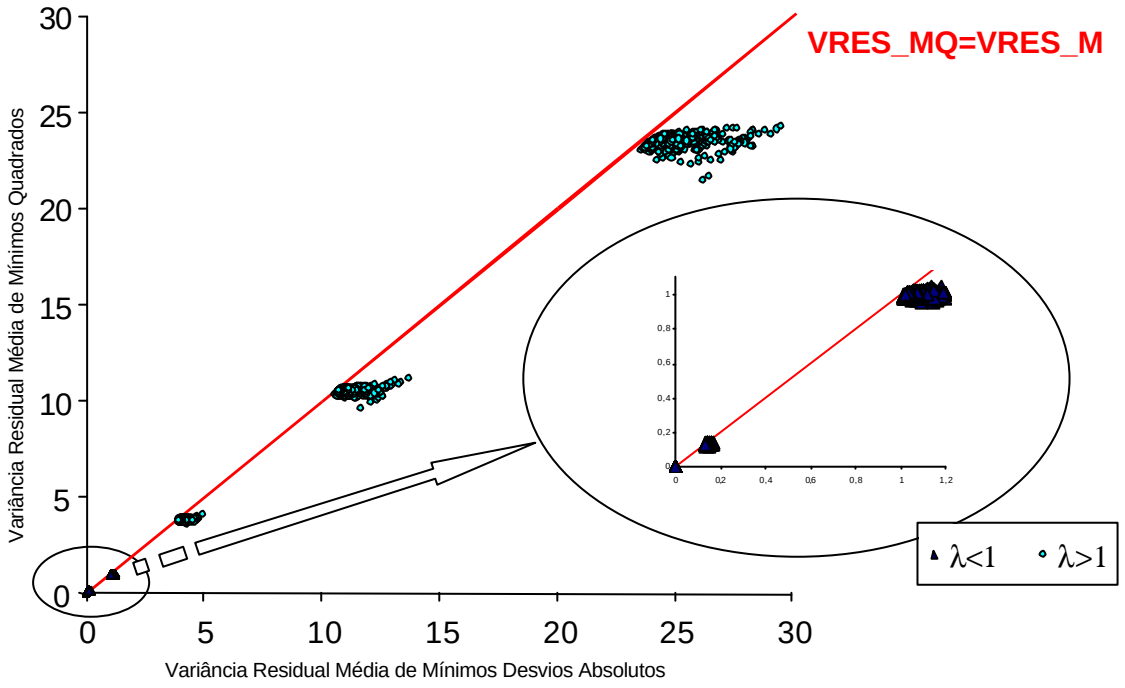


Figura 7 – Variâncias residuais médias dos ajustes dos modelos de regressão linear simples utilizando os métodos de mínimos quadrados e mínimos desvios absolutos, obtidos por simulação (n=5400), considerando tamanho amostral, coeficiente linear e angular do modelo de regressão linear, parâmetro de escala e de forma do modelo Gama e o coeficiente de assimetria (1000 repetições para cada combinação).

6.3 Aplicação

Nesta seção, será realizada a análise dos resultados obtidos para o conjunto de observações referentes à aplicação do estudo realizado neste trabalho.

De acordo com a seção 5.3, os conjuntos de observações utilizados como aplicação foram obtidos de Guiscem (2001). Naquele trabalho analisou-se a qualidade fisiológica das sementes de milho doce em função do teor de água na colheita e da temperatura de secagem em espigas, utilizando três cultivares de milho com 11 tratamentos, compostos pela combinação colheita e secagem.

Analisando esses dados, verificou-se que os conjuntos de observações referentes ao cultivar1-safra2, cultivar7-safra1 e cultivar7-safra3 são os que melhor se ajustam a modelos de regressão linear simples e as distribuições dos seus erros fogem da normal (conforme Figura 12 do Apêndice 4). Assim, esses conjuntos foram os escolhidos descartando-se os demais.

Seguindo os passos realizados na teoria (simulação de Monte Carlo), realizou-se o ajuste para esses dados através do método de mínimos quadrados e do método de regressão L_1.

O Quadro 2 mostra a análise de variância referente ao ajuste do modelo de regressão linear simples aos conjuntos de observações considerados como aplicação. Neste quadro, pode-se verificar que o erro quadrático médio referente a cada conjunto de observações (4,1054; 1,7217 e 5,0643) é consideravelmente alto, concluindo-se que se a estimativa dos parâmetros fosse obtida por mínimos quadrados, estes estimadores não seriam de boa qualidade.

Quadro 2 – Análise de variância referente ao ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, segundo cultivar safra (Guiscem, 2001).

