UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Tathiana Leite de Sousa

Modelos de regressão aplicados a dados de DAP do Estado da Bahia

Campina Grande - PB

Abril de 2016

Tathiana Leite de Sousa

Modelos de regressão aplicados a dados de DAP do Estado da Bahia

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Estatística Aplicada do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de especialista em Estatística.

Orientador: Tiago Almeida de Oliveira

Campina Grande - PB

Abril de 2016

R696m Rodrigues, Tathiana Leite de Sousa
Modelos de regressão aplicados a dados de DAP do estado da Bahia [manuscrito] / Tathiana Leite de Sousa Rodrigues. - 2016.
39 p.

Digitado.
Monografia (Estatística Aplicada) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2016.
"Orientação: Prof. Drº Tiago de Almeida de oliveira, Departamento de Estatística".

1. Modelo linear. 2.Transformação Potência. 3.Territórios da Bahia. 4. Regressão logística. I. Título.

21. ed. CDD 519.72

Tathiana Leite de Sousa

Modelos de regressão aplicados a dados de DAP do Estado da Bahia

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Estatística Aplicada do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de especialista em Estatística.

Trabalho aprovado em 22 de abril de 2016.

BANCA EXAMINADORA

Tiago Almeida de Oliveira

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Elias Dias Coelho Neto

Prof. Dr. Elias Dias Coelho Neto
Universidade Estadual da Paraíba

DMEsteves

Profª. Drª. Divanilda Maia Esteves
Universidade Estadual da Paraíba

Á Minha mãe, Graça e ao meu Pai, Benedito, por
me ensinarem que o conhecimento é o unico
"bem intangível"que o tempo não consome.
Ao meu esposo, Pedro Rodrigues, pelo incentivo
e apoio nos momentos mais difíceis.

Agradecimentos

Em primeiro lugar, agradeço a Deus pela sua infinita misericórdia, concedeu-me graça, disposição, inteligência, conhecimento, força, saúde, enfim, tudo proporcionou para que eu chegasse onde cheguei, pois sem ELE eu nada seria.

Aos meus pais Benedito Araujo de Sousa e Maria das Graças Leite de Sousa, que deram de todo seu amor, de toda uma base familiar e ética, para que eu pudesse crescer e ser alguém com sabedoria e humildade.

Ao meu marido Pedro Rodrigues Batista pelo apoio, que sempre me incentivaram na busca e realização dos meus sonhos.

Ao professor orientador Tiago Almeida de Oliveira pela dedicação oferecida para realização do presente trabalho, e a Professora Ana Patricia Basto Peixoto pela a atenção e dedicação concedida para ajudar na realização deste trabalho.

A todos os professores(as) pela contribuição no meu aprendizado ao passar pela especialização.

"Para se ter sucesso, é necessário amar de verdade o que se faz.
Caso contrário, levando em conta apenas o lado racional, você simplesmente desiste.
É o que acontece com a maioria das pessoas."
(Steve Jobs)

Resumo

Neste trabalho foi estudado a análise de regressão linear múltipla que é um método utilizado para conhecer os efeitos que algumas variáveis exercem sobre outras. Para o estudo foram utilizados base de dados e sistema de informações do censo agropecuário 2006, do IBGE, do Banco Central (BCB), do Ministério de Desenvolvimento Agrário (MDA), da Empresa Baiana de Desenvolvimento Agrícola (EBDA). Nesse aspecto, adotou-se como delimitação espacial 7 Territórios da Bahia. Estes Territórios baianos estão localizados no semiárido, isto se dá não somente possuírem baixo índice do desenvolvimento humano (IDH), mas também devido a sua localização na região semiárida do Estado. Com o objetivo de ajustar o modelo de regressão linear múltipla que melhor se adequasse aos dados aplicou-se primeiramente um teste de adequação de distribuição com o uso do pacote (vcd) do software R, em que se é calculado os valores ajustado segundo a distribuição discreta de interesse. Ajustou-se o modelo de regressão linear múltipla e posteriormente fez o teste e gráficos para verificar pressuposições do modelo (Shapiro-Wilks para Normalidade; Durbin-watson para independência). Realizou-se também a transformação de Box-cox chegando-se ao modelo final com efeitos significativos de Ano, Grupo e Recursos.

Palavras-chaves: Modelos Lineares; Transformação Potência; Territórios da Bahia.

Abstract

In this work it was studied a linear regression analysis Multiple. That is a method used to know that effects some variables exert about other. For the Study were used base data and Census Information System Agricultural 2006, to IBGE, Central Bank (BCB), make Ministry of Agrarian Development (MDA), the Bahian Agricultural Development (EBDA). In this aspect, it was adopted as spatial boundaries Territories 9 of Bahia. These nine Bahian Territories, seven are located in the semiarid region, Give This is not only possess Low Human Development Index (HDI), but Also because of its location in the semiarid region of the state. With the goal of adjusting the linear regression model Multiple what better would fit data applied first hum Distribution adequacy test with the Package use (VCD) of R software in See and calculated set values.According to Distribution discrete Interest. Set the linear regression model Multiple and later auditioned and Graphics paragraph assumptions of the model (Shapiro-Wilks paragraph Normality, Durbin-Watson paragraph independencia). Also held a Box-Cox transformation Coming up the final model with significant effects Year, Group, and Resources.

Key-words: Linear Model; transformation power; Bahia territorials.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A agricultura familiar (AF) não é uma categoria social recente, é um conceito genérico que incorpora uma diversidade de situações específicas e particulares. Fundada pela relação entre trabalho, propriedade e família. O instrumento utilizado que o identifica como agricultor, para ter acesso às políticas públicas, é chamado de DAP (Declaração de aptidão ao PRONAF).Os órgãos legitimados e responsáveis pela emissão da DAP no estado da Bahia são: EBDA (Empresa Baiana de Desenvolvimento Agrícola S.A), CEPLAC (Comissão Executiva do Plano da Lavoura Cacaueira) e STR (Sindicato dos Trabalhadores Rurais).

A ideia é estabelecer uma relação funcional entre variáveis, com o intuito de se predizer mudanças nos valores das variáveis que se estuda. A análise de regressão é um método utilizado para conhecer os efeitos que algumas variáveis exercem sobre outras. Até mesmo quando não existe uma relação casual entre as variáveis, elas podem se relacionar por meio de algumas expressões matemáticas, que são uteis para a estimação do valor de uma das variáveis, quando se tem conhecimento dos valores das outras variáveis (HOFFMANN, 2006).

Algumas vezes há interesse não apenas em saber se existe associação entre duas variáveis quantitativas X e Y , mas também em conhecer uma provável relação de causa e efeito entre variáveis. Deseja-se saber se Y depende de X. Neste caso, Y é chamado de variável dependente ou variável resposta e X é chamado de variável independente ou explanatória. A regressão é dita linear, quando considera-se que a relação da resposta às variáveis é uma função linear de alguns parâmetros. A análise de regressão linear simples é utilizada quando a predição da variável dependente é realizada em apenas uma variável independente, enquanto a análise de regressão linear múltipla diz respeito à predição da variável dependente com base em duas ou mais variáveis independentes.

O objetivo desse trabalho é ajustar o modelo de regressão linear múltipla que melhor se adequasse aos dados para o números de contratos efetuados por meio do PRONAF B (crédito rural via DAP) com as variáveis explicativas Ano, Grupo e Recursos.

2 Fundamentação Teórica

Encontram-se nesta seção as principais metodologias que servirão de base para este trabalho, no que se refere aos métodos de análise de regressão linear múltipla.

2.1 Regressão Múltipla

De acordo com Gujarati e Porter (2011) e Montgomery e Runger (2003), a análise de regressão é uma técnica estatística que se ocupa do estudo de dependência de uma variável (dependente) em relação a uma ou mais variáveis (independentes ou explicativas). O objetivo principal deste modelo é estimar e ou prever a média(da população) ou o valor médio da variável dependente em relação aos valores conhecidos (ou fixo) das variáveis independentes.

