Univesidade Federal Rural do Rio de Janeiro

Instituto de Ciências Exatas

Departamento de Matemática

Utilização do Modelo de Regressão Linear Múltipla Aplicado na Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba

João Paulo de Castro Antunes

Orientador: Prof. Dr. Wagner de Souza Tassinari

Orientadora: Prof. Dr. Maria Cristina Lorenzon

Seropédica

2009

João Paulo de Castro Antunes

Utilização do Modelo de Regressão Linear Múltipla Aplicado na Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba

Sob a orientação do Prof. Dr. Wagner de Souza Tassinari

e Prof. Dr. Maria Cristina Lorenzon

Monografia submetida como requisito parcial para obtenção do grau de Licenciado e Bacharel em Matemática.

Seropédica

2009

João Paulo de Castro Antunes

Utilização do Modelo de Regressão Linear Múltipla Aplicado na Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba

Monografia submetida como requisito parcial para obtenção do grau de Licenciado e Bacharel em Matemática, submetida à aprovação da banca examinadora composta pelos seguintes membros:

Prof. Dr. Wagner de Souza Tassinari

Prof. Dr. Maria Cristina Lorenzon

Prof. Mestre Adriana Oliveira Andrade

Seropédica, 14 de julho de 2009.

Primeiramente agradeço a Deus, que é a luz que ilumina o meu caminho.

Aos meus pais que sempre souberam dizer as palavras certas nas horas certas.

Aos meus irmãos que mesmo de longe nunca deixaram de me apoiar.

A minha familia em geral.

Aos professores e amigos da minha universidade.

A todos muito obrigado.

Resumo

O estado do Rio de Janeiro é um dos maiores centros consumidores de mel do país, mas no que diz respeito à produção continua estagnada, favorecendo a importação de mel de outros estados. O estado é também um dos que apresentam maior devastação ambiental e índices muitos pobres de suporte a agricultura familiar (Sebrae, 2006) . A região da Costa Verde do estado do Rio de Janeiro é uma das regiões menos expressivas na produção apícola.

Este trabalho objetiva identificar alguns fatores que contribuem para variabilidade do preço do mel nos municípios de Angra dos Reis e Mangaratiba. Tal estudo foi feito utilizando técnicas de modelagem de regressão linear múltipla, fornecendo como resultado uma reta de regressão que nos permite prevê o preço médio de 100g de mel, auxiliando assim o consumidor na pesquisa e compra de mel na região da Costa Verde - RJ

Sumário

Introdução

Os modelos de regressão linear vem sendo largamente utilizados em diversas áreas do conhecimento, tais como: computação, administração, engenharias, biologia, agronomia, saúde, sociologia, etc.

Tomando como base o trabalho feito por Scudino (2007), o objetivo dessa monografia é analisar potênciais fatores determinísticos do preço do mel na região da Costa Verde do estado do Rio de Janeiro, através da utilização de modelos de regressão linear múltipla. O estudo foi realizado nas cidades de Angra dos Reis e Mangaratiba, no período compreendido entre janeiro e julho de 2007.

Essa monografia está dividida em cinco capítulos. No primeiro serão apresentados os modelos de regressão linear simples. No segundo serão apresentados os modelos de regressão linear múltipla, enfatizando o uso de variáveis dummeis. No terceiro abordaremos o uso de variáveis Dummeis.

No quarto capítulo serão mostrados alguns métodos de diagnósticos de resíduos sobre o modelo de regressão linear. No quinto capítulo será abordado um pouco sobre o mercado de mel, sendo feita uma contextualização do assunto.

E, finalmente, no sexto capítulo será feita uma aplicação desses modelos e dos respectivos métodos de diagnósticos no estudo dos fatores determinantes da vari-

abilidade do preço do mel na região da Costa Verde - RJ.

Capítulo 1

Regressão Linear Simples

1.1 Introdução

A finalidade da análise de regressão é estudar a relação funcional entre duas ou mais variáveis, as quais assumem valores quantitativos ou qualitativos, de tal forma que elas possam ser preditas a partir de outras, ou seja, podem projetar ou estimar uma nova observação (Martins, 2005). Muitos são os exemplos que podemos citar/enumerar onde uma variável é predita através de outra ou de outras, onde existe esta relação funcional entre as variáveis.

A) Preço do aluguel de um imóvel a partir de sua localização, área total do imóvel, número de quartos e estado de conservação do imóvel;

B) Aumento na produção de uma cultura em relação à quantidade de adubo fornecido às plantas;

C) Número de pessoas internadas com sintomas da Dengue em relação ao índice pluviométrico da região estudada.

Observando esses exemplos o objetivo é medir a relação entre elas, essa medida terá as seguintes características:

1 - se houver relação entre as variáveis, esta relação será forte ou fraca;

2 - caso observe-se esta relação, o objetivo seguinte é construir um modelo que possa interpretar a função existente entre as variáveis envolvidas no estudo;

3 - obtendo este modelo, pode-se utilizá-lo para estimar valores futuros da variável independente e simultaneamente verificar o efeito de cada variável independente em relação a variavel dependente.

Assim podemos concluir que o conjunto de variáveis pode ser constituido por vários elementos, os quais poderão estar relacionados entre si. Quando tiver uma relação entre duas variáveis, utilizaremos a regressão linear simples, quando a relação for entre mais de duas variáveis utilizaremos a regressão linear múltipla.

Para Akaike (1974) os modelos de regressão obedecem algumas características:

1 - buscar relações de causa e efeito das variáveis;

2 - predição de valores;

3 - estabelecer uma explicação sobre uma população a partir da amostra.

1.2 Modelo de Regressão Linear Simples

Martins (2005) explica que a natureza da relação entre as variáveis pode tomar várias formas, desde uma simples relação linear até uma complexa função matemática.

O modelo de regressão linear simples pode ser representado como:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{1.1}$$

onde: Y_i = valor predito pelo modelo para a i-ésima observação.

α = intercepto

β = inclinação da reta

ε = erro aleatório de Y_i para a i-ésima observação.

Onde α representa o valor da variável Y quando a variével $X = 0$; β representa a inclinação da reta de regressão, isto é, a mudança de Y por cada unidade de X ; ε representa uma variável aleatória que descreve o erro de Y para cada observação i.

Com base em uma amostra pretendemos expecificar uma equação de regressão, ao estimar os valores dos coeficientes da reta de regressão:

$$\hat{Y}_i = a + bX_i \tag{1.2}$$

onde: $\hat{Y}$ = o valor da previsão de Y para uma observação X_i

X_i = o valor de X para cada observação i

a = o estimador de α

b = o estimador de β

Agora basta ajustar os parâmetros a e b de tal forma que a reta se ajuste da melhor forma ao nosso conjunto de pontos. Existem várias formas de estimar estes parâmetros. Nesse estudo utilizaremos o Método dos Mínimos Quadrados (MMQ), que consiste em minimizar a soma dos quadrados dos erros, ou seja minimizar a

função de mínimos quadrados (Bussab, 2002):

O método de ajuste dos mínimos quadrados é preferível por que:

1. Onera os desvios maiores, fato desejável que evita grandes desvios.

2. Permite realizar testes de significância na equação de regressão.

3. A reta de regressão passa pelo ponto formado pelos valores das médias das duas séries de observações.

O erro de previsão ε corresponde a diferença entre um valor observado Y e o valor correspondente de $\hat{Y}$ da reta. O que se deseja e temos que fazer é torna esse erro pequeno, ou seja, $(Y - \hat{Y})$ o menor possível. Vamos somar todos os erros e fazer que essa soma de zero, isto é: $\sum_{i=1}^{n} e_i = 0$.

Na reta os pontos que estão acima dão erros positivos e os que estão abaixo dão erros negativos por isso usaremos os quadrados dos erros em nossos cáculos, daí o nome de Mínimos Quadrados e como $\hat{Y}_i = a + bX_i$, vamos minimizar os erros (Montgomery & Runger, 2003):

Sendo a soma dos quadrados dos desvios das observações em relação à reta de regressão dada por

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Yi - a - bX_i)^2 \tag{1.3}$$

Os estimadores dos mínimos quadrados têm que satisfazer

$$\frac{\partial L}{\partial a} = -2 \sum_{i=1}^{n} (Yi - a - bX_i) = 0 \tag{1.4}$$

$$\frac{\partial L}{\partial b} = -2\sum_{i=1}^{n}(Yi - a - bX_i)X_i = 0 \tag{1.5}$$

A simplificação dessas duas equações resulta em

$$na + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \tag{1.6}$$

$$a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i \tag{1.7}$$

Resolvendo o sistema para a e b, temos:

$$b = \frac{S_{XY}}{S_{XX}} \qquad e \qquad a = \bar{y} - b\bar{x} \tag{1.8}$$

onde:

$$S_{XY} = \sum xy - \frac{\sum x \sum y}{n} \qquad e \qquad S_{XX} = \sum x^2 - \frac{(\sum x)^2}{n} \tag{1.9}$$

$$\bar{x} = \frac{\sum x}{n} \qquad e \qquad \bar{y} = \frac{\sum y}{n} \tag{1.10}$$

Após encontrarmos os coeficientes é necessário fazer algumas hipóteses sobre a variável aleatória ε para que esse modelo tenha validade.