Cultivar	Safra	Causa de variação	Graus de Liberdade	Soma de Quadrados	Quadrado médio	Valor p F	R^2 ajustado
1	2	Regressão	1	2635,4352	2635,4352	0,0001	0,9092
		Resíduo	63	258,6419	4,1054		
		Total	64	2894,0771			
7	1	Regressão	1	2494,7167	2494,7167	0,0001	0,9514
		Resíduo	73	125,6860	1,7217		
		Total	74	2620,4027			
7	3	Regressão	1	1516,6706	1516,6706	0,0001	0,8640
		Resíduo	46	232,9590	5,0643		
		Total	47	1749,6296			

A fim de melhor visualizar o efeito do ajuste do modelo de regressão linear simples aos conjuntos de observações em estudo, considere as Figuras 8, 9 e 10 que correspondem, respectivamente, ao ajuste realizado para os conjuntos cultivar1-safra2, cultivar7-safra1 e cultivar7-safra3. Nestas figuras pode-se ver que o ajuste por regressão linear simples não é um bom ajuste, pois os pontos que representam os teores de umidade do milho observados estão relativamente distantes da reta que representa o teor esperado, mostrando visualmente que o quadrado médio dos resíduos é consideravelmente grande. Esta situação é ainda mais evidente na Figura 10, que corresponde ao cultivar7-safra3, cujo quadrado médio dos resíduos foi 5,0643 como mostra o Quadro 2.

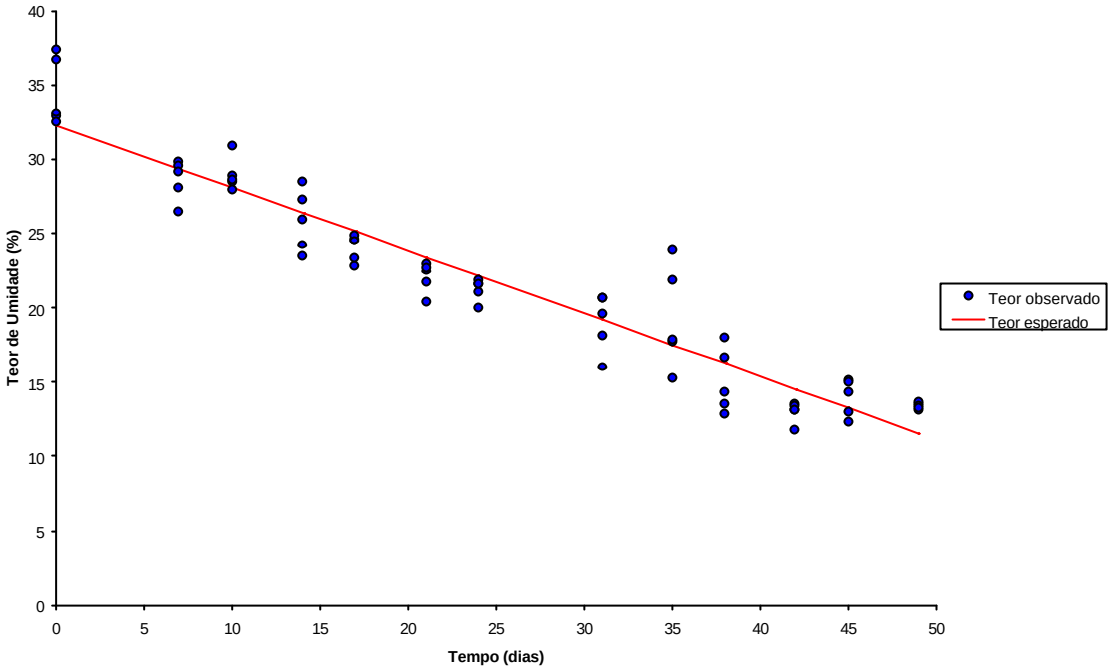


Figura 8 – Ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, cultivar1-safra2 (Guiscem, 2001).