A análise de regressão é um dos modelos mais usados, sobretudo para fazer previsões. Por isso, excessos são comuns em suas utilização. Embora trate da dependência de uma variável em relação as outras variáveis, ela não implica necessariamente em modelos causais. Deve-se, portanto, tomar cuidado na seleção das variáveis que serão utilizadas para construir a equação de regressão e para determinar a forma do modelo, pois é muito difícil modelar uma real relação de causa e efeito (GUJARATI e PORTER, 2011; MONTGOMERY e RUNGER,2003)

2.1.1 Modelo teórico

O resultado final de uma Regressão Múltipla (RM) é uma equação da reta que representa a melhor predição de uma variável dependente a partir de diversas variáveis independentes. Esta equação representa um modelo aditivo, no qual as variáveis preditoras somam-se na explicação da variável critério. A equação da regressão linear pode ser representada por: $y = a + bxi + \hat{\text{I}}$ em que, y é a variável dependente, ou critério; a é a constante, ou o intercepto entre a reta e o eixo ortogonal; "b" é o parâmetro, coeficiente padronizado de regressão, ou peso; xi são as variáveis independentes (preditoras) e Î é o erro ou resíduo, que se refere à diferença entre os valores observados e preditos. O formato geral da equação de regressão linear múltipla é

$$y = a + b_1x_1 + b_2x_2 + ... + b_kx_k + \hat{\text{I}}.$$

em que, Y é a variável dependente; a corresponde ao coeficiente técnico fixo, a um valor de base a partir do qual começa Y; b_k corresponde aos coeficientes técnicos atrelados às variáveis independentes; e X_k as variáveis independentes; Î é o erro do modelo.

Em notação matricial, o modelo de regressão linear múltipla pode ser escrito na forma:

$$Y = X\beta + \varepsilon$$

sendo,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \; e \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

em que, Y é um vetor, de dimensão $n \times 1$, da variável aleatória Y, X é a matrix de dimensões, $n \times p$, denominada matriz do modelo, β é o vetor, de dimensão $p \times 1$, de parâmetros desconhecidos, ε é o vetor, de dimensão $n \times 1$ de variáveis aleatórias não observáveis. Tal representação simplifica a notação e os cálculos a serem realizados futuramente (HOFFMANN, 2006).

2.1.2 Pressuposições para modelo

Para o ajuste de um modelo de regressão linear múltipla é necessário seguir as seguintes pressuposições:

i) A variável Y é função linear das variáveis explicativas X_j, $j = 1; 2; \ldots; k$;

ii) Os valores das variáveis explicativas X_j são consideradas fixas;

iii) $E(\varepsilon_i) = 0$, ou seja, $E(\varepsilon) = 0$, sendo o 0 um vetor de zeros dimensão $\eta \times 1$;

iv) Os erros são homocedásticos, isto é, $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$;

v) Os erros são independentes, isto é, $Cov(\varepsilon_i; \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0, i \neq j$;

vi) Os erros têm distribuição normal.

A suposição de normalidade dos erros não utiliza-se na estimação, e sim na elaboração dos testes de hipóteses e na obtenção de intervalos de confiança.

2.1.3 Estimação dos Parâmetros

Considerando todos os pressupostos, o estimador dos coeficientes de regressão (representado por $\widehat{\beta}$, conforme Rencher e Schaalje (2007), é aquele que minimiza a soma de quadrados dos erros (que representa os desvios das observações da variável resposta em

relação ao correspondentes valores preditos pelo modelo), a qual é obtida pelo método dos minimos quadrados ordinários, dado por:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon$$

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})$$

$$\varepsilon_i^2 = [y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})]^2$$

$$Ł = \Sigma \varepsilon_i^2 = \Sigma (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_k x_{ik})^2$$

Derivando L em função dos β's obtemos

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}],$$

$$\frac{\partial L}{\partial \beta_j} = -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}] x_{ji}, \quad j = 1, 2, \ldots, p.$$

Igualando as derivadas parciais a zero e substituindo $\beta_0, \beta_1, \ldots, \beta_p$ por $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$, temos o sistema de equações

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \widehat{\beta}_p \sum_{i=1}^{n} x_{ip} = \sum_{i=1}^{n} Y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + \widehat{\beta}_p \sum_{i=1}^{n} x_{i1} x_{ip} = \sum_{i=1}^{n} x_{i1} Y_i \\ \vdots\vdots \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{ip} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{ip} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{ip} x_{i2} + \ldots + \widehat{\beta}_p \sum_{i=1}^{n} x_{ip}^2 = \sum_{i=1}^{n} x_{ip} Y_i. \end{cases}$$

Resolvendo este sistema, obtemos os estimadores de mínimos quadrados $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$ dos parâmetros do modelo em questão.

O método de mínimos quadrados tem como objetivo encontrar o vetor $\widehat{\beta}$ que minimiza

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (Y - X\beta)'(Y - X\beta) =$$

$$= Y'Y - Y'X\beta - \beta' X'Y + \beta' X'X\beta = Y'Y - 2\beta' X'Y + \beta' X'X\beta,$$

sendo que, $Y'X\beta = \beta'X'Y$ pois o produto resulta em um escalar. A notação X' representa o transposto da matriz X enquanto que Y' e β' representam os transpostos dos vetores Y e β, respectivamente. Usando a técnica de derivação (em termos matriciais) obtemos

$$\frac{\partial L}{\partial \beta} = -2X'Y + 2X'X\beta.$$

Igualando a zero e substituindo o vetor β pelo vetor $\widehat{\beta}$, temos

$$(X'X)\widehat{\beta} = X'Y.$$

Assim, o vetor $\widehat{\beta}$, obtido pelo Método dos Mínimos Quadrados Ordinários (MMQ), é dado por (FOX, WEISGERG,2010):

$$\widehat{\beta} = (X'X)^{-1}X'y,$$

cuja matriz de covariâncias é:

$$V(\widehat{\beta}) = \sigma^2(X'X)^{-1}.$$

O estimador não viciado da variância, S^2, é obtido com base no estimador de MMQ de $\widehat{\beta}$ sendo sua fórmula (RENCHER e SCHAALJE, 2007):

$$S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - X'\widehat{\beta})^2.$$

Considerando a Soma de Quadrados do Erro (SQE),

$$\sum_{i=1} n(y_i - X\prime\widehat{\beta})^2 = (y - X\widehat{\beta})'(y - X\widehat{\beta}) = y'y - \widehat{\beta}X'y = SQE,$$

é possível expressar s^2 através do Quadrado Médio do Erro (QME) por

$$QME = s^2 = \frac{SQE}{n-k-1}$$

2.1.4 Soma de quadrados

i) Soma de quadrados de resíduos (SQRes) - Para calcular soma de quadrados dos desvios, ou soma de quadrados residual, é necessário relembrar a Equação, em que:

$$\begin{aligned}\varepsilon'\varepsilon &= Y'Y - 2\beta'X'\,Y + \beta'\underbrace{X'X\beta}_{X'\,Y}\\ &= \;Y'\,Y - 2\beta'X'\,Y + \beta'X'\,Y\\ SQRes &= \;Y'\,Y - \beta'X'\,Y\end{aligned}$$

ii) Soma de quadrados total (SQTotal)- A soma de quadrado total, mede a variação total das observações em torno da média. Tem-se a expressão:

$$SQtotal = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}\; Y_i^2 - \frac{(\sum_{i=1}^{n}\; Y_i)^2}{n} = \;Y'\,Y - \frac{(\sum_{i=1}^{n}\; Y_i)^2}{n}.$$

iii) Soma de quadrados de regressão (SQReg) - A soma de quadrado de regressão, mede a quantidade de variação da variável dependente explicada pela equação de regressão linear múltipla. Então a expressão é definida por:

$$SQreg = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$$
$$= \sum_{i=1}^{n} \widehat{Y}_i^2 - \frac{\sum_{i=1}^{n} \widehat{Y}_i{}^2}{n} = \widehat{Y}'\ \widehat{Y} - \frac{(\sum_{i=1}^{n} \widehat{Y}_i)^2}{n}$$
$$= \left(X\hat{\beta}\right)' X\hat{\beta} - \frac{\sum_{i=1}^{n} \hat{Y}_i{}^2}{n} = \hat{\beta}' X' X\hat{\beta} - \frac{\sum_{i=1}^{n} \widehat{Y}_i{}^2}{n} = \widehat{\beta'} X' Y$$

então,

$$SQreg = \hat{\beta}'\, X' Y - \frac{\sum_{i=1}^{n} \widehat{Y}_i{}^2}{n}.$$

2.1.5 Teste de hipótese e intervalo de confiança

Segundo Queiroz (2011), após a estimação dos parâmetros, em geral, realizam-se testes afim de determinar se hipóteses realizada sobre tais parâmetros são suportadas por evidências obtidas por meio de dados amostrais. Ou melhor, é importante avaliar se existe uma boa correlação entre a variável resposta e a variável explicativa. Por exemplo, se o aumento da variável explicativa acarretará em uma mudança significativa ou não no valor esperado da variável resposta. Há dois testes que podem ser aplicados para verificar a tal mudança significativa, o teste t de Student e o F de Snedecor.

i) Teste de significância para o modelo de regressão (Teste F)

O teste F é utilizado para verificar se as variáveis independentes conjuntamente, contribuem significativamente para explicar a variação da variável resposta. Definindo-se as hipóteses

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_0 = 0 \\ H_1 : \beta_j \neq 0 \textit{ para pelo menos um } j = 1, 2, \ldots, k \end{cases}$$

A estatística teste será

$$F = \frac{QMreg}{QMres} \sim F_{(K, n-p),}$$

em que, k é o número de variáveis independentes e $p = k + 1$. Então, após encontrar o valor F calculado, o F tabelado, e atribuir o nível α de significância, pode-se decidir que, se $F_{calculado} > F_{tabelado}$, rejeita-se a hipótese H_0 e conclui-se ao nível α de significância que há regressão. Se $F_{calculado} < F_{tabelado}$, aceita-se a hipótese H_0 ao nível de significância e conclui-se ao nível α de significância que não há indícios de relação linear entre as variáveis.

Pode-se resumir o procedimento descrito em uma Tabela da Análise de Variância (ANOVA), conforme representado na Tabela 1.

Tabela 1 – Análise da variância para o modelo de regressão linear múltipla

Fonte de Variação	GL	SQ	QM	F
Regressão	K	SQ_{Reg}	SQ_{Reg}/k	QM_{Reg}/QM_{Res}
Resíduo	$n-p$	SQ_{Res}	$SQ_{Res}/n-p$	-
Total	$n-1$	SQ_{Total}	-	-

ii) Teste de significância para os coeficientes de regressão (Teste t-Student)

Muitas vezes é de interesse do pesquisador testar hipóteses acerca dos coeficientes de regressão, para determinar o potencial de cada regressor no modelo. Segundo Charnet et al. (2008), para medir a significância

das variáveis do modelo individualmente, para cada $j = 1; 2; ...; k$, testa-se

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 \ : \beta_j \neq 0 \end{cases}$$

Neste caso, a estatística do teste é

$$T = \frac{\widehat{\beta}_j - \beta_j}{S(\widehat{\beta}_j)} \sim t_{n-p} \ \ j = 1, 2, ..., k,$$

e a regra de decisão é se $t_{calc} > t_{tab}$, rejeita-se a hipótese H_0, e conclui-se ao nível α de significância, que a variável não pode ser eliminada do modelo, pois explica bem a regressão linear. Se $t_{calc} < t_{tab}$, aceita-se a hipótese H_0, e ao nível α de significância, conclui-se que a variável pode ser eliminada do modelo sem tanto dano para a explicação da regressão linear.

2.1.6 Intervalo de confiança para os coeficientes de regressão

Outra forma de se avaliar a significância dos parâmetros do modelo é por meio da construção de intervalos de confiança. Podendo-se encontrar um intervalo de confiança que contenha o verdadeiro valor do parâmetro β_j com $j = 1; 2; ...; k$, a um certo nível de significância, que se queira.

Considerando-se a estatística teste dada em um intervalo com $100(1-\alpha)\%$ de confiança para o coeficiente da regressão β_j, $j = 1; 2; ...; k$, é definido por

$$IC = \left[\widehat{\beta}_j - t_{(\frac{\alpha}{2}, n-p-1)}\sqrt{S(\widehat{\beta}_j)}; \widehat{\beta}_j + t_{(\frac{\alpha}{2}, n-p-1)}\sqrt{S(\widehat{\beta}_j)}\right]$$

2.1.7 O coeficiente de determinação

Segundo (REENCHER;SCHAALJE,2007), uma maneira de avaliar o poder de explicabilidade do modelo é através do coeficiente de determinação dado por:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}.$$

O R^2 reflete a proporção das variações na variável resposta explicada pelo modelo, assumindo-se valores entre zero e um, de modo que valores próximos a um indicam uma boa qualidade do ajuste do modelo aos dados.Segundo Reencher e Schaalje (2007) o coeficiente de determinação aumenta com a inclusão de variáveis regressoras, visto que, proporcionalmente, a soma de quadrados o erro diminui. Para o modelo de regressão múltipla, um ajuste na fórmula que penaliza o acréscimo indiscrimado de variáveis explicativas fornece o R^2 ajustado.

$$R^2_\alpha = \frac{(n-1)R^2 - k}{n-k-1}.$$

O R^2_α consiste numa importante indicação preliminar para a qualidade do ajuste e para a decisão de buscar novas variáves explicativas.Todavia, seus resultados devem ser relativados para cada caso e analisado de maneira conjunta com as técnicas de seleção de variáveis e de diagnóstico do modelo.

2.1.8 Seleção de variáveis

Uma importante questão relacionada à análise de regressão múltipla se refere a obtenção de um modelo que proporcione maior eficiência na explicabilidade da variável dependente sem adição indevida de variáveis independentes. Dentre os métodos de seleção de váriaveis, destacam se o maior R^2_P, o de menor s^2_p, *backward*, *forwards*, *stepwise* e o Critério de Akaike (AIC). Nenhum dos métodos dísponivel é consistente, segundo Paula (2004), de modo que,mesmo para grandes amostras, não selecionam uma variável explicativa com probabilidade 1.

O método *forward* inicia com o modelos mais simples, em que y é uma constante, isto é: $y = \beta_0$. Supondo que seja consideradas q variáveis explicativas, para cada uma, ajusta se o modelo:

$$y = \beta_0 + \beta_j x_j,\ j = 1, ..., q$$

seja p o menor nível descritivo dos q modelos ajustados para o teste de hipóteses:

$$H_0 :\ \beta_j = 0$$
$$H_1 :\ \beta_j \neq 0$$

a variável correspondente entra no modelo se $p < p_e$, em que p_e é um nível descritivo crítico, escolhido como critério de entrada. O processo ocorre de maneira iterativa, levando em consideração o ajuste com variáveis já selecionadas em passos anteriores (isto é, uma vez que uma variável seja selecionada para o modelo, está não será mais descartada), até que ocorra $p > p_e$ (PAULA, 2004).

O método *backward* ,parte do modelo mais completo, que inclui todas as q possíveis variáveis explicativas consideradas:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q.$$

Realiza-se o teste de significância dos coeficientes e, sendo p o maior p-valor dos q teste, retira-se a variável correspondente se $p > p_s$, em que p_s é o nível descritivo crítico escolhido como critério de descartes já realizados(de modo que nenhuma variável descartada pode ser reconsiderada), até que não haja descarte, quando verificar-se $p < p_s$ (PAULA 2004).