1.3 Hipóteses Sobre o Modelo

Martins (2005) explica que para validação das inferências feitas, são necessárias algumas hipóteses sobre o comportamento da variável aleatória que representa os possíveis erros da variável dependente Y, no modelo:

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

São elas:

1. A distribuição de probabilidade dos erros segue uma distribuição normal, ou seja,
   $\varepsilon \sim N(0, \sigma^2)$.

2. A média da distribuição de probabilidade da variável ε é zero, isto é: $\mu_\varepsilon = E[\varepsilon] = 0$. Assim, para cada observação X, a média dos erros para uma grande série de experimentos é zero.

3. A variância da distribuição de probablilidade da variável ε é constante para todos os valores de X, e é igual a σ^2. Isto é: $Var[\varepsilon] = \sigma^2$.

4. Os erros associados a duas observações quaisquer são independentes. Isto é, o erro associado com um valor de ε_i não afeta o erro associado com outro valor de ε_{i+1}, ou seja, $cor(\varepsilon_i, \varepsilon_{i+1}) = 0$

Para algumas aplicações, é necessário um estimador para a variância do erro que ajudará a inspecionar o comportamento do resíduo no diagrama de inspeção, e observar o grau de precisão do ajuste. O método para calcular esta estimativa é obitida a soma dos quadrados dos desvios pela número de graus de liberdade da soma, onde os desvios são as diferenças entre as observações e a estimativa do modelo proposto. Observe a fórmula abaixo:

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\phi} = \frac{\sum_{i=1}^{n}(\varepsilon_i)^2}{\phi} \tag{1.11}$$

onde ϕ representa o número de graus de liberdades, que no nosso caso é $n-2$ pois já temos dois parâmetros estimados a e b da reta $\hat{Y}_i = a + bX_i$. Substituindo temos:

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(\varepsilon_i)^2}{n-2} \tag{1.12}$$

A raiz quadrada de S^2 é o desvio padrão, ou erro padrão da estimativa.

Assim:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(\varepsilon_i)^2}{n-2}} \tag{1.13}$$

Denomina-se variação residual a soma dos quadrados dos desvios $\sum(y_i - \hat{y}_i)^2$:

$$VR = \sum_{n=i}^{n}(y_i - \hat{y}_i)^2 = S_{YY} - (b)(S_{XY}) \tag{1.14}$$

Desta maneira também poderemos escrever:

$$S^2 = \frac{VR}{n-2} = \frac{S_{YY} - (b)(S_{XY})}{n-2} \quad e \quad S = \sqrt{\frac{VR}{n-2}} = \sqrt{\frac{S_{YY} - (b)(S_{XY})}{n-2}} \tag{1.15}$$

onde:

$$S_{YY} = \sum_{i=1}^{n} y^2 - \frac{(\sum_{i=1}^{n} y)^2}{n} \tag{1.16}$$

1.4 Coeficiente de Determinação (R^2)

Diz respeito a um indicador que mensura a qualidade do ajuste do modelo, ou seja, verifica a aderência da reta de regressão estimada ao conjunto de dados.

Desta maneira, o coeficiente de determinação é dado por:

$$R^2 = \frac{VE}{VT} \tag{1.17}$$

$$R^2 = \frac{b(S_{XY})}{S_{YY}} \tag{1.18}$$

$$R^2 = \frac{VT - VR}{VT} \tag{1.19}$$

onde: $0 \leq R^2 \leq 1$, ou se multiplicarmos o R^2 por 100 teremos: $0\% \leq R^2 \leq 100\%$

Sejam:

- Variação Total

$$VT = S_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y^2 - \frac{(\sum_{i=1}^{n} y)^2}{n} \tag{1.20}$$

- Variação Explicada pela Variável Independente

$$VE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b(S_{XY}) \tag{1.21}$$

- A variação residual será calculada como a diferença entre a variação total e a variação explicada:

$$VR = VT - VE = S_{YY} - b(S_{XY}) \tag{1.22}$$

Pode dizer-se que R^2 representa a percentagem da variabilidade de Y que é explicada pela regressão. Este coeficiente é a medida de quão bem a reta de regressão se ajusta aos dados. R^2 varia entre zero e um, e quanto mais próximo de um estiver, melhor será o ajuste do modelo.

$R^2 = 1$ indica que o ajuste é perfeito, pois todos os pontos observados estão na reta ajustada, ou seja as variações de $y_i e \hat{y}_i$, são completamente explicadas pelas variações de X .Se $R^2 = 0$, conclui-se que as variações de Y são devidas a outras razões e a inserção de X no modelo não afetará as variações de novas variáveis Y, ou seja, não existiria regressão linear no conjunto de dados.

1.5 Inferência Sobre o Coeficiente de inclinação (β)

1.5.1 Teste de Hipóteses

Com este teste poderemos verificar a existência, ou não, associação entre as variáveis X e Y(Martins, 2005). O teste com o parâmetro β verifica a contribuição da variável X na explicação da variabilidade de Y.

Procedimento para realização do teste:

1. Formulação da hipóteses

   $H_0 : \beta = 0$

   $H_1 : \beta \neq 0$

   $H_1 : \beta > 0$

   $H_1 : \beta < 0$

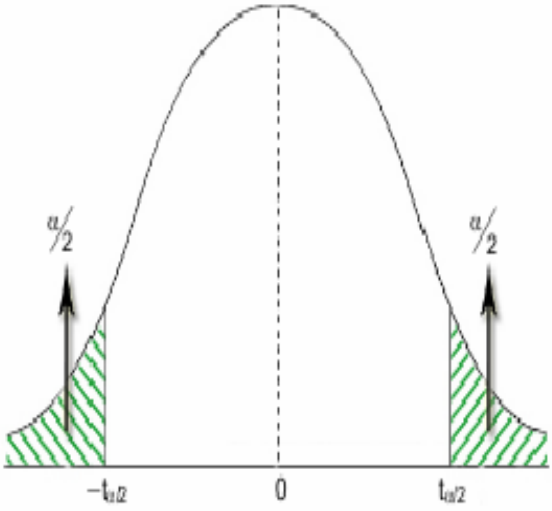
2. Fixar α (probabilidade de erro) e escolher a variável do teste, no caso, a distribuição t de Student com ϕ = n - 2

3. Com auxílio da *tabela* t, construir as regiões para rejeição (RC) e aceitação (não-rejeição) RA para H_0.

4. Com os dados amostrais, calcular o valor da variável:

$$t_{cal} = \frac{b}{\frac{S}{\sqrt{S_{XX}}}} \tag{1.23}$$

Figura 1.1: Regiões Críticas



Fonte: Martins (2005)

5. Conclusão do teste:

   - Caso (a): se $t_{cal} > t_{\frac{\alpha}{2}}$, ou $t_{cal} < -t_{\frac{\alpha}{2}}$,rejeita-se H_0, com risco α, ou seja, que a contribuição de X é significativa.
   - Caso (b): Se $-t_{\frac{\alpha}{2}} \leq t_{cal} \leq t_{-\frac{\alpha}{2}}$, não rejeita H_0 com risco α, ou seja, que a contribuição é nula.

1.5.2 ANOVA da Regressão (teste F)

Podemos verificar a existência da contribuição da variável X também através da utilização da análise de variância (ANOVA), estudando as variações totais, explicadas e residuais. Para executar o teste F seguimos os seguintes procedimentos:

(a) Enunciar as hipóteses: $H_0 : \beta = 0$

$H_1 : \beta \neq 0$

(b) Fixar o nível de significância do teste (α) e escolher uma variável F com 1 grau de liberdade no numerador (devido ao uso de uma variável independente) e $(n-2)$ graus de liberdade no denominador.

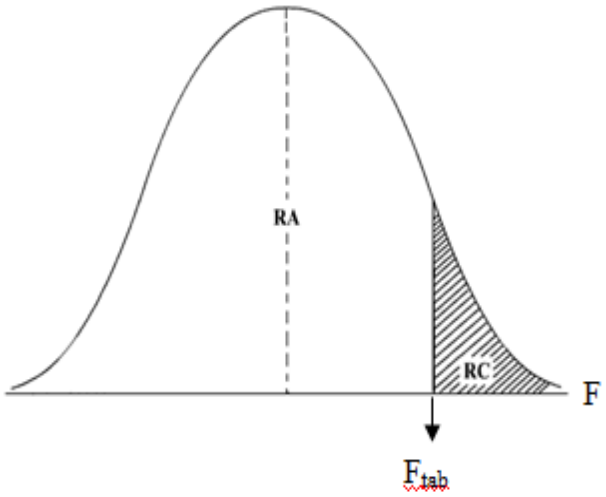
(c) Com o auxílio da tabela de distribuição F, determinan-se as regiões de reijeição e não rejeição (aceitação).