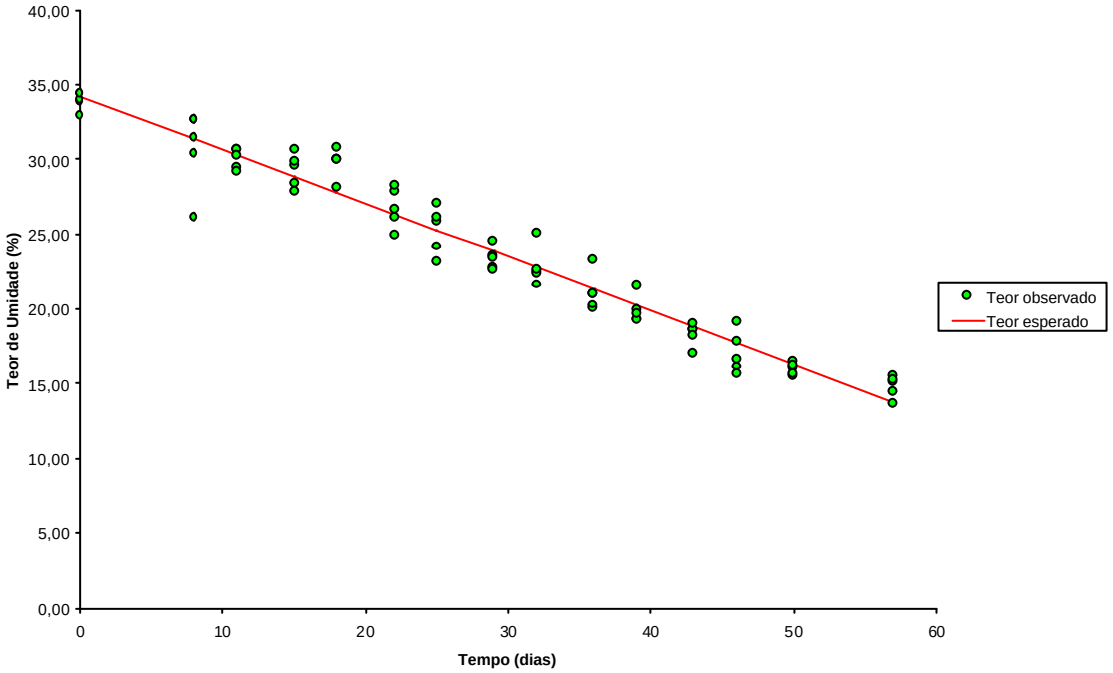


Figura 9 – Ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, cultivar7-safra1 (Guiscem, 2001).

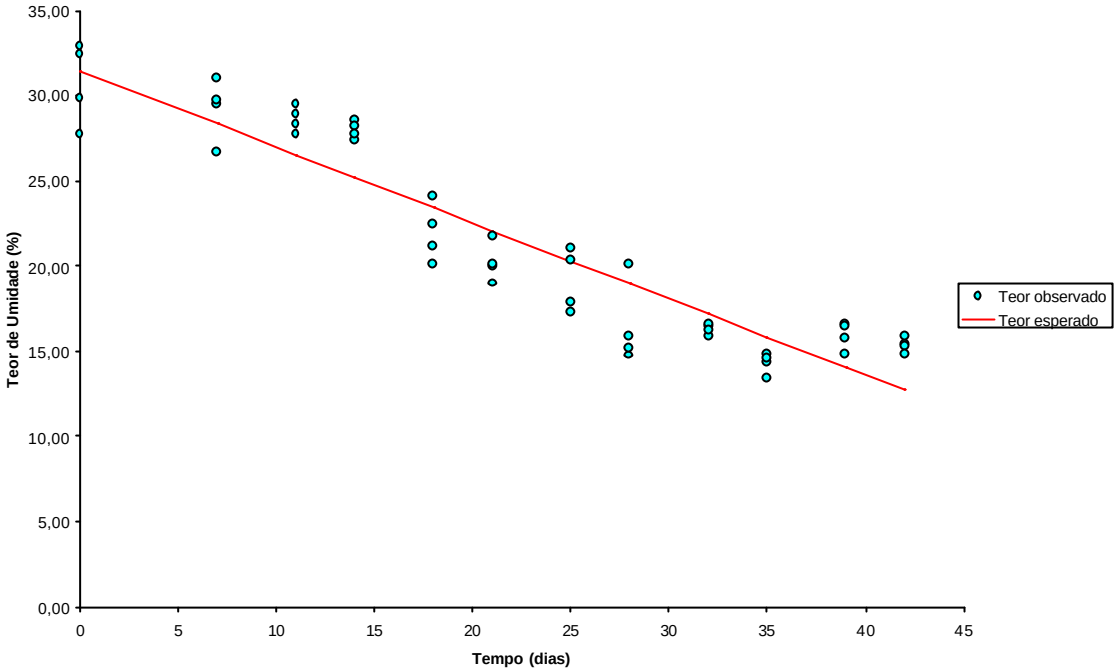


Figura 10 – Ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade em milho, cultivar7-safra3 (Guiscem, 2001).