0 método *stepwise* é uma mistura dos dois procedimentos anteriores conforme (PAULA 2004), iniciamos o processo com o modelo $[y = \beta_0]$. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira não sai do modelo. O processo continua até que nenhuma variável seja incluída ou seja retirada do modelo. Geralmente adotamos $0,15 \leq P_E, P_S \leq 0,25$. Uma sugestão seria usar $P_E = P_S = 0,20$. (PAULA,2004,p.78)

O método AIC (*Akaike's Information Criterion*) foi introduzido em 1973 e se popularizou no meio acadêmico rapidamente pelas possibilidades de aplicações práticas. Como uma extensão do principio da máxima verossimilhança, propondo uma combinação entre estimação pontual e teste de adequação do modelo em um principio único de comparação entre os valores da log-verossimilhança média (DELEEUW,1992). Conforme Akaike (1973), a esperança da verossimilhança é dada por:

$$E(\log f(X/\widehat{\theta}) = E(\int f(x/\theta) \log f(x/\widehat{\theta}) dx),$$

em que, $\widehat{\theta}$ são estimadores do vetor de parâmetro θ da distribuição de probabilidade cuja densidade é

$$f(x/\theta)$$

sendo, X é uma variável aleatória que segue esta distribuição. Partindo da ideia de maximizar uma razão média de log-verossimilhanças para maximizar a entropia do modelo escolhido, chega-se a uma fórmula prática de cálculo do AIC, dada por

$$AIC_P = -2 \sum_{t=1}^{n} \log(L_P) + 2(p+1),$$

onde, L_P é uma função de verossimilhança do modelo; p o número de parâmetros; $K = 2$, para aplicações usuais do AIC ou $K = \log(n)$ (n representa o tamanho amostral), para abordagens bayesianas. O modelo a ser selecionado vai ser aquele com menor AIC.

2.1.9 Análise de resíduos no MRLM

Para que os resultados de uma análise de regressão sejam confiáveis, tanto no MRLS quanto MRLM, é fundamental que as suposições do modelo ajustado sejam válidas. Se as suposições são violadas, têm-se falhas sistemáticas, ou seja, não linearidade, não normalidade, heterocedasticidade, não independência dos erros, e presença de pontos atípicos, e então, levando-se á análises com conclusões duvidosas. Desta forma, a análise de resíduos desempenha um papel fundamental, pois oferece técnicas que nos ajudam a verificar a presença de indícios da adequabilidade do modelo por meio dos resíduos.

Então, o vetor de resíduos é definido por

$$\varepsilon = Y - X\beta$$

e lembrando alguns resultados importantes:

$$y \sim N(X\beta, \sigma^2\ I),$$

$$\widehat{\beta} \sim N(\beta, \sigma^2(X'\ X)^{-1}),$$

$$\widehat{y_i} = x_i'\widehat{\beta} \sim N(x_i'\beta, \sigma^2 x_i'(X'X)^{-1}x_i)$$

então, a esperança e a variância dos resíduos são definidas respectivamente por

$$E[\varepsilon] = E[Y - X\beta] = 0.$$

e

$$Var[\varepsilon] = var[Y - X\beta]$$

$$= \sigma^2[I - X(X'\ X)^{-1}X']$$

o que pode ser reescrito da seguinte forma:

$$\varepsilon \sim N(0, \sigma^2[I - X(X'X)^{-1}X']).$$

Segundo Hoffmann (2006), a matriz $X(X'X)^{-1}X'$ é considerada matriz de projeção H, a qual, é simétrica e idempotente e os valores da diagonal principal da matriz H são h_{ii}, com $0 < h_{ii} < 1$ e $i = 1, 2, ..., n$. Neste caso, h_{ii} é o valor observado da influência de x_i a $\bar{x}$.

Utiliza-se algumas técnicas para verificar as suposições do modelo, podem ser informais (como gráficos) ou formais (como testes), sendo que estes, são mais indicadas para a tomada de decisão. O ideal é combinar as técnicas disponíveis, para o diagnóstico de problemas nas suposições do modelo. Para cada suposição do modelo, descreve-se com detalhes as técnicas para diagnóstico.

2.1.9.1 Diagnóstico de normalidade

A normalidade nos resíduos é uma dedução muito importante para que sejam confiáveis os resultados a respeito do ajuste do modelo de regressão linear. Essa dedução pode ser verificada por:

i) Gráfico de probabilidade normal - Q-Q Plot - Quantil de probabilidade esperado para a distribuição normal, em função dos resíduos.

ii) Teste Shapiro-Wilk - Segundo Ferreira (2009) o teste de Shapiro-Wilk é baseado em estatísticas de ordem da distribuição normal e de seus respectivos valores

esperados. Supondo-se que a partir de uma população normal sejam retirada amostras aleatória de tamanho n, com $(X_1, X_2, ..., X_n)$, em que, os valores das amostras são ordenadas em forma crescente.

Testando-se as hipóteses

$$\begin{cases} H_0: & \text{A amostra provém de uma população Normal} \\ H_1: & \text{A amostra não provém de uma população Normal} \end{cases}$$

A estatística do teste de Shapiro-Wilk (1965), é representado pela seguinte expressão:

$$W = \frac{[\sum_{i=1}^{n} a_i X_{(i)}]^2}{(n-1)S^2}$$

em que, a_i é o melhor estimador linear não-viesado normalizado do valor esperado das estatísticas de ordem da distribuição normal padrão e $X_{(i)}$ os valores das amostras ordenadas de forma crescente $(X_{(1)}, X_{(2)}, ..., X_{(n)})$. Realiza-se o teste escolhendo o nível de significância.

2.1.9.2 Diagnóstico de homoscedasticidade

A falta de homoscedasticidade é chamada de heteroscedasticidade, ou seja, quando há heteroscedasticidade as variâncias não são constantes em diferentes observações, daí o modelo sofre alguns efeitos no seu ajuste. A suposição de homoscedasticidade é testada pelas técnicas a seguir:

i) Gráfico dos resíduos versus valores ajustados - Por meio de alguma tendência nos pontos pode-se identificar se há detecção de heteroscedasticidade da variância dos erros, se os pontos estão aleatoriamente distribuídos em torno do 0, sem nenhum comportamento, há indícios de que a variância dos resíduos é homoscedástica.

ii) Teste de Goldfeld-Quandt - A exigência do teste de Goldfeld-Quandt é de que a amostra seja relativamente grande.

Segundo Rodrigues e Diniz (2006), as n observações são ordenadas de acordo com os valores da variável regressora, divide-se a amostra ordenada em 3 partes, em que, a parte do meio deve ter 25% dos dados, a 1º contendo os menores valores da variável explicativa e a 3º contendo os maiores valores da variável explicativa, em que deve-se apresentar praticamente a mesma quantidade de dados. De posse dessas três partes, ajustam-se dois modelos de regressão, um com os dados da 1º parte e outro com os dados da 3º parte.

Enfim, utiliza-se o teste F, testando-se as seguintes hipóteses

$$\begin{cases} H_0: \sigma_1^2 = \sigma_3^2; \\ H_1^2: \sigma_3^2 > \sigma_1^2 \end{cases}$$

em que, σ_1^2 com $i = 1, 2, 3$ são as variância dos resíduos dos três modelos de regressão.

A estatística de teste é definida por

$$F_{GQ} = \frac{SQ_{Res}/(n_3 - (p+1))}{SQ_{Res^1}/(n_1 - (p+1))},$$

em que, SQ_{Res^1} e SQ_{Res^1} , são as somas de quadrados dos resíduos dos modelos de 1º e 3º parte, n_1 e n_3 são os números de observações da 1º e 3º parte dos valores da variável regressora e p o número de observações da 2º parte. O $F_{tabelado} = F_{(n3(p+1)),n1-(p+1))}$, então rejeita-se a hipótese nula se $F_{GQ} > F_{(\alpha)}$.