(d) Montar o quadro de Análise das Variâncias (QAV)

Fonte de Variação	Soma dos Quadrados	GL	Quadrados Médios	$F_{calculado}$
Devido a variável	VE = bS_{XY}	1	$bS_{XY}/1$	
Residual	VR = S_{YY} - S_{XY}	n - 2	$S^2 = \frac{VR}{(n-2)}$	Fcal = $\frac{bS_{XY}}{S^2}$
Total	$VT = S_{YY}$	n - 1		

- Conclusão:Obtemos o F_{tab} por F(α, n - 2) e daí se $F_{cal} > F_{tab}$ rejeita-se H_0, concluindo-se, com risco α, que existe regressão linear simples, isto é, o modelo pode explicar e prever a variável Y.

Figura 1.2: Regiões Críticas



Fonte: Martins (2005)

1.6 Intervalos de Confiança

Uma forma alternativa ao teste do parâmetro β_1 do modelo é a contribuição de intervalos de confiança em torno do parâmetro β ou do valor previsto de Y (Martins, 2005).

Quanto maior for o intervalo em torno da reta ajustada, menor será a precisão do ajuste é uma medida de qualidade do ajuste.

Para o intervalo de confiança do coeficiente de inclinação (β) temos:

$$P(b - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{XX}}} \leq \beta \leq b + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{XX}}}) = (1 - \alpha)100 \tag{1.24}$$

Para o intervalo de confiança do valor previsto Y temos:

$$P(\hat{Y}_{(x)} \pm t_{\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}}) = (1 - \alpha)100 \tag{1.25}$$

onde a distribuição t é tomada com $(n - 2)$ graus de liberdade, e os valores $\pm t_{\frac{\alpha}{2}}$ obtidos na tabela t de Student.

Capítulo 2

Regressão Linear Múltipla

2.1 Introdução

Na prática o modelo de regressão linear simples é pouco utlizado no mundo real, pois geralmente não existe apenas uma variável que influência a variável que queremos prever e/ou explicar. Modelos que contém mais de uma variável independente recebem o nome de modelos de regressão linear múltipla (Gujarati, 2000).

A regressão múltipla envolve três ou mais variáveis. Ou seja, ainda uma única variável dependente e duas ou mais variáveis independentes (explanatórias).

A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples.

Mesmo quando estamos interessados no efeito de apenas uma das variáveis, é aconselhavel incluir as outras capazes de afetar Y, efetuando uma análise

de regressão múltipla, por 2 razões:

(a) Para reduzir os resíduos. Reduzindo-se a variância residual (erro padrão da estimativa), aumenta a força dos testes de significância;

(b) Para eliminar a tendenciosidade causada pela omissão de uma variável que afeta Y substancialmente.

Uma estimativa é tendenciosa quando, por exemplo, numa pesquisa em que se deseja investigar a relação entre a aplicação de fertilizante e o volume de safra, atribuímos erroneamente ao fertilizante os efeitos do fertilizante mais a precipitação pluviométrica.

O ideal é que o modelo propossto seja capaz de explicar a maior parte da variabilidade da variável resposta com o menor número de variáveis independentes, sobretudo em virtude do custo na obtenção de dados para muitas variáveis e também pela necessidade de observações adicionais para compensar a perda de graus de liberdade decorrente da introdução de mais variáveis independentes.

2.2 Modelo de Regressão Linear Mútipla

A equação da regressão múltipla tem a forma seguinte:

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki}. \tag{2.1}$$

onde: $\hat{Y}_i$ é a variável dependente;

$X_{1i}, X_{2i}, ..., X_{ki}$ são as variáveis independentes;

a é o intercepto, também conhecido como média geral;

b_i determina os efeitos(contribuição) das variáveis independentes X_k (coeficiente de inclinação);

Agora precisamos estimar os nossos parâmetros para podermos escreve a equação do modelo, para isso podemos reescrever nosso modelo como (Montgomery & Runger, 2003) :

$$\hat{Y}_i = a + \sum_{j=1}^{k} b_j X_{ij} \qquad\qquad i = 1, 2, ..., n \tag{2.2}$$

A função dos míninos quadrados é

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (\hat{Y}_i - a - \sum_{j=1}^{k} b_j X_{ij})^2 \tag{2.3}$$

A função L deve ser minimizada com relação aos parâmetros a, b_1, ..., b_k. e as estimativas dos mínimos quadrados têm que satisfazer

$$\frac{\partial L}{\partial a} = -2 \sum_{i=1}^{n} (Yi - a - \sum_{j=1}^{k} b_j X_{ij}) = 0 \tag{2.4}$$

e

$$\frac{\partial L}{\partial b_j} = -2 \sum_{i=1}^{n} (Yi - a - \sum_{j=1}^{k} b_j X_{ij}) X_{ij} = 0 \tag{2.5}$$

Simplificando as equações (2.4) e (2.5), obtemos as equações normais de mínimos quadrados

$$na + b_1 \sum_{i=1}^{n} X_{i1} + b_2 \sum_{i=1}^{n} X_{i2} + ... + b_k \sum_{i=1}^{n} X_{ik} = \sum_{i=1}^{n} Y_i$$

$$a\sum_{i=1}^{n} X_{i1} + b_1 \sum_{i=1}^{n} X_{i1}^2 + b_2 \sum_{i=1}^{n} X_{i1}X_{i2} + ... + b_k \sum_{i=1}^{n} X_{i1}X_{ik} = \sum_{i=1}^{n} X_{i1}Y_i$$

. . . . .

. . . . .

. . . . .

. . . . .

$$a\sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_{ik}X_{i1} + b_2 \sum_{i=1}^{n} X_{ik}X_{i2} + ... + b_k \sum_{i=1}^{n} X_{ik}^2 = \sum_{i=1}^{n} X_{ik}Y_i$$

Note que há $p = k + 1$ equações normais, uma para cada um dos coeficientes desconhecidos de regressão. A solução para as equações normais serão os estimadores de mínimos quadrados dos coeficientes de regressão, a, $b_1, b_2, ..., b_k$. As equações normais podem ser resolvidas por qualquer método apropriado para resolver um sistema de equações lineares (Montgomery & Runger, 2003).

2.3 Hipóteses Sobre o Modelo

Para garantir a validade sobre os resultados obtidos pelo Método dos Mínimos Quadrados (MMQ), bem como a significância das inferências sobre

o modelo de regressão linaer múltipla, são necessárias as mesmas hipóteses que foram feitas para o modelo de regressão linear simples (Martins, 2005).

2.4 Estimação da Variância S^2

Como na regressão linear simples, a variância σ^2 do erro do modelo de regressão linear múltipla também pode ser estimador, constituindo-se em um indicador da qualidade do ajustamento. A expressão da variância amostral é dada por (Martins, 2005):

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1} \tag{2.6}$$

onde: n = número de observações no estudo.

k = número de variáveis independentes.

Quanto a estimativa do desvio padrão teremos:

$$S = \sqrt{S^2} \tag{2.7}$$

2.5 Coeficiente de Determinação (R^2)

2.5.1 Coeficiente de Determinação Clássico

Todo os cáculos para Variação Total (VT) (Eq. 1.18), Variação Explicada (VE) (Eq. 1.19) e Variação Residual (VR) (Eq. 1.15) são iguais aos utilizados

para o modelo de regressão linear simples. Mas no caso do Coeficiente de determinação para os modelos de regressão mútipla (R^2) apesar de ter mesma equação (Eq. 1.21) ele tende a superestimar o verdadeiro valor, assim quado necessário podemos ajusta-lo pelo coeficiente de determinação ajustado.

2.5.2 Coeficiente de Determinação Ajustado

O coeficiente de determinação ajustado tem por finalidade caracterizar a redução da variabilidade total de Y com o conjunto de variáveis X_i , onde $1 \leq i \leq n$ e $1 \leq j \leq k$. Se $R^2 = 1$ temos que $Y_i = \hat{Y}_i$, e além disso observa-se que R^2 aumenta com a adição de variáveis independentes, por isso usa-se o coeficiente de determinação ajustado neste caso, onde é dado por (Charnet, 1999):

$$R_a^2 = 1 - \frac{\frac{VE}{n-p}}{\frac{VT}{n-1}} = 1 - (\frac{n-1}{n-p})\frac{VE}{VT} \tag{2.8}$$

onde: p = o número de variáveis independentes.

2.6 Inferência sobre os Coeficientes de inclinação (β_i)

2.6.1 Existência de Regressão Linear Múltipla

Estamos interessados em testar o modelo, visando realizar previsões para Y com certa segurança. Será apresentado o roteiro para o teste do modelo com duas ou mais variáveis (Martins, 2005):

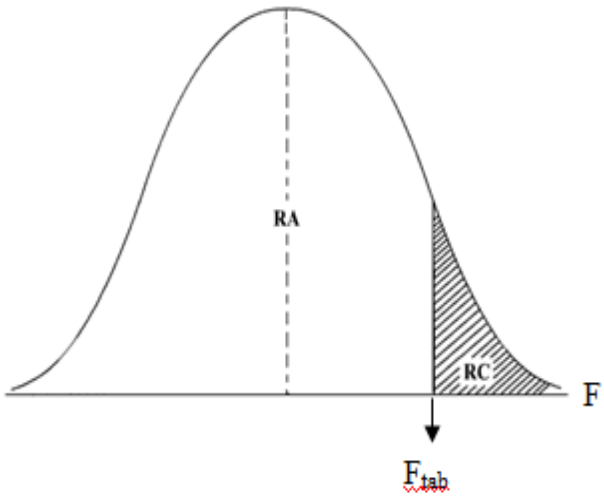
(a) Formulação das hipóteses

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

H_1: Todos os parâmetros $\beta_1, \beta_2, ..., \beta_k$ são diferentes de zero.