O Quadro 3 mostra as estimativas dos parâmetros obtidas pela regressão L_1 e por mínimos quadrados. O conjunto de dados cultivar7-safra1 não convergiu em 1000 iterações e o resultado apresentado é o obtido para esse número de iterações, portanto, não é o melhor.

Quadro 3 – Estimativa dos parâmetros referente ao ajuste do modelo de regressão linear simples ao conjunto de dados de teor de umidade do milho, segundo cultivar safra (Guiscem, 2001) obtidos por regressão L_1 ($\hat{\beta}_{0q}$, $\hat{\beta}_{1q}$) e por mínimos quadrados ($\hat{\beta}_{0mq}$, $\hat{\beta}_{1mq}$).

Cultivar	Safra	Estimativa de parâmetros			
		$\hat{\beta}_{0q}$	$\hat{\beta}_{1q}$	$\hat{\beta}_{0mq}$	$\hat{\beta}_{1mq}$
1	2	32,1386	-0,4279	32,2704	-0,4234
7	1	34,3571	-0,3671	34,2791	-0,3598
7	3	32,8500	-0,5084	31,4381	-0,4457

O Quadro 4 mostra a razão entre o erro quadrático médio da regressão L_1 e o erro quadrático médio de mínimos quadrados. Neste quadro, pode-se observar que para os dois conjuntos de dados a razão (R_{qmq}) é menor que 1, mostrando que nestes casos a regressão L_1 é melhor que a regressão de mínimos quadrados. Nota-se também, que o conjunto de dados cultivat7-safra1 não foi considerado, pois o método não convergiu para esses dados e, assim a razão (R_{qmq}) referente a eles pode estar comprometida.

Quadro 4 – Razão entre o erro quadrático médio da regressão L_1 e o erro quadrático médio dos mínimos quadrados para o conjunto de dados de teor de umidade do milho, segundo cultivar safra (Guiscem, 2001).

Cultivar	Safra	R_{qmq}
1	2	0,8913
7	3	0,0837

A Figura 11 mostra um sistema de classificação de distribuições baseado nos coeficientes de assimetria e curtose representado em um sistema de eixos cartesianos $(\alpha_3)^2\alpha_4$, sendo possível identificar vários modelos probabilísticos, tais como Normal, Gama, Weibull, Inversa Gaussiana e Exponencial. Nela pode-se ver que, as amostras utilizadas na aplicação estão espalhadas neste sistema. O cultivar1-safra2 é a amostra mais próximo da Gama, o cultivar7-safra3 é a amostra mais próxima da distribuição uniforme e a outra amostra corresponde ao cultivar7-safra1.