2.1.9.3 Diagnóstico de independência

Independência dos erros é um acontecimento aleatório que ocorre em um determinado período de tempo, sendo, um resíduo não afeta nos resíduos seguintes. Esse diagnóstico é verificado da seguinte forma:

i) Gráfico dos resíduos versus a ordem de coleta - Ao avaliar o gráfico e perceber alguma tendência nos pontos, ou seja, se os pontos repetem-se em um determinado ambiente do gráfico há indícios de dependência dos resíduos.

ii) Teste de Durbin-Watson- De acordo com Montgomery (2003) o teste de Durbin-Watson é utilizado para a detectar a presençaa de autocorrelação nos resíduos de um modelo de regressão. Testa-se a presença de autocorrelação por meio da hipótese

$$\begin{cases} H_0 : \rho = 0; \\ H_1 : \rho \neq 0. \end{cases}$$

A estatística teste é representada por:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

em que, $e_i = y_i - y_i$ é o resíduo associado a i-ésima observação. O valor da estatística deve variar de $0 \leq DW \leq 4$. Para a tomada de decisão compara-se o valor da estatística DW com os valores críticos D_L e D_U, daí toma-se a decisão de acordo com:

-Se $DW < D_L$ rejeita-se a hipótese $H_0 : \rho = 0$;

- Se $DW > D_L$ aceita-se a hipótese $H_o : \rho = 0$;

- Se $D_L < DW < D_U$ o teste é inconclusivo.

2.1.9.4 Diagnóstico de outliers

Outlier é uma observação com o comportamento diferente das demais. Desta forma, pode ser um outlier em relação a Y ou aos X, e pode ou não ser um ponto influente.

i) Outliers com relação a variável X

Para identificar um outliers em X, utilizam-se os valores h_{ii} da matriz de projeção, observa se há valor extremo do h_{ii} em um box-plot.

ii) Outliers com relação a variável Y

Os resíduos são definidos, para detectar melhor outliers na variável Y, os resíduos foram modificados por

i) Resíduos padronizados

O resíduo padronizado não tem boas propriedades, por não ter variância constante, muda cada valor de X_i.

Se os erros seguem uma distribuição normal, 95% dos resíduos padronizados devem estar no intervalo entre (-3,3), se não, podem indicar a presença de outlier. O resíduo padronizado é definido por

$$d_i = \frac{\varepsilon_i}{\sqrt{QM_{Res}}}, i = 1, 2, ..., n.$$

ii) Resíduos estudentizados

Os resíduos estudentizados tem variância constante e igual a 1, ajudando-se a encontrar com maior facilidade outliers. Desta forma, os resíduos estudentizados são definidos por:

$$r_i = \frac{\varepsilon_i}{\sqrt{QM_{Res}(1 - h_{ii})}}, i = 1, 2, ..., n.$$

Se após a realização da análise de resíduo, constata-se que não foi possível satisfazer presuposição para o modelo linear clássico, é possível que uma transformação não linear dos dados possa produzir a homogeneidade da variância e a distribuição aproximadamente normal dos resíduos.

A transformação Box-Cox identifica uma transformação a partir de uma família de transformação de potência de Y ,a fim de encontrar a transformação que estabilize ou reduza a varacibilidade existente e normalidade dos resíduos.

Box e Cox (1964) propuseram uma família e transformação definida por:

$$Y(\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, \lambda \neq 0; \\ log(\lambda), \lambda = 0. \end{cases}$$

em que, λ é o parâmetro de transformação e Y a variável resposta. Quando $\lambda = 1$ não é necessário a realização de transformação e quando $\lambda = 0$ utiliza-se a transformação logaritímica.

Nas observações $(Y_i; x_i^{'})$, $i = 1, 2, ..., n$ e $x_i^{'} = (X_{1i}, X_{2i}, ..., X_{ki})$, tem-se que

$$Y_i(\lambda) \sim N(x_i^{'}\beta, \sigma^2), i = 1, 2, ..., n.$$

Para escolha da melhor potência para λ consideram-se valores no intervalo de $[-2, 2]$, conforme descrevem Draper e Smith (1998). Se no gráfico da verossimilhança perfilhada o valor 0 estiver contido no intervalo, é indicado a utilização da transformação logaritímica da variável, pois os resultados serão bem próximos dos obtidos com a transformação préviamente adotada.

Ao realizar a transformação na variável Y, as estimações e predições são expressas em novas unidades, de acordo com cada transformação admitida. Portanto um problema que não pode ser esquecido é o retorno à escala normal.

Para facilitar esse retorno, Miller (1984) sugere o estimador $E[Y/x]$, definido por

$$E[Y/X] = exp(\hat{\beta_0} + \hat{\beta_1}X_1 + ... + \hat{\beta_K}X_K)exp(\frac{\hat{\sigma}^2}{2})$$

sendo, $\hat{\beta_0} + \hat{\beta_1}X_1 + ... + \hat{\beta_K}X_K$ é o ajuste do modelo na escala transformada e $\hat{\sigma}^2$ é o quadrado médio do resíduo também na escala transformada.

Se o valor de $\hat{\sigma}^2$ for pequeno, há uma outra linha de desenvolvimento, em que, Taylor (1986) propôs

$$E[Y/3] = exp(\hat{\beta_0} + \hat{\beta_1}X_1 + ... + \hat{\beta_K}X_K)exp(\frac{1+\sigma^2}{2}), \tag{2.1}$$

dado que, $\sigma^2 \approx 0$ os dois estimadores acima praticamente coincidirão.

Outras transformações são adotadas de acordo com a necessidade dos dados, algumas delas são destacadas a seguir:

i) Raiz quadrada ($\hat{Y} = \sqrt{Y}$) é utilizada para estabilizar a variância quando é proporcional a média do $Y^{'}$s. O estimador para $E[Y/x]$ proposto por Miller(1984), será:

$$E[Y/X] = (\hat{\beta_0} + \hat{\beta_1}X_1 + ... + \hat{\beta_k}X_K)^2 + \hat{\sigma}^2$$

ii) Transformação recíproca ($\hat{Y} - Y^{-1}$) é utilizada para estabilizar a variância, minimizando possíveis altos valores da variável Y. O estimador para $E[Y/X]$ proposto por Miller (1984), será

$$E[Y/X] = (\hat{\beta_0} + \hat{\beta_1}X_1 + ... + \hat{\beta_k}X_K)^{-1} + \frac{\hat{\sigma}^2}{(\hat{\beta_0} + \hat{\beta_1}X_1 + ... + \hat{\beta_k}X_K)^2}$$

2.1.10 Algumas transformações usuais

Quando após a análise gráfica dos resíduos constatamos a violação de uma ou mais suposições, algumas vezes é possível amenizar esse problema fazendo alguma transformação pertinente na variável resposta (Y). Considerando aqui algumas transformações possíveis que são normalmente usadas com o objetivo de estabilizar as variâncias dos erros e, em circunstâncias especiais, também para normalizar os dados. Porém estas transformações são indicadas para situações específicas, sendo que algumas delas só podem ser utilizadas se a variável Y assume somente valores positivos.

i) **log(Y)** Transformação logarítmica na base e: Essa transformação estabiliza a variância, quando esta tende a crescer à medida que Y também cresce. Em algumas situações pode também a ajudar a normalizar os dados. Se a relação entre X e Y é do tipo exponencial, essa transformação introduz uma relação linear entre X e a variável transformada, $log(Y)$.

ii) $\sqrt{Y}$ Transformação raiz quadrada: É usada para estabilizar a variância quando esta é proporcional à média dos $Y^{'}$s. Em particular, quando a variável Y for uma contagem, com distribuição Poisson, a variável transformada, $\sqrt{Y}$, pode ser considerada com distribuição normal.

iii) $\frac{1}{Y}$ Transformação recíproca: É usada para estabilizar a variância, no sentido de minimizar o efeito de possíveis valores muito altos de Y.

iv) Y^2 Transformação quadrática: Usada para estabilizar as variância, quando estas tendem a decrescer com a média dos $Y^{'}s$. Normaliza os dados quando os resíduos se mostram com assimetria negativa. Linearidade os dados quando esses tem uma relação curvilínea. Se a relação entre X e Y é do tipo curvilínea, essa transformação introduz uma relação linear entre X e variável transformada Y^2.

v) $arcsen\sqrt{Y}$ Transformação arcseno: Estabiliza a variância quando os dados são proporções.