(b) Fixar α (probabilidade de erro) e escolher uma variável $F(\alpha; n-(k-1))$.

Figura 2.1: Teste de Hipótese F



Fonte: Martins (2005)

(c) Com auxílio da tabela de distribuição F, construir as regiões para rejeição (RC) e aceitação (não-rejeição) RA.

(d) Com os dados amostrais, calcular a estatística do teste:

$$F_{cal} = (\frac{R^2}{1-R^2})[\frac{n-(k-1)}{k}] \tag{2.9}$$

(e) Caso $F_{cal} > F_{tab}$, rejeita-se H_0, concluindo-se, com risco α, que existe regressão, isto é, o modelo é capaz de explicar e prever Y.

2.7 Nível de significância e p-valor

Para testar uma hipótese estabelecida, a probabilidade máxima com o qual se pode ocorrer o erro, é denominada nível de significância do teste (Spiegel, 1993). Normalmente, o nível de significância é representado por α e, geralmente, é especificado antes da extração das amostras e das hipóteses, de modo que os resultados obtidos não influênciem a escolha.

Usualmente são escolhidos os seguintes níveis $\alpha = 0,01$ ou $0,05$, isto é, se escolhido o índice de $0,01$, então existe 1 chance em 100, da hipótese ser rejeitada. Da mesma maneira podemos dizer que existe uma confiança de 99% de que se tome a decisão certa. Supondo que a hipótese nula seja verdadeira e que a probabilidade de se obter um efeito devido ao erro amostral seja menor do que 1%, o achado é dito significativo. Se a probabilidade for maior que 1%, o achado é dito não-significativo (Dancey & Reidy, 2006).

Na resposta dos testes de hipóteses, um valor é comparado com o nível

de significância previamente escolhido, sendo chamado de p-valor ou valor p, isto é, valor do poder do teste. O p-valor (nível de significância observado) é o menor nível de significância em que H_0 seria rejeitada, quando um procedimento de teste específico é usado em um determinado conjunto de dados.

Assim, quando p-valor $\leq \alpha$ implica na rejeição de H_0 no nível α. Ou se p-valor $> \alpha$ implica na não rejeição de H_0 no nível α. Então, em vários estudos as respostas poderão vir referenciando o nível de significância ou p-valor.

2.8 Técnica de seleção de variáveis explicativas ou independentes *Stepwise*

Um problema importante em muitas aplicações da análise de regressão envolve selecionar o conjunto de variáveis independentes ou preditoras a ser usado no modelo. Algumas vezes, experiência prévia ou considerações teóricas em foco podem ajudar o analista a especificar o conjunto de preditoras(Lins & Moreira, 1999).

Uma grande quantidade de julgamento e de experiência com o fenômeno sendo modelado é geralmente necessária para selecionar um conjunto apropriado de variáveis preditoras para um modelo de regressão múltipla.

Porém existem algumas técnicas que fazem essa seleção de maneira "automática", a mais usual se chama *stepwise*. Essa técnica é classificada de duas maneiras, *stepwise BackWard* e *stepwise ForWard*.

A regressão *stepwise BackWard* provavelmente, é a técnica mais utilizada para seleção de variáveis, o procedimento constrói iterativamente uma seqüência de modelos de regressão pela adição ou remoção de variáveis em cada etapa. O critério para adicionar ou remover uma variável em qualquer etapa é geralmente expresso em termos de um teste parcial F(Lins & Moreira, 1999).

A regressão *stepwise BackWard* começa formando um modelo com todas as variáveis independentes, e vai se retirando uma a uma as que não foram significativas ($p - valor \geq 0,05$).

O procedimento de seleção *Stepwise ForWard* é uma variação da regressão *stepwise BackWard*, e está baseado no princípio de que as variáveis preditoras devem ser adicionadas ao modelo uma de cada vez, desde que inclusas no modelo sejam significativas ($p - valor \geq 0,05$).

A seleção progressiva é uma simplificação da regressão *stepwise BackWard* que omite o teste parcial F de remoção do modelo das variáveis que foram adicionadas em etapas prévias. Essa é uma potencial fraqueza da seleção *stepwise ForWard*(Lins & Moreira, 1999).

Capítulo 3

Variáveis *Dummies*

3.1 Introdução

A finalidade desse capítulo é considerar o papel das variáveis explicativas qualitativas na análise de regressão. Na análise de regressão, a variável dependente é muitas vezes influênciada não somente pelas variáveis que podem ser facilmente quantificadas em alguma escala bem definida (por exemplo, renda, produto, preços, custos, altura e temperatura), mas também por variáveis de natureza essencialmente qualitativa (por exemplo, sexo, raça, cor, religião, nacionalidade)(Gujarati, 2000).

Como tais variáveis qualitativas geralmente indicam a presença ou a ausência de uma qualidade ou atributo, tais como homem ou mulher, negro ou branco, católico ou não-católico, um método para quantificar tais atributos é contruir variáveis artificiais que assumam valores de 1 ou 0. O 0 indicando a ausência de um atributo e 1 indicando a presença desse atributo. As variáveis que

assumem tais valores 0 e 1 são chamadas de variáveis *Dummies*.

As variáveis *Dummies* podem ser usadas nos modelos de regressão tão facilmente quanto as variáveis quantitativas. Aliás, um modelo de regressão pode conter variáveis explicativas que são excusivamente dummies, que é o caso do objeto de estudo nesta monografia.

3.2 Regressão Sobre uma Variável Explicativa Qualitativa com duas Classes ou Categorias (Variável *Dummie*)

Para uma melhor compreensão vamos utilizar um exemplo onde queremos prever o salário anual de um professor universitário e analisamos para isso seus anos de experiência de ensino e o sexo. Daí temos:

$$Y_i = \alpha + \delta D_i + \beta X_i + \varepsilon_i \tag{3.1}$$

em que: Y_i = salário anual de um professor universitário

X_i = anos de experiência de ensino

D_i = 1 se homem e 0 se for mulher

O modelo contém uma variável explicativa quantitativa (anos de experiência de ensino) e uma variável qualitativa (sexo) que tem duas classes (ou dois níveis), uma *Dummie*, homem e mulher.

3.3 Regressão Sobre uma Variável Explicativa Quantitativa e uma Qualitativa com Mais de duas Classes ou Categorias

Suponhamos que queiramos modelar a despesa anual com saúde feita por um indivíduo sobre a renda e o nivel de intrução do indivíduo. Como a variável "nivel de instrução"é de natureza qualitativa, suponha que consideremos três níveis de intrução mutualmente excludentes: formação superior, secundária e menos que secundária. Uma estratégia de modelagem seria tranformar a variável qualitativa de D categorias em D - 1 variáveis *Dummies* que representam cada categorias, a categoria não inclusa no modelo será a categoria de referênccia. Portanto podemos usar o seguinte modelo(Gujarati, 2000):

$$Y_i = \alpha + \delta D_{1i} + \eta D_{2i} + \beta X_i + \varepsilon_i \tag{3.2}$$

em que: Y_i = despesa anual com saúde

X_i = renda anual

D_{1i} = 1 se tiver formação secundária

= 0 caso contrário

D_{2i} = 1 se tiver formação superior

= 0 caso contrário

Note que na atribuição de variáveis *Dummies* do modelo, estamos tratando arbitrariamente a categoria "menos de formação secundária"como categoria-base ou de referência. Portanto o intecepto α reflitirá o intercepto dessa categoria.

3.4 Regressão Sobre duas Variáveis Qualitativas Explicativas

A técnica da variável dummy pode ser facilmente estendida para manipular mais de uma variável qualitativa, vontando agora ao primeiro exemplo sobre variáveis dummies, mas admitindo agora que, além dos anos de experiência de ensino e do sexo, a cor da pele da professor seja também um importante determinante da variável salário. Por simplicidade, suponha que haja duas categorias de cor, negro e branco. Podemos agora escrever o modelo como(Gujarati, 2000):

$$Y_i = \alpha + \delta D_{1i} + \eta D_{2i} + \beta X_i + \varepsilon_i \tag{3.3}$$

em que: Y_i = salário anual de um professor universitário

X_i = anos de experiência de ensino

D_{1i} = 1 se homem

= 0 se mulher

D_{2i} = 1 se for branco

= 0 for negro

Observe que cada uma das variáveis qualitativas, sexo e cor, tem duas categorias e as categorias base (ou referência) agora são: ser mulher e ser negro.