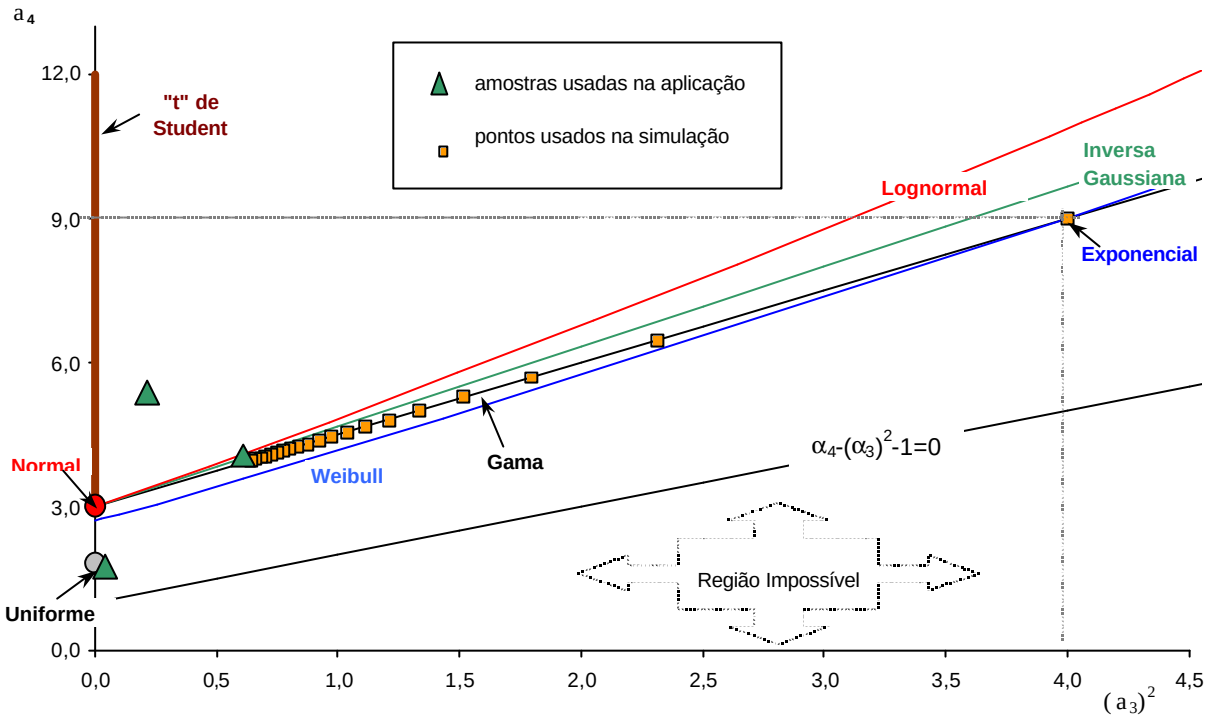


Figura 11 – Representação de alguns modelos probabilísticos no plano assimetria curtose ($(\alpha_3)^2\alpha_4$), pontos sob a curva do modelo Gama usado para a simulação de Monte Carlo e as amostras usadas na aplicação.

7 CONCLUSÕES

Com relação à comparação da qualidade dos estimadores de mínimos quadrados e de mínimos desvios absolutos em modelos de regressão linear simples, com erro gama, nas situações estudadas por simulação: tamanho amostral 20(20)100; coeficiente linear do modelo de regressão 20(10)40; coeficiente angular do modelo de regressão –0,3(-0,1)-0,5; parâmetro de forma do modelo Gama 0,25(0,50)9,75 e o parâmetro de escala 0,2(0,4)2,2, pode-se concluir:

1) a razão e a diferença de erros quadráticos médios, além de poderem ser usados como critérios para comparação da qualidade de estimadores, não diferindo entre si, produzem resultados diferentes do critério usual da variância residual média;

2) o parâmetro (λ) de escala do modelo Gama de probabilidade é responsável por diferenciar a qualidade dos estimadores: $\lambda \leq 1$ o estimador de mínimos quadrados produz menor erro quadrático médio, caso contrário, o melhor estimador é o mínimos desvios absolutos;

8 BIBLIOGRAFIA CONSULTADA

ABDELMALEK, N. N., An Efficient Method for the Discrete Linear L_1 Approximation Problem. *Mathematics of Computation*, v.29, n.131, p.844-50, 1975.

ARENALES, M.N., On Techniques for Curve Fitting Based on Minimum Absolute Deviations. In: ESCOLA DE MODELOS DE REGRESSÃO, 2, 1991, Universidade Federal do Rio de Janeiro. *Atas* ... São Paulo: Associação Brasileira de Estatística, 1991. p.06-11.

BARRODALE, I., ROBERTS, F.D.K., An Improved Algorithm for Discrete L_1 Linear Approximation. *SIAM J. Numer. Anal.*, v.10, n.5, p.839-48, 1973.

BLOOMFIELD, P., STEIGER, W., Least Absolute Deviations Curve – Fitting. *SIAM J. Sci. Stat. Comput.*, v.1, n.2, p.290-301, 1980.

BOLFARINE, H., RODRIGUES, J., CORDANI, L., O Modelo de Regressão com Erros nas Variáveis. *10º Simpósio Nacional de Probabilidade e Estatística*, 30p, 1992.

COHEN, A. C., WHITTEN, B. J., *Parameter Estimation in Reliability and Life Span Models*. New York: Marcel Dekker, Inc, 1988, 394p.

GUISCEM, J.M., *Qualidade Fisiológica de Sementes de Milho Doce em Função do Teor de Água na Colheita e da Temperatura de Secagem*. Botucatu, 2001. 123p. Tese (Doutorado em Agronomia/Agricultura) – Faculdade de Ciências Agronômicas, Universidade Estadual Paulista.

HAHN, G.J., SHAPIRO, S.S., *Statistical Models in Engineering*. New York: John Wiley & Sons, Inc., 1967. 355p.

HOFFMANN, R., VIEIRA, S., Regressão Linear Simples. In: __. *Análise de Regressão: Uma Introdução à Econometria*. 2.ed. São Paulo: Editora "HUCITEC", 1987. p.39-76.

KOENKER, R., BASSETT, G.Jr., Regression Quantiles. *Econometrica*, v.46, n.1, p.33-50, 1978.

KOENKER, R., PORTNOY, S., Quantile Regression.. In: ESCOLA DE MODELOS DE REGRESSÃO, 5, 1997, Campos do Jordão. *Minicurso* ... São Paulo: Associação Brasileira de Estatística, 1997. 77p.