Quando fazemos uma transformação na variável original Y, as estimações e predições estão expressas em novas unidades, conforme a transformação usada. Muitas vezes o nosso objetivo é fazer estimações e predições na escala original, portanto, este é o problema que não pode ser esquecido.

Vamos exemplificar esta situação com o uso da transformação log(Y). Se obtivermos um bom ajuste do MRLS para log(Y), temos uma reta que nos fornece estimativas dos valores esperados de log(Y), para valores da variável regressora. No entanto, suponha que precisamos obter estimativas dos valores esperados de Y e não log(Y). Intuitivamente,

pensamos em fazer a transformação inversa dos valores preditos, neste caso a função exponencial, mas precisamos investigar o que está ocorrendo.

Seja g uma função monótona não decrescente. Seja y_α o quantil α da variável aleatória Y, isto é, $Prob\{Y < y_\alpha\} = \alpha$. Então, $g(y_\alpha)$ é o quantil α da variável aleátoria $g(Y)$.

Prova:

$$Prob\{Y < y_\alpha\} = Prob\{g(Y) < g(y_\alpha)\} = \alpha.$$

Considere a transformação logarítmica. Supomos que MRLS se ajusta bem a log(Y). Assim, temos o modelo na escala transformada,

$$log(Y) = \beta_0 + \beta_1 X + \epsilon \quad \text{em que} \quad \epsilon \sim N(0; \sigma^2).$$

Vejamos o que ocorre com a variável original Y. O modelo acima implica

$$exp(log(Y)) = exp(\beta_0 + \beta_1 X + \epsilon) = exp(\beta_0 + \beta_1 X) exp(\epsilon).$$

Como ϵ tem distribuição normal, a distribuição de $exp(\epsilon)$ é lognormal. Pelo resultado acima, a mediana de $exp(\epsilon)$ é igual a $exp(mediana(\epsilon)) = exp(0) = 1$. Por resultados da teoria de probabilidade, a esperança de $exp(\epsilon)$ é dada por $exp(\sigma^2/2)$. Então, podemos explicitar a esperança e a mediana de Y, para dado valor x, de X.

Ou seja,

$$\begin{aligned} E[Y/x] &= E[exp(\beta_0 + \beta_1 x) exp(\epsilon)] \\ &= exp(\beta_0 + \beta_1 x) \underbrace{E[exp(\epsilon)]}_{exp(\sigma^2/2)} \\ &= exp(\beta_0 + \beta_1) exp(\sigma^2/2). \end{aligned} \tag{2.2}$$

e

$$\begin{aligned} mediana[Y/x] &= exp(\beta_0 + \beta_1 x) \underbrace{mediana[exp(\epsilon)]}_{1} \\ &= exp(\beta_0 + \beta_1 x). \end{aligned} \tag{2.3}$$

Note que, se σ^2 for uma valor próximo de zero , $exp(\sigma^2/2)$ será aproximadamente igual a 1 e a esperança e a mediana de Y/x coincidirão.

O problema agora está em como estimar E[Y/x]. Miller (1984) segere que usemos

$$exp(\widehat{\beta_0} + \widehat{\beta_1} x) exp(\widehat{\sigma^2}/2),$$

sendo $\widehat{\beta_0}$, $\widehat{\beta_1}$ e $\widehat{\sigma^2}/2$ do ajuste do modelo na escala transformada.

Por outra linha de desenvolvimento, Taylor(1986) propõe o estimador

$$exp(\widehat{\beta_0} + \widehat{\beta_1}x)(1 + \hat{\sigma}^2/2),$$

sob a suposição de que o valor de σ^2 seja pequeno. Observe que, se σ^2 for pequeno, os dois estimadores acima praticamente coincidirão. Estudos diversos mostram que o valor σ^2, a variância do erro do modelo na escala transformada, deve ser um valor bem proximo de zero, ou seja, o ajuste do modelo transformado, quando usamos a transformação logarítmica. O artigo de Miller (1984) sugere estimadores para E[Y/x], sob outras transformações:

i) para transformações $\sqrt{Y} \Rightarrow (\widehat{\beta_0} + \widehat{\beta_1}x)^2 + \widehat{\sigma^2}$

ii) para a transformação $Y^{-1} \Rightarrow (\widehat{\beta_0} + \widehat{\beta_1}x)^{-1} + \frac{\widehat{\sigma^2}}{(\widehat{\beta_0}+\widehat{\beta_1}x)^2}$

Para está última transformação, é necessário que $\sigma < (\beta_0 + \beta_1 x)^2$. Também é interessante notar que nos três tipos de transformações os estimadores de $E[Y/x]$ são definidos pela transformação inversa dos valores preditos na escala transformada com uma correção, e esta correção será tanto menor quanto menor for o valor observado de $\widehat{\sigma^2}$.

Duan (1983) propõe um estimador não paramétrico para $E[Y/x]$, após transformação $g(Y)$, com inversa g^{-1}. O modelo na escala transformada não exige normalidade do erro. O estimador proposto é

$$\frac{1}{n}\sum_{i=1}^{n} g^{-1}(\widehat{\beta_0} + \widehat{\beta_1}x + e_i),$$

sendo $\widehat{\beta_0}$ e $\widehat{\beta_1}$ do ajuste do modelo na escala transformada e e_i, $i = 1, ..., n$, os resíduos do modelo na escala transformada. quando faz-se uma transformação na variável Y, deve-se verificar o ajuste do modelo na escala transformada. Com isso, pode-se obter um intervalo de predição na escala original, simplesmente usando a transformação inversa nos limites dos intervalos de predição na escala transformada.

3 Material e Métodos

A pesquisa foi realizada com base nos dados advindos de agricultores familiares situados nos Territórios da Cidadania da Bahia (Figura 1), sendo eles: Velho Chico, Litoral Sul, Baixo Sul, Chapada Diamantina, Irecê, Sertão do São Francisco, Itaparica, Semiárido Nordeste II, Sisal. O recorte sugerido se dá por que os mesmos possuem baixo índice de desenvolvimento humano (IDH). Por conta do período de pior estiagem dos últimos 50 anos, esta pesquisa será realizada com os dados resultantes do período entre 2010 e 2013.

Para este trabalho foram utilizadas base de dados e sistemas de informações do Censo Agropecuário 2006, do IBGE, do Banco Central do Brasil (BCB), do Ministério de Desenvolvimento Agrário (MDA), da Empresa Baiana de Desenvolvimento Agrícola (EBDA), dentre outras. Nesse aspecto, adotou-se como delimitação espacial de análises os territórios da cidadania, pois os mesmos apresentam indicadores sociais menores em relação aos demais Territórios de Identidade da Bahia. Dos nove Territórios da Cidadania da Bahia, sete estão localizados no semiárido. Este recorte sugerido se dá não somente por conta dos mesmos possuírem baixo índice de desenvolvimento humano (IDH), mas também devido a sua localização na região semiárida do Estado. As áreas estudadas foram divididas em 7 territórios, Chapada (CH), Irecê (IR), Itaparica (IT), Semi-Árido (SA), Sertão do São Francisco (SE), Sisal (SI) e Velho Chico (VC). As variáveis respostas foram número de contratos e recursos e as Declaração de Aptidão ao Pronaf (DAP) ativas e desativadas durante o período de estudo, que abrangeu os anos de 2010 à 2013. Na tabela 2 tem-se uma prévisualização da disposição dos dados, em que as áreas estudadas foram renomeadas para valores quantitativos crescentes, bem como os anos.