3.5 Mudança de inclinação da Reta por Variáveis *Dummies*

Supusemos até agora, nos modelos considerados, que as variáveis qualitativas afetam o intercepto, mas não o coeficiente de inclinação das variáveis. Assim outro formato possível que a variável *Dummy* pode assumir refere-se a mudanças na inclinação. A variável, assim, assume o valor zero para o período sem a mudança e o valor igual ao da variável cuja inclinação mudou para o período com mudança(Gujarati, 2000).

$$Y_i = \alpha + tX_i + \beta X_i + \varepsilon_i \tag{3.4}$$

Um exemplo onde utilizamos esse tipo de variável *Dummy* é quando vamos analisar dois tipos de grupos diferentes, como dois periodos diferentes de tempo de uma determinada pesquisa. Cada período de tempo receberá uma reta de regressão com coeficiente de inclinação diferente, assim para um período de tempo t receberá valor 1 e para o outro t receberá valor 0.

Capítulo 4

Análises de Resíduos

4.1 Introdução

O modelo de regressão linear estabelece, para seu desenvolvimento, um conjunto de pressupostos básicos, dos quais a maioria refere-se a variável ε_i. Desta forma os resíduos estimados devem passar por um processo de análise para que seja verificado se não há violação dos pressupostos. Deve ser realizado amplo processo de validação dos resultados baseado em testes complementares (Hochheim, 1999).

Já vimos o papel fundamental dos resíduos em várias ocasiões, como no método dos mínimos quadrados e na estimação das variâncias. Vamos agora ver o seu papel crucial na avaliação dos pressupostos do modelo, e consequentemente na "qualidade"do ajuste.

4.2 Testes para Normalidade dos Resíduos

Para a verificação dos pressupostos do modelos de regressão, devemos verificar a normalidade na distribuição dos residuos, ou seja, $\varepsilon_i \sim N(0, \sigma^2)$. Recordamos que os testes t e teste F utilizados anteriomentes exigem que o termo de erro siga uma distribuição normal. Caso contrário o procedimento dos testes não será válido (Gujarati, 2000).

Os resíduos devem seguir aproximadamente uma distribuição normal ou seja, com aproximadamente 95 % dos resíduos dentro do intervalo $[-2\sigma; +2\sigma]$.

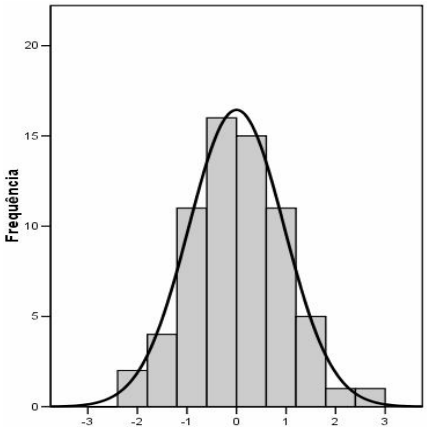
Existem várias formas de inspecionar a normalidade dos resíduos e veremos aqui algumas delas. A inspeção gráfica é uma forma bastante utilizada através de histograma dos resíduos e a comparação destes resíduos com distribuições normais simuladas, com mesma média e variânça, mas também existe alguns testes que podemos utilizar como o teste *Shapiro Wilk* e o teste *Komogorov*(Siegel & Castellen Jr, 2006)..

4.2.1 Técnicas gráficas

- **Histograma**

  Procuram-se afastamentos evidentes em relação à forma simétrica e unimodal da distribuição Normal, em geral, grandes assimetrias ou mais do que um máximo local.

Figura 4.1: Histograma de distribuição Normal

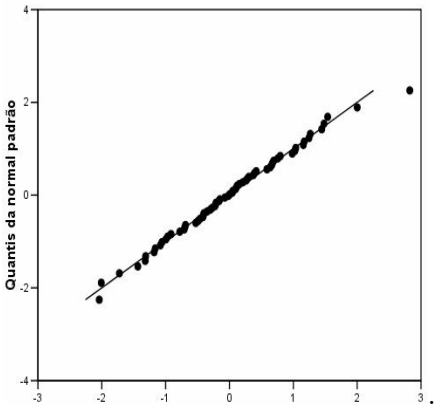


Fonte: Montgomery & Runger (2003)

- **QQ-plots**

  Representação gráfica dos quantis teóricos de uma normal reduzida vs. quantis empíricos dos resíduos observados.

Figura 4.2: Gráfico de Quantis QQ-plot



Fonte:Montgomery & Runger (2003)

4.2.2 Teste de Shapiro Wilk

O teste Shapiro-Wilk é utilizado para amostras com menos de 50 casos, ele calcula uma variável estatítica (W) que investiga se uma amostra aleatória provém de uma distribuição normal(Siegel & Castellen Jr, 2006).

Hipóteses sobre o teste:

H_0: A amostra provém de uma população Normal.

H_1: A amostra não provém de uma população Normal.

A variável W é calculada da seguinte forma:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{4.1}$$

sendo,

- x_i os valores ordenados de amostras (x_1 é o menor).

- a_i constantes geradas a partir de meio, variâncias e covariâncias da ordem estatística de uma amostra de tamanho n e uma distribuição normal.

Rejeitar H_0 ao nível de significância α se:

$$W_{cal} < W_{\alpha}$$

4.2.3 Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov pode ser aplicado para testar se a característica estudada da amostra é oriunda de uma população com distribuição normal. O teste é de execução simples, quando comparado ao qui-quadrado, e baseado na maior diferença absoluta entre a freqüência acumulada observada e a estimada pela distribuição normal(Siegel & Castellen Jr, 2006).

(a) Formulação das hipóteses

H_0 : A característica em estudo da população ou os erros (desvios) segue a distribuição normal.

H_1 : A característica em estudo da população ou os erros (desvios) não segue a distribuição normal.

(b) Escolha da significância α

(c) Estatística apropriada

A estatística apropriada do teste é baseada na maior diferença absoluta entre a função de distribuição normal acumulada, $\hat{F}(z_i)$, e a freqüência relativa observada acumulada e ajustada, F_i . As expressões encontram-se a seguir:

$$D_{máx} = g_{máx} + \frac{1}{2n} \tag{4.2}$$

onde:

$g_{máx}$: maior valor calculado de g (diferença absoluta);

n : tamanho da amostra ou número de parcelas. Sendo:

$$g = \mid \hat{F}(z_i) - F_{0,5} \mid ; sendo F_{0,5} = \frac{(i - 0,5)}{n} \tag{4.3}$$

onde: $F(z_i)$: Função de distribuição normal acumulada ; $F_{0,5}$:freqüência relativa observada acumulada e ajustada; i : número da amostra; n : tamanho da amostra ou número de parcelas.

(d) Conclusão

Para $n \leq 100$, quando o valor $D_{máx}$ for maior que o valor crítico tabelado $D_t(D_{máx} > D_t)$, para um tamanho de amostra n, δ = 0,5 e significância α (tabelado), a hipótese H_0 é rejeitada e conclui-se que a característica em estudo da população não segue a distribuição normal. Por outro lado, se $D_{máx}$ for menor que o valor crítico tabelado ($D_{máx} > D_t$), a hipótese H_0 é aceita e conclui-se que a característica em estudo da população segue a distribuição normal.

Para n > 100, o valor crítico D_t é obtido diretamente da expressão, sem o auxílio da tabela.

$$D_t = \sqrt{\frac{-ln(\frac{\alpha}{2})}{2n}} \tag{4.4}$$

onde: ln : logaritmo natural; α : significância estabelecida; n : tamanho da amostra ou número de parcelas.

Da mesma forma, quando o máximo valor de Dmáx for maior que o valor obtido pela expressão ($D_{máx} > D_t$), a hipótese H_0 é rejeitada e conclui-se que a característica em estudo da população não segue a distribuição normal; caso contrário ($D_{máx} > D_t$), a hipótese H_0 é aceita.

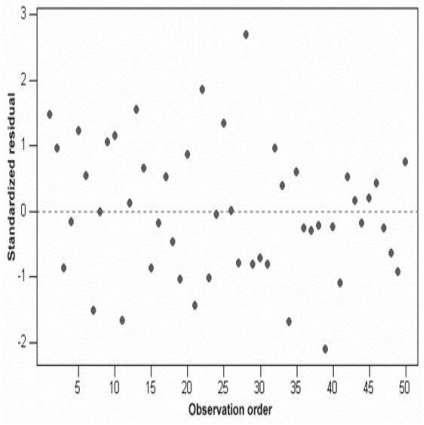
4.3 Resíduos Padronizados

Podemos também padronizar os resíduos, calculando:

$$d_i = \frac{\epsilon_i}{\sqrt{\sigma^2}} \qquad i = 1, 2, 3, ..., n. \tag{4.5}$$

Se os erros forem distribuídos normalmente, então aproximadamente 95% dos resíduos padronizados devem cair no intervalo (-2,+2). Os resíduos que estiverem fora desse intervalo podem indicar a presença de um outlier ; ou seja, uma observação que não é típica do resto dos dados. Várias regras têm sido propostas para descartar outliers. Entretanto algumas vezes, outliers fornecem informações de interesse para experimentalistas sobre circunstâncias não usuais, não devendo assim ser descartados (Montgomery & Runger, 2003). Abaixo um gráfico que ajuda a observar a presença de outliers e a normalidade da distribuição:

Figura 4.3: Resíduos Padronizados



Fonte: Montgomery & Runger (2003)

Capítulo 5

Mercado do Mel

5.1 Introdução

No Brasil, como em qualquel outro lugar do mundo, a criação de abelhas propicia a geração de inúmeros postos de trabalho, empregos e fluxo de renda, principalmente no que diz respeito à agricultura familiar, que desamparada, encontrou nesta atividade uma diversificação de sua produção.