MEYER, P.L., *Probabilidade: Aplicações à Estatística*. Rio de Janeiro: Ao Livro Técnico S.A., 1969. 391p.

MOOD, A.M., GRAYBILL, F.A., BOES, D., *Introduction to the Theory of Statistics*. 3 ed. Singapore: McGraw-Hill Book Company, 1974, 564p.

NARULA, S.C., STANGENHAUS, G., Análise de Regressão L_1. *8º Simpósio Nacional de Probabilidade e Estatística*, 67p, 1988.

NARULA, S.C., WELLINGTON, J.F., *An Algorithm to Find All Regression Quantiles Using Bicriteria Optimization*. s.n.t.

SAS, *Statistical Analysis System for Windows Release 6.12* – SAS System Help. Cary, North Carolina, USA. 1996.

SHAPIRO, S.S., GROSS, A.J., *Statistical Modeling Techniques – Statistical: textbooks and monographs*. v.38. New York: Marcel Dekker, Inc., 1981. 367p.

SILVA, M.A.Z.M., *O Problema de Aproximação Linear no L_1 e Extensões*. São Carlos, 1994. 70p. Dissertação (Mestrado em Ciências/Ciências de Computação e Matemática Computacional) Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo.

SILVA, M.A.Z.M., Resolução do Problema de Regressão Quantil Através do Método Primal Simplex. In: CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 20, 1997, Gramado-Rio Grande do Sul, *Resumos* ... Sociedade Brasileira de Matemática Aplicada e Computacional, 1997. p.411-12.

SILVA, M.A.Z.M., Uma Extensão do Método Simplex. Para a Resolução do Problema de Regressão Quantil. *Revista de Matemática e Estatística, São Paulo*, v.18, p.125- 44, 2000.

SILVEIRA JUNIOR, P.S., MACHADO, A.A., ZONTA, E.P., SILVA, J.B., *Curso de Estatística*. v.2. Pelotas: Editora Universitária, 1992. 233p.

STANGENHAUS, G., NARULA, S.C., Inference Procedures for the L_1 Regression. *Computational Statistics & Data Analysis.* v.12, p.79-85, 1991.

WAGNER, H.M., Linear Programming Techniques for Regression Analysis. *Amer. Statist. Assoc. J.*, v.54, p.206-12, 1959.

APÊNDICE 1

MOMENTOS

Momentos são quantidades que auxiliam na descrição de uma distribuição de probabilidade. Assim, tanto melhor se conhece uma distribuição quanto mais momentos dessa distribuição são determinados. Os tipos mais importantes de momentos são dois: os momentos ordinários, representados por μ'_r, e os momentos centrados na média, representados por μ_r, sendo r a ordem dos momentos (Silveira Júnior et al., 1992).

Se X é uma variável aleatória contínua, com função de densidade f, o r-ésimo momento em torno de zero (na origem) é definido como:

$$\mu'_r = E(X^r) = \int_{-\infty}^{+\infty} x^r f(x)\,dx$$

O r-ésimo momento em torno da média é definido por:

$$\mu_r = E(X - \mu)^r = \int_{-\infty}^{+\infty} (X - \mu)^r f(x)\,dx .$$

Nota-se facilmente que, para r = 1, tem-se:

$$\mu'_1 = E(X) = \mu$$
$$\mu_1 = E(X - \mu) = E(X) - E(\mu) = \mu - \mu = 0$$

e, para r = 2,

$$\mu_2 = E(X - \mu)^2 = V(X)$$

ou seja, a média e a variância são casos particulares dos momentos que, pela sua importância, são estudados separadamente.

De uma maneira geral é quase sempre preferível colocar os momentos centrados na média em função dos momentos centrados na origem, facilitando assim o seu cálculo. Por exemplo, considere os casos de μ_2 e μ_3:

$$\begin{aligned}\mu_2 &= E(X - \mu)^2 \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= \mu'_2 - 2\mu'_1 \mu'_1 + (\mu'_1)^2 \\ &= \mu'_2 - (\mu'_1)^2\end{aligned}$$

e

$$\begin{aligned}\mu_3 &= E(X - \mu)^3 \\ &= E(X^3 - 3X^2\mu + 3X\mu^2 - \mu^3) \\ &= E(X^3) - 3\mu E(X^2) + 3\mu^2 E(X) - \mu^3 \\ &= \mu'_3 - 3\mu'_1 \mu'_2 + 3(\mu'_1)^2 \mu'_1 - (\mu'_1)^3 \\ &= \mu'_3 - 3\mu'_1 \mu'_2 + 2(\mu'_1)^3\end{aligned}$$