Figura 1 – Mapa da Bahia, identificando os territórios pesquisados. (Fonte: Extraído de *www.seplan.ba.gov/modules/conteudo/conteudo.php?conteudo* = 17/ - Acessado em: 20 de março de 2016

Tabela 2 – Representação tabular dos seis primeiros valores do banco de dados.

Obs	Ncont	Ano	Recursos	Grupo
1	382,00	1,00	517120,00	1,00
2	44,00	1,00	57448,50	1,00
3	742,00	1,00	1745950,01	1,00
4	1104,00	1,00	1089150,00	1,00
5	517,00	1,00	1401441,11	1,00
6	313,00	1,00	1023803,34	1,00

Ncont: Número de Contratos

As análises estatísticas foram realizadas no software R (R Core Team, 2016). Primeiro se aplicou um teste de adequação de distribuição com o uso do pacote (vcd), em que são calculados os valores ajustados segundo a distribuição discreta de interesse e estima-se os parâmetros via método de máxima verossimilhança. A estatística de razão de verossimilhança é calculada, com seu Valor P. Ajustou-se um modelo de regressão linear múltipla e posteriormente foram feitos os testes e gráficos para verificar as pressuposições do modelo (shapiro-wilks para Normalidade e Durbin-watson para independência.). Realizou-se também a transformação de Box-cox por meio do pacote (MASS).

4 Resultados e discussão

Os dados de DAP foram analisados por meio de um modelo de regressão múltipla em que foram considerados o totail de recursos alocados para cada região, os anos que foram feitas estas alocações de recursos e as regiões (grupos) do estado da Bahia. Inicialmente apresenta-se na Figura 2, gráficos de dispersão de todas as variáveis envolvidas no estudo. Percebe-se que a relação via gráfico de dispersão entre o número de contratos e a quantidade de recursos é linear e crescente, indicando que quanto maior os recursos maior o número de contratos. Com relação aos anos nota-se que até o ano 3 (2012), houve aumento do número de contratos, com leve diminuição no ano de 2013. Os territórios do semi-árido (Grupo 4), Sertão do São Francisco (Grupo 5) e Velho Chico (Grupo 7), obtiveram maior quantidade de número de contratos.

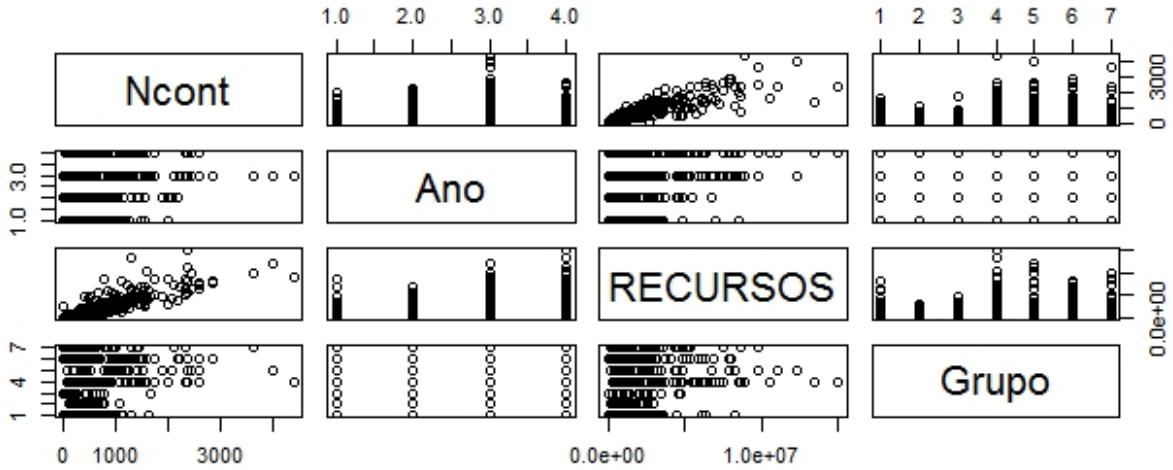

Figura 2 – *Scatterplot* dos dados de DAP, anos, recursos e regiões da Bahia.

Devido à natureza dos dados ser discreta (Número de Contratos) levantou-se a hipótese de estas observações poderiam ser consideradas seguir em uma distribuição de Poisson. Para tanto utilizou-se como ferramenta para tal análise as funções implementadas no pacote vcd (Meyer,Zeileis, Hornik, 2015). Por meio destas funções calculou-se os valores da distribuição discreta Poisson para o conjunto de dados amostrado (Número de contratos), o parâmetro da Poisson foi estimado por máxima verossimilhança e em seguida foi calculada a estatística de qui-quadrado de Pearson para se verificar se os dados pertenciam ou não a distribuição candidata, o valor P para o ajuste da distribuição Poisson aos dados de número de contratos foi <0,05 (Valor P=0,0001), indicando que esta distribuição não é adequada para estes dados. Para visualização dos dados foi utilizado o

método de visualização via rootograma das observações dos valores ajustados Figura 2.

Na Figura 3 de acordo com Kleiber e Zeileis (2014) o gráfico de rootograma compara os valores observados e os valores esperados pela visualiazação do histograma dos valores observados versus a curva das frequências ajustadas (raiz das frequências) segundo alguma distribuição de interesse.

Figura 3 – Rotoograma das observações dos valores ajustados.

Percebe-se pela Figura 3 que não há uma caracterização padrão da distribuição Poisson, bem como que as colunas do gráfico não foram justapostas indicando grande dispersão do número de ocorrências, de modo que a distribuição de Poisson não caracteriza bem este tipo de dados.

O modelo de regressão múltipla foi ajustado para o Número de Contratos com as variáveis explicativas Ano, Grupo e Recursos, em que as variáveis foram significativas segundo o teste F (valor P $< 0{,}05$), porém suas estimativas não são apresentadas devido ao fato que o teste de Shapiro-Wilk para os resíduos rejeitou a hipótese de normalidade dos resíduos (W $= 0{,}82$; Valor P $< 2,2 \times 10^{-16}$). E pelo teste de Durbin-Watson obteve-se os seguintes valores (DW $= 1{,}8662$, Valor P $= 0{,}03964$ em que a hipótese alternativa é que a verdadeira autocorrelação é maior que zero, ou seja, ausência de independência dos resíduos). Na Figura 4 percebe-se que os valores dos resíduos não estão perfeitamente em uma reta (quantil quantil), indicando a falta de normalidade.

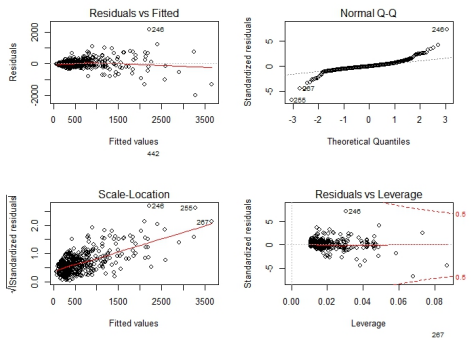

Figura 4 – Representação gráfica dos resíduos *versus* valores ajustados, quantil quantil da normal, resíduos *versus* leverage.

Devido à falta de independência dos resíduos, procedeu-se um ajuste de séries temporais dos resíduos, para ver se os mesmos poderiam ser modelados segundo um modelo autorregressivo integrado de médias móveis, modelo misto de Box-Jenkins (2015), Figura 5.

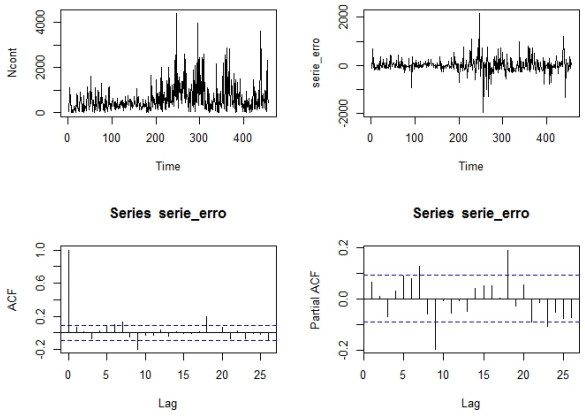

Figura 5 – Série temporal do Número de contratos, dos resíduos do modelo de regressão, função de autocorrelação e função de autocorrelação parcial dos resíduos do modelo de regressão multiplo.