A apicultura é uma atividade de fácil manutenção e de baixo custo inicial em relação às demais atividades agropecuárias e vem se tornando um importante segmento para reduzir os índices de pobreza de seus produtores, ao gerar alimentos, ocupação e renda. O setor apícola ganhou maior dimensão a partir da africanização das abelhas melíferas, que ao propiciar maior resistência das abelhas às doenças e ao ataque de inimigos naturais (De Jong, 1992), repercutiram no aumento da produção e no aperfeiçoamento de técnicas de seu manejo (Gonçalves, 1994). Em muitas regiões houve mudanças significativas

do perfil sócio-econômico, ora ampliando a margem de lucro como atividade complementar, ora figurando como atividade principal.

5.2 Produção Apícola no Brasil

No Brasil, 80% do total das propriedades rurais pertencem a grupos familiares, abrangendo um universo de 13,8 milhões de pessoas, que são responsáveis pela produção de grande parte dos alimentos consumidos no país (FIBGE, 2005).

A criação de animais é um segmento muito valorizado pela agricultura familiar, envolvendo-a em ações econômicas, ecológicas e sócio-culturais (Guelber, 2005), e a apicultura é uma dessas criações de destaque familiar, especialmente quando destinada à regiões e épocas desfavoráveis para a agricultura e outras atividades pecuárias.

Há 60 anos, o Brasil sofreu a africanização de suas abelhas, ou seja, a missigenação entre as sub-espécies de abelhas introduzidas da Europa e da África. Isto resultou na formação de uma abelha po-híbrida com mais produtividade e resistência.

No atual sistema produtivo fracionamos os fatores de produção apícola, verifica-se que alguns foram bem estudados, tais como, sanidade e genética, que permitem posicionar o poli-híbrido africanizado como bom produtor. No entanto, o baixo controle zoocténico das criações pode revelar índices de produção conflitantes (Gonçalves, 1994).Os avanços tecnológicos no se-

tor apícola estão se ampliando, mas a maioria dos apicultores ainda trabalha artesanalmente e a demanda técnica a nível de produtor não atende ao crescimento da classe.

Estudo realizado com 570 apicultores de vários municípios do estado do Ceará (SEBRAE, 2006) revelou que os problemas ainda são básicos. Os produtores obtêm uma renda média baixa, girando entre R$ 120,00 a R$ 240,00 por mês e a baixa produtividade e qualidade dos produtos é decorrente de uso de tecnologia inadequada, falta de assistência técnica, descrédito junto às cooperativas, desarticulação de comunidade, presença de atravessadores, inadimplência junto à instituições financeiras, fomento insuficiente, baixa diversificação da produção, alto índice de perdas de enxames Similar problemática foi apresentada por Reis (2003) em Mato Grosso do Sul.

Neste perfil, a Apicultura ainda não se distingue das demais atividades agrárias familiares, ainda exclusas. Cerca de cinco milhões de famílias rurais vivem com menos de dois salários mínimos mensais (FIBGE, 2008), o que os destinam às formas rústicas de vida para sobreviver. Embora as novas tecnologias sejam de conhecimento dos muitos produtores, nem todos a adotam, muitas vezes por fatores socioeconômicos relacionados Certamente, a adoção das tecnologias preconizadas pode elevar os níveis de produtividade deste segmento agrário, beneficiando positivamente a economia. (Khan et al., 1991).

5.3 A Apicultura no Rio de Janeiro

A Apicultura do estado do Rio de Janeiro foi uma das pioneiras no Brasil, auxiliou entre as décadas de 40 a 60 a difusão desta cultura em todo o país, ao oferecer cursos, palestras e insumos (Rangel, 2006). A atividade no estado soma mais de 60 anos e seus entraves e avanços devem ser um alerta para outras regiões no atual cenário apícola.

O Rio de Janeiro é considerado um dos maiores centros consumidores de mel no país (Lorenzon et al., 2008). Em dez anos a classe produtora dobrou, mas a produção de mel, em torno de 400 toneladas, continua estagnada, favorecendo a importação de muitas marcas de méis de outros estados. Para o SEBRAE (2006), o estado do Rio apresentou uma alta devastação ambiental e índices muito pobres de suporte à agricultura familiar, fatores estes que contribuem para a improdutividade. Dentro do estado, a região da Costa Verde é uma das menos expressivas na produção apícola. Desta forma tal região merece atenção especial no que tange a comercialização dos produtos apícolas, para favorecer o seu marketing e estratégias de produção.

No próximo capítulo mostraremos os resultados obtidos a partir de pesquisa realizada nos municípios de Angra dos Reis e Mangaratiba no período entre janeiro e julho de 2007, onde foram obtidas as seguintes variáveis: preço do mel, tipo de embalagem, pureza do mel, cidade onde foi vendido o mel, bairro, inspeção feita no produto, tipo de estabelecimento e estado de origem do produto.

Capítulo 6

Análise dos Dados

Observando os resultados obtidos pela tabela 6.1 é possivel verificar que usando como base as variáveis de referência que representam, os méis sem aditivo, vendidos em farmácia, em ambalagem de plástico, de origem dos estado de Minas Gerais e comercializado no município de Angra dos Reis, chegamos a um preço médio de 1,99 R$ por 100g de mel.

Essas variáveis foram escolhidas para servirem como referência, pois apresentavam o maoir número de observações da amostra investigada. No processo de seleção de variáveis *stepwise*, foram retiradas as variáveis, bairro e inspeção.

Ainda nos referindo a tabela 6.1, podemos observar que as variáveis de referência, assim como o mel composto, vendido em supermercado de embalagem de vidro e origem dos estados de RJ e SP, apresentam maior significância (p-valor$\leq$ 0,05).

.

Tabela 6.1: Modelo de Regressão Linear Múltipla utilizando "stepwise"sobre o preço de 100g de mel.

Variáveis		Coeficiente(R$)	IC 95%(R$)	*p - valor*
Intercepto		1.99	[1.75 ; 2.23]	$\leq 0.05*$
Pureza do mel	Sem aditivo	-	-	-
	Composto	0.71	[0.47 ; 0.95]	$\leq 0.05*$
Tipo de Estabelecimento	Farmácia	-	-	-
	Supermercado	-0.45	[-0.71 ; - 0.19]	$\leq 0.05*$
	Mercado popular	-0.71	[-1.5 ; 0.08]	0.20
	Horti-Fruti	0.54	[-0.64 ; 1.72]	0.91
	Produtos Naturais	-0.41	[-0.91 ; 0.09]	0.62
Embalagem	Plástica	-	-	-
	Vidro	0.17	[-0.06 ; 0.40]	$\leq 0.05*$
Origem	MG	-	-	-
	RJ	0.77	[0.37 ; 1.17]	$\leq 0.05*$
	SP	1.56	[1.04 ; 2.08]	$\leq 0.05*$
	Outros	0.28	[-0.44 ; 1.00]	0.33
Municipio	Angra dos Reis	-	-	-
	Mangaratiba	-0.18	[-0.68 ; 0.32]	0.95

Fonte:Angra dos Reis e Mangaratiba. ERJ. 2008

Tabela 6.2: ANOVA do Modelo de Regressão.

Fonte de Variação	Soma dos Quadrados	Graus de liberdade
Composição	64.13389	1
Embalagem	1.96799	1
Tipo de Estabelecimento	22.31910	4
Origem	59.38124	1
Município	1.49566	3
Resíduos	242.23534	331
$R^2 = 0.3813$	$R^2_{ajustado} = 0.3626$	

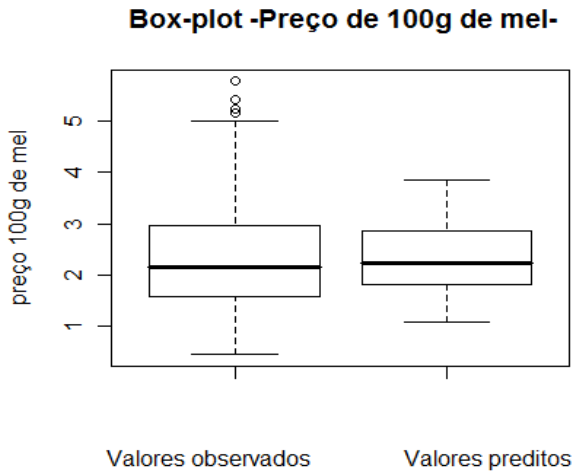
Fonte:Angra dos Reis e Mangaratiba. ERJ. 2008

.