Para esses e para outros momentos de ordem superior, no entanto, pode ser usada a seguinte forma de recorrência (Binômio de Newton):

$$\mu_r = \sum_{i=0}^{r} (-1)^i \binom{r}{i} \mu'_{r-i} (\mu'_1)^i$$

Para r = 4, por exemplo:

$$\begin{aligned}\mu_4 &= \sum_{i=0}^{4} (-1)^i \binom{4}{i} \mu'_{4-i} (\mu'_1)^i \\ &= \binom{4}{0}\mu'_4 - \binom{4}{1}\mu'_3 \mu'_1 + \binom{4}{2}\mu'_2 (\mu'_1)^2 - \binom{4}{3}\mu'_1 (\mu'_1)^3 + \binom{4}{4}(\mu'_1)^4 \\ &= \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 4\mu'_1 (\mu'_1)^3 + (\mu'_1)^4 \\ &= \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^4\end{aligned}$$

APÊNDICE 2

PROGRAMA FONTE EM LINGUAGEM SAS PARA A SIMULAÇÃO

E A FUNÇÃO RANGAM

2.1 Programa fonte em linguagem SAS (SAS, 1996) para a simulação

O programa fonte apresentado a seguir gera os conjuntos de dados para o caso em que $\beta_0 = 20$ e $\beta_1 = -0,3$.

```
data marcia.gera_aa (keep=tamos lambda eta repete x y);
  beta0 = 20;
  beta1 = -0,3;
  do tamos = 20 to 100 by 20;
    do lambda = 0,2 to 2,2 by 0,4;
      do eta = 0,25 to 9,75 by 0,5;
        do repet e= 1 to 1000;
          do x = 1 to tamos;
            ei = lambda*rangam(0,eta);
            wi = (ei-eta/lambda)/sqrt(eta/(lambda**2));
            y = beta0+ beta1*x+wi;
            output;
          end;
        end;
      end;
    end;
  end;
```

2.2 Função RANGAM

O programa SAS possui uma função pré-definida que gera variáveis aleatórias segundo modelo Gama, com parâmetro λ de escala e parâmetro η de forma; e é dada por:

λ RANGAM (semente, η).

Como neste trabalho, a variável erro segue uma distribuição Gama padronizada é necessário gerar o erro (ei) seguindo distribuição Gama e, em seguida, utilizar as equações (4.8) e (4.9) a fim de obter o erro (wi) seguindo Gama padronizada. Isto deve ser feito considerando a variação do parâmetro de escala λ e a variação do parâmetro de forma η. Assim, considerando semente=0, foram executados os seguintes comandos:

```
ei = lambda*rangam(0,eta);
wi = (ei-eta/lambda)/sqrt(eta/(lambda**2));
```

APÊNDICE 3

ALGORITMO DE REGRESSÃO L_1

ALGORITMO

Determine uma partição básica inicial, isto é: $X = \begin{pmatrix} B \\ N \end{pmatrix}$, com $B \in \Re^{2x2}$ e $N \in \Re^{(m-2)x2}$.

Passo 1: Iniciação

- Escolha uma solução básica viável não degenerada:

$$\begin{cases} \hat{\beta} = B^{-1} y_B \\ \hat{\varepsilon}_N = y_N - N\hat{\beta} \\ \hat{\varepsilon}_B = 0 \end{cases}$$

Passo 2: Teste de Otimalidade

- Calcular

$$\hat{v}^t = -\text{sinal}\,(\hat{\varepsilon}_N)^t \, NB^{-1}$$

- Se $-1 \leq \hat{v}_i \leq 1 \;\; \forall i \in B$

Então

- $(\hat{\beta}, \hat{\varepsilon})$ é ótima. Pare.

Senão

- o i-ésimo ponto que deixará de ser interpolado é calculado tal que:

$$\min\{1 - \hat{v}_i,\ \hat{v}_i + 1\}$$

Passo 3: Determinação do Tamanho do Passo

- Considere $\Delta_0 = 1 - \hat{v}_i$ (se $\Delta_0 = 1 + \hat{v}_i$ apenas 3.1 se altera)

3.1 Calcule os pontos críticos $k \; I \; N$

$$\delta_k = - \frac{\hat{\varepsilon}_k}{\hat{N}^i_k} \geq 0$$

(Se $\hat{\varepsilon}_k = 0$, $\delta_k = 0$ é considerado apenas se $\text{sinal}\left(\hat{N}_k^i\right) = -\text{sinal}\left(\hat{\varepsilon}_k\right)$)

3.2 Ordenação dos δ_k (r pontos)

$$0 < \delta_{k_1} \leq \delta_{k_2} \leq \ldots \leq \delta_{k_r}$$

3.3 Passo ótimo

$j = 0$

Enquanto $(\Delta_{j-1} + 2\left|\hat{N}_{k_j}^i\right|) < 0$ faça

$j = j + 1$

$\Delta_j = \Delta_{j-1} + 2\left|\hat{N}_{k_j}^i\right|$

(k_j é o novo ponto a ser interpolado)

Passo 4: Atualização

- Atualização da partição básica: B e N são alteradas por $x_{B_i} \leftrightarrow x_{N_{k_j}}$
- Volte para o *Passo 2.*

APÊNDICE 4

HISTOGRAMAS DO TEOR DE UMIDADE DO MILHO

SEGUNDO CULTIVAR SAFRA

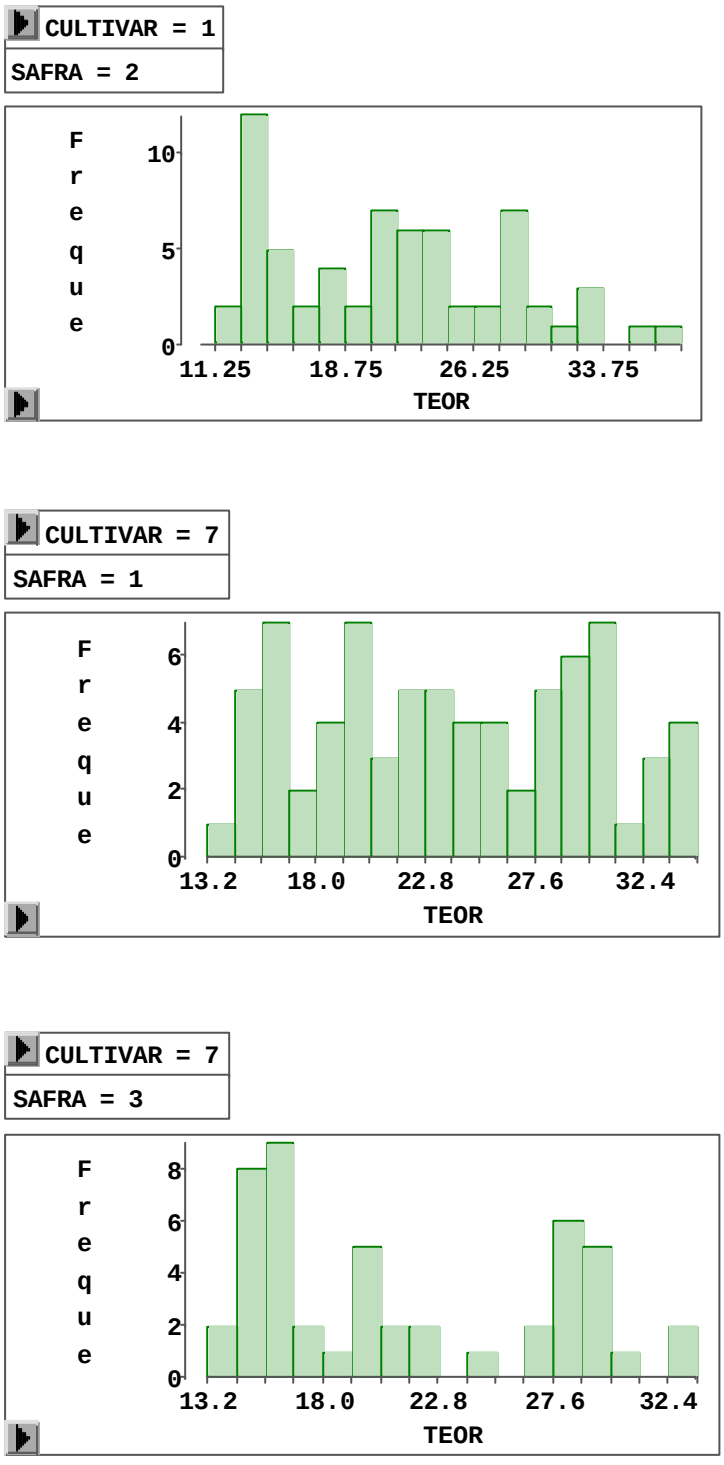


Figura 12: Histogramas do teor de umidade do milho segundo cultivar1-safra2, cultivar7-safra1 e cultivar7-safra3.