A partir da tabela 6.2 podemos observar que nossa reta de regressão só explica 36,25% dos dados, que ainda pode ser melhorado.

Precisamos agora validar o modelo de regressão, a figura 6.1 apresenta dois box-plot com os valores observados do preço de 100g de mel o os valores preditos pelo nosso modelo, podemos observar que eles são bem parecidos.

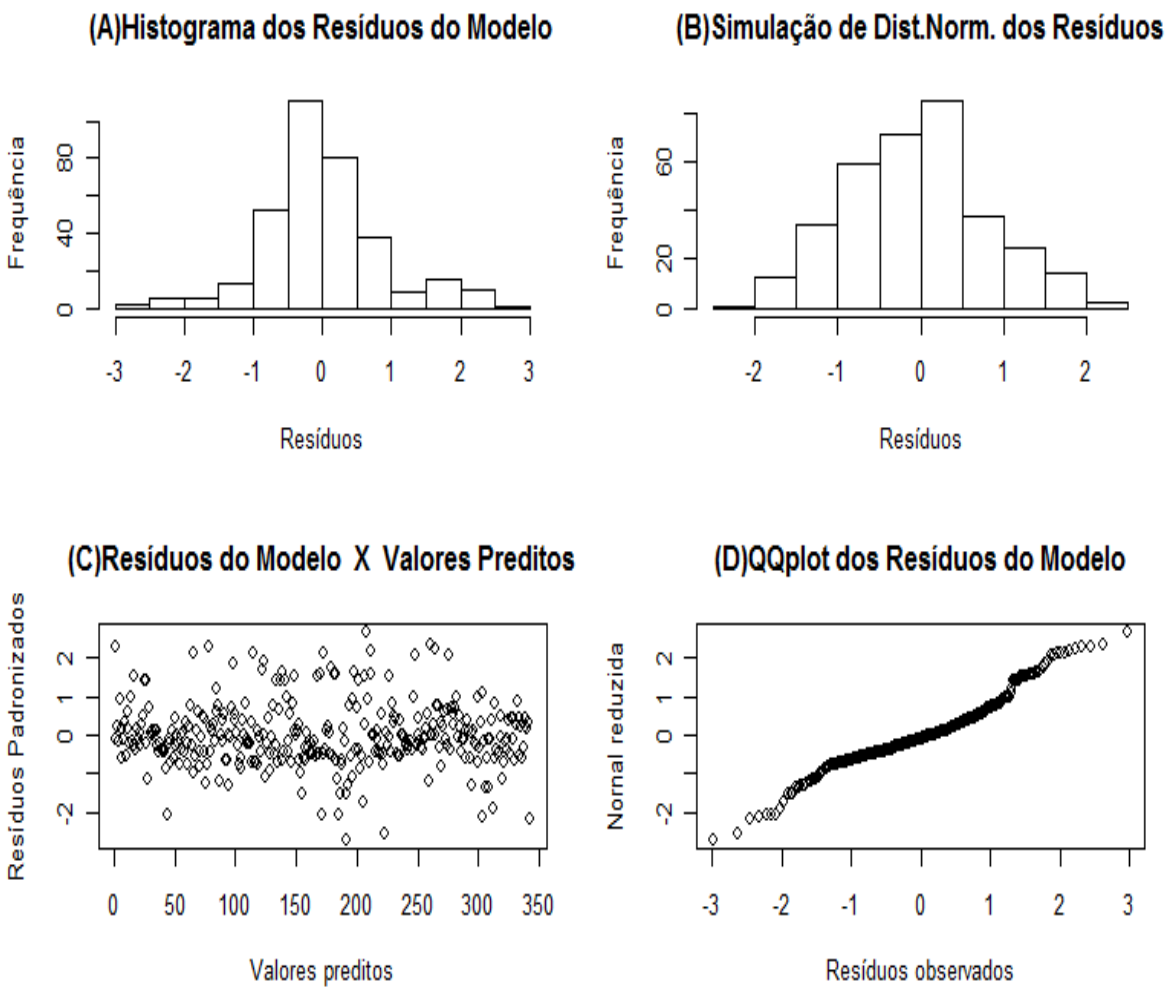
Figura 6.1: Box-plot



Fonte:Angra dos Reis e Mangaratiba. ERJ. 2008

Temos também que verificar se há normalidade na distribuição dos resíduos do modelo. A figura 6.2 (A) nos mostra o histograma dos resíduos do modelo, que é bem parecido com uma de distribuição normal. A figura 6.2 (B) simula uma distribuição normal a parti da média e desvio padrão dos resíduos do nosso modelo.Na figura 6.2 (C) podemos analisar como os resíduos se comportam ao longo do eixo das abcissas, onde a maioria dos pontos se encontram no intervalo (-2,2), nesta figura podemos observar também observações extremas (outliers). Na figura 6.2(D) analisamos o gráfico QQ-plot, onde observamos que a grande maioria dos pontos se encontra agrupados em formato de reta .

Figura 6.2: Análise de Resíduos.



Fonte:Angra dos Reis e Mangaratiba. ERJ. 2008

Capítulo 7

Conclusão

Hilmi (2005) em "O marketing de produtos apícolas"explica que os preços do mel podem mudar frequentemente. Por exemplo, os preços variam em consequência de alterações na produção e demanda em diferentes épocas do ano. O fornecimento de produtos apícolas muitas vezes varia de tempos em tempos, devido às condições meteorológicas,variação da área da flora as doenças das plantas, doenças das abelhas, problemas no deslocamento do mel ao mercado consumidor e muitos outros fatores, que afetam por exelência o fornecimento no mercado.

Referente a influência de cada variável no preço do mel, notamos que o Mercado Popular é onde se encontra o mel mais barato, isso pode está acontecendo pelo fato de não haver atravessadores neste tipo de comércio. Agora levando em consideração os municípios, observamos que o mel vendido em Mangaratiba é quase 20 centavos por 100g mais barato do que o vendido em Angra dos Reis, explica-se este fato por ser de Angra dos Reis um município

onde há grande concentração de turistas, elevando assim o poder aquisitivo da população . Já o mel vindo do estado de São Paulo, se destaca como sendo o mais caro, cerca de 1,56 R$ em referência ao de origem de Minas Gerais que tem maior oferta no Rio de Janeiro.

O tipo de embalagem do mel Vidro obteve um acréscimo de R$ 0,52 em relação à referência (plástico), por ser um material de maior custo, mas em compensação de maior higiene (Barreto et al., 2006; Costa et al., 2005)

O modelo de regressão nos permite avaliar quanto cada variável influência no preço final do mel, contribuindo para que os consumidores possam escolher o local e quais são as características que fazem o mel ter um preço menor. O produtor, dependendo de seu objetivo, pode fazer uma combinação dessas características para projetar o escoamento de sua produção.

São várias as suposições que submetemos o modelo de regressão e o analista deve sempre duvidar da validade dessas suposições e conduzir análises para encaminhar a adequação do modelo que está sendo testado. Neste sentido que foram adotados alguns gráficos (QQ-plot, Histograma dos resíduos, Resíduos padroninados versus Valores previstos) que nos ajudaram à aceitar como verdadeiras essas hipóteses

A análise de regressão é uma técnica largamente utilizada nos tempos atuais, que tem sido difundida vastamente entre os mais diversos campos da ciência. Esta trabalho também enfoca a interdisciplinariedade, no trabalho de pesquisadores de diferentes areas do conhecimento, o que proporciona a troca de experiência entre estes.

Deixa-se como proposta para um novo trabalho a sugestão de se utilizar outros tipos de modelo de transformações não lineares e/ou outras familias de modelos, como: modelos lineares generalizados (GLM), Modelo de equações silmutaneos(Fahrmeir & Tutz,2001), modelos aditivos generalizados(McCullag & Nelder, 1989), entre outros, podendo se obter um melhor ajustamento.

Referências Bibliográficas

[1] AKAIKE, HIROTUGU **A New Look at the Statistical Model Identification.** *Ieee Transaction on Automatic Control,*,1974.

[2] BUSSAB,W. O.; MORRETTIN,P. A. **Estatística Básica.** 5ª ed. São Paulo: *Saraiva*, 2002.

[3] BARRETO, L. M. R. C.; PEÃO, G. F. R.; DIB, A. P. da S.**Higienização e Sanitização na produção Apícola** Taubate *Cabral Editora*, 2006.

[4] CHARNET, R. et al. **Análise de Modelos de Regressão Linear com Aplicações.**Campinas/SP *Editora da Unicamp*,1999.

[5] DANCEY, C. P.; REIDY, J. **Estatística sem Matemática para Psicologia: usando SPSS para Windows. [Tradução VIALI, L.].** 3ª ed. Porto Alegre: Artmed, 2006.

[6] DE JONG D.; GONÇALVES L.S. **The africanízed bees of Brazil have become tole rant to varroa**. *Apiacta*, v.33,, 1992.

[7] DEVORE, J. L. **Probabilidade e Estatística: para Engenharia e Ciências. [Trad. SILVA, J. P. N.].** São Paulo: *Pioneira Thomson*

Learning, 2006.

[8] FAHRMEIR, L. AND TUTZ, G. **Multivariate Statistical Modelling Based on Generalized Linear Models.** 2nd edition*Springer-Verlag*,2001.

[9] FARIAS A. A., SOARES J. F.; CÉSAR C. C. **Introdução á Estatística 2 ed.**Rio de Janeiro/RJ,*LTC*, 2003.

[10] FIBGE (Fundação Instituto Brasileiro de Geografia). **Censo Agropecuário: Rio de Janeiro, 1990-2005.** *http: // www. sidra. ibge. gov.br/.* Acesso em: 28 jan. 2005.

[11] GONÇALVES, L.S. **Genetic improvement of Apis mellifera and tecnological developments in Brazil.** *Apiacta*, v.15, 1994

[12] GUELBER, V.M.O; KERR, W.E.. **Influência da troca de rainhas entre colônias de abelhas africanizadas na produção de pólen.** *Bioscience Journal*, v.22, 2005.

[13] GUJARATI, D. N. **Econometria Básica, 3 ed.** São Paulo/SP,*Markron Books*, 2000.

[14] HILMI, M. **The Marketing of bee products**,*http: // www.beekeeping.com* . Acesso em 20/10/2008.

[15] HOCHHEIM, N. **Avaliações por inferência estatística** Florianópolis: *IBAPE*, 1999.

[16] HOFFMAN, R. **Estatistica para Economistas, 4 ed**.São Paulo/SP ,*Pioneira Thomson Learning*, 2006.

[17] KHAM, M.V. introduction of the african bees (Apis mellifera adansonii) into Brasil and some comments on their spread in South America. *American Bee Journal, v.114*,1974.

[18] LEVIN, J. **Estátistica Aplicada a Ciencias Humanas, 2 ed.** São Paulo/SP,*Harbra*, 1987.

[19] LINS, M. P. E.; MOREIRA. M. C. B.**Método I-O Stepwise para Seleção de Variáveis em Modelos de Análise Envoltória de Dados** *Pesquisa Operacional*, 1999.

[20] LORENZON, M C.A: PEIXOTO, E. T., GONÇALVES, E.B. **. Censo Apícola do estado do rio de Janeiro.** Rio de Janeiro. *SESCOOP,*,2008.

[21] MARTINS, G. A.**Estatistica Geral e Aplicada 3 ed.**- São Paulo : *Atlas*,2005.

[22] MCCULLAGH, P. NELDER, J.A.**Generalized Linear Models. Chapman and Hall** Oxford, 1989.

[23] MONTGOMERY, D. C.; RUNGER, G. C.**Estatística e Probabilidade para Engenheiros** 2ª ed. Rio de Janeiro: *LTC*, 2003.

[24] MUSTAFA C.; ALPHAY F.; MICHAEL T. R. **Inflation, price dispersion, and market structure**. *European Economic Review* , 2008.

[25] R DEVELOPMENT CORE TEAM.R **Foundation for Statistical Computing**, Vienna, AUS,2000.

[26] SCUDINO, P. A.**A utilização de alguns testes estatísticos para a análise da variabilidade do preço do mel nos municípios de Angra dos Reis e Mangaratiba**Monografia de licenciatura em Matemática*UFRRJ*, 2008.

[27] SEBRAE. **Desafios da Apicultura brasileira** *Revista SEBRAE Agronegócio n.324*,,2006.

[28] SIEGEL, S.; CASTELLAN JR, N. J. **Estatística não-paramétricapara ciências do comportamento**; [Tradução: CARMONA, S. I. C.], 2ªed. Porto Alegre: *Artmed*, 2006.

[29] SPIEGEL, M. R. **Estatística**. [Tradução: CONSENTINO, P.] (Coleção Schaum), São Paulo: Makron Books, 1993.

[30] STEPHENS, M. A.**FED para a Bondade de Estatística Fit e algumas comparações**,*Jornal da Associação Americana Estatística* Vol. 69 1974.

[31] VENABLES, W. N. AND RIPLEY, B. D**Modern Applied Statistics with** *S. Fourth edition.* Springer,2002

ANEXOS

ANEXO 1

```
      ######  SALVANDO O PROGRAMA ###########

save.image('bancojoao.RData')

load('bancojoao.RData')

######## CARREGANDO E EDITANDO BANCO##############


banco = read.table('dado17042008.csv', sep="\t", header=T)

ls()

edit(banco)

summary(banco)

str(banco)

attach(banco)

banco = subset(bancosemml, preco100grama < 6 & preco100grama > 0.20)

### TRANSFORMANDO AS VARIAVEIS EM FACTOR  ##############

######### COMPOSICAO  ###################

banco$composicaofat = factor(banco$composicao, labels=c("Mel Puro",
"Mel Composto"), exclude=NA)
banco$composicaofat= relevel(banco$composicaofat , ref="Mel Puro")

########## EMBALAGEM  ####################

banco$embalafat = factor(banco$embala, labels=c("Vidro", "Plastico"),
exclude=NA)
banco$embalafat= relevel(banco$embalafat , ref="Plastico")



########## INSPECAO  #################
```

```
banco$inspecaofat = factor(banco$inspecao, labels=c("Sif", "Sie" ,
"Sie-er" , "Sie-rj"  , "Visa sim" ,"Nao inspecionado"), exclude=NA)
banco$inspecaofat= relevel(banco$inspecaofat , ref="Sif")

# Agregando
banco$inspecao2=banco$inspecao
banco$inspecao2[banco$inspecao >= 2] = 2
banco$inspecaofat2 = factor(banco$inspecao2, labels=c("Sif", "Sie"),
 exclude=NA)
banco$inspecaofat2= relevel(banco$inspecaofat2 , ref="Sif")

########## BAIRRO  ###################

banco$bairrofat = factor(banco$bairro, labels=c("Centro de Angra dos Reis",
 "Japuíba","Frade", "Pereque","Bracuí","Camorim","Monsoaba",
"Jacuenganga","Village","Centrode Mangaratiba","Itacuruça","Muriqui",
"Conceicao do Jacareí"),exclude=NA)
banco$bairrofat= relevel(banco$bairrofat , ref="Centro de Angra dos Reis")

########## ORIGEM  ###################

banco$origemfat = factor(banco$origem, labels=c("MG", "RJ", "SP","ES",
 "SC","RN","PE","CE"), exclude=NA)
banco$origemfat= relevel(banco$origemfat , ref="SP")

# Agregando
banco$origem2=banco$origem
banco$origem2[banco$origem >= 4] = 4
banco$origemfat2 = factor(banco$origem2, labels=c("MG", "RJ",
 "SP","outros"), exclude=NA)
banco$origemfat2= relevel(banco$origemfat2 , ref="MG")


########## MUNICIPIO  ###################

banco$municipfat = factor(banco$municip, labels=c("Angra dos Reis",
 "Mangaratiba"), exclude=NA)
banco$municipfat= relevel(banco$municipfat , ref="Angra dos Reis")

########## TIPO DO ESTABELECIMENTO  ###################

banco$Tipoestabfat = factor(banco$Tipoestab, labels=c("Farmacia", "Feira",
"Supermercado","Horti-Fruti", "Produtos Naturais"), exclude=NA)
banco$Tipoestabfat= relevel(banco$Tipoestabfat , ref="Farmacia")
```

```
summary(banco)

###### MODELO COM AS VARIAVEIS EM FATORES ##############

modelofull1= (lm(preco100grama ~  composicaofat + embalafat  +
Tipoestabfat + municipfat + origemfat2  + inspecaofat2 , data=banco))


modelofull2= step(lm(preco100grama ~  composicaofat + embalafat  +
Tipoestabfat + municipfat + origemfat2  + inspecaofat2 , data=banco))

modelofull1

modelofull2

summary(modelofull2)

aov(modelofull2)


######  TESTES DE NORMALIDADE DOS RESIDUOS #######

## shapiro-wilk###
shapiro.test(resid(modelofull2))

####  wilcoxon  #####

wilcox.test(resid(modelofull2))


#### GERANDO ALGUNS GRAFICOS ######################

boxplot(banco$preco100grama, modelofull2$fitted.values,
 main = " Box-plot -Preço de 100g de mel-",
ylab="preço 100g de mel",xlab = "Valores observados x  Valores preditos" )

hist(modelofull2$fitted.values)

plot(fitted(modelofull2),resid(modelofull2))

hist(resid(modelofull2),main="Histograma dos Resíduos do Modelo",
ylab="Frequência",xlab="Resíduos")

qqnorm(resid(modelofull2),main="QQplot dos Resíduos do Modelo",
ylab="Nornal reduzida", xlab="Resíduos observados")
```

```
plot(resid(modelofull2), main="Resíduos do Modelo  X  Valores Preditos",
ylab="Resíduos",xlab="Valores preditos")

### SIMULANDO A DISTRIBUIÇÃO NORMAL ###

mean(resid(modelofull2))

sd(resid(modelofull2))


dist.normal = rnorm(342,2.280882e-17,0.842833)

hist(dist.normal,main="Simulação de Dist.Norm. dos Resíduos",
ylab="Frequência",xlab="Resíduos")
```