Como pode ser ver pela Figura 5, não é possível estabelecer uma série temporal para modelar os resíduos do modelo de regressão. Sendo o modelo adequado para ajustar estes resíduos dado por um ARIMA (0,0,0), ou seja, sem presença de parâmetros autoregressivos, médias móveis e diferença por integração. Devido a isto, procedeu-se uma transformação ótima de Box-Cox para contornar os problemas encontrados nas suposições dos resíduos (Figura 6).

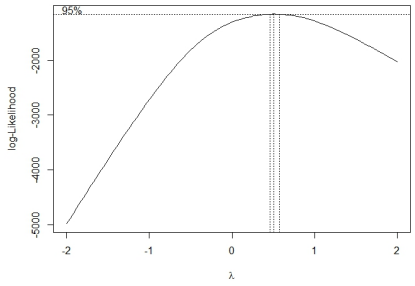

Figura 6 – Transformação box-Cox do número de contratos do modelo de regressão múltipla

De acordo com a Figura 6, o valor ótimo para transformação é $\lambda = 0,5$, o que sugere a transformação raiz quadrada para os dados. Após a transformação ótima de Box-Cox, realizou-se uma nova análise com a variável transformada. Na Tabela 3, tem-se as estimativas dos parâmetros do modelo ajustado com a variável transformada e pode-se

afirmar que a variável ano não foi significativa para o modelo, bem como o efeito do grupo 7.

Tabela 3 – Estimativas dos parâmetros, erros-padrão e Valor t do modelo para a variável transformada.

	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
(Intercepto)	13,9295	0,8595	16,21	0,0000
Ano	-0,2699	0,2625	-1,03	0,3044
grupo2	1,8211	0,8885	2,05	0,0410
grupo3	-0,6657	1,3392	-0,50	0,6194
grupo4	1,6279	0,9638	1,69	0,0919
grupo5	4,1120	1,1342	3,63	0,0003
grupo6	3,0300	0,8927	3,39	0,0007
grupo7	-0,3381	0,9537	-0,35	0,7231
Recursos	0,0000	0,0000	28,15	0,0000

O teste de Shapiro-Wilks para os resíduos rejeitou a hipótese de normalidade dos resíduos. (W = 0,96043, Valor P = $9,975 \times 10^{-}10$), porém o teste de Durbin-Watson não rejeitou a hipótese nula, (DW = 2,1186, Valor P = 0, 8284), indicando que a transformação de Box-Cox corrigiu uma das suposições do modelo de regressão. Na Figura 7 apresenta uma análise gráfica dos resíduos e por meio dela é possível verificar que houve melhora no ajuste da distribuição aos dados e que existe alguns pontos identificados que podem ser importantes e interferir na análise.

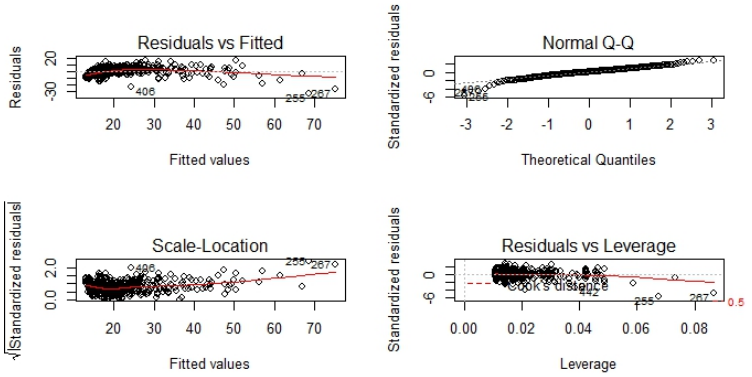

Figura 7 – Análise de resíduos para a variável transformada

Dado a necessidade de se verificar quais pontos poderiam ser considerados outliers, influentes, etc. Foi realizado um teste para verificar os pontos outliers.

Na Figura 8, tem-se um gráfico que mede influência de acordo com o tamanho do raio dos círculos visualizados no gráfico, percebe-se que alguns pontos que já foram considerados outliers na Tabela 4 aparecem na Figura 7, bem como alguns outros pontos. E nas Figuras 9 e 10 tem-se o gráfico de cada efeito em separado *versus* a variável resposta, em

Tabela 4 – Resíduos estudantizados e teste de bonferroni para indentificar outliers no modelo de regressão múltipla com a variável transformada.

Obs.	Resid. estudent.	Valor P não Ajustado	P Bonferonni
255	-5,99	$4,25 \times 10^{-09}$	$1,93 \times 10^{-06}$
267	-4,85	$1,63 \times 10^{-06}$	$7,46 \times 10^{-04}$
406	-4,06	$5,71 \times 10^{-05}$	$2,60 \times 10^{-02}$

que percebe-se o efeito linear de grupo e recursos e ausência de efeitos para ano. Realizou-se na análise de diagnóstico gráficos suplementares com a identificação dos pontos atípicos deste ajuste verificou-se que as observações 92,200,255,256,262,267,307,394,406,410,442, referentes as regiões Chapada (CH), 4 observações de Itaparica (IT), 1 do Semi-Árido (SA), 3 do observações Velho Chico (VC), respectivamente, se configuraram como atípicas. Essas 11 observações foram retiradas e o modelo foi novamente ajustado. A retirada das observações melhorou significativamente o ajuste do modelo, reduzindo o Erro Padrão e o p-valor relacionado ao parâmetro de cada variável. As medidas de diagnóstico para o novo modelo sem as 11 observações citadas acima foram realizadas. Surgiram novos pontos atípicos, porém, a retirada dessas observações não causaram grandes mudanças na variação percentual das estimativas dos parâmetros. Logo, optou-se por conservar essas observações no modelo. O gráfico de probabilidade normal com envelope simulado (Figura 11), mostra que quase todos os pontos se encontram dentro da banda de confiança, indicando um ajuste satisfatório do modelo aos dados.

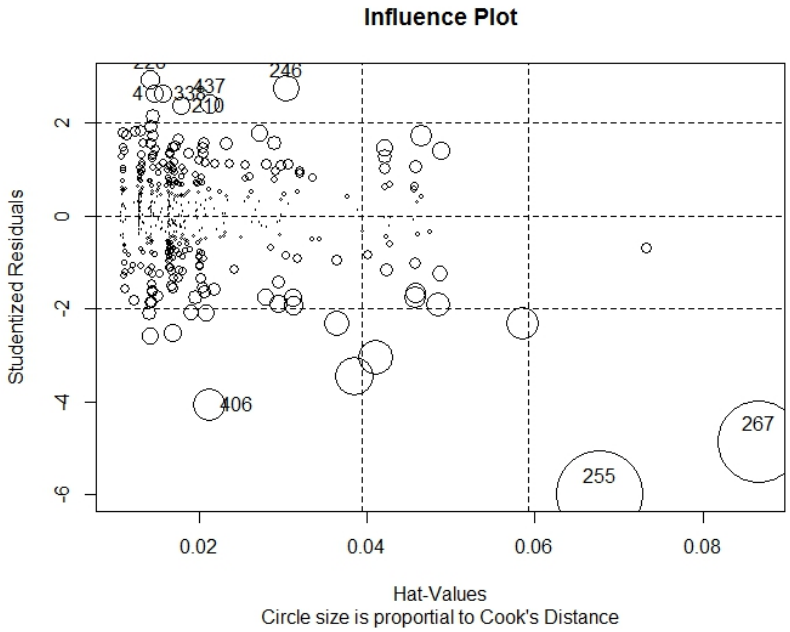

Figura 8 – Gráfico de Influência para os resíduos estudentizados *versus* os valores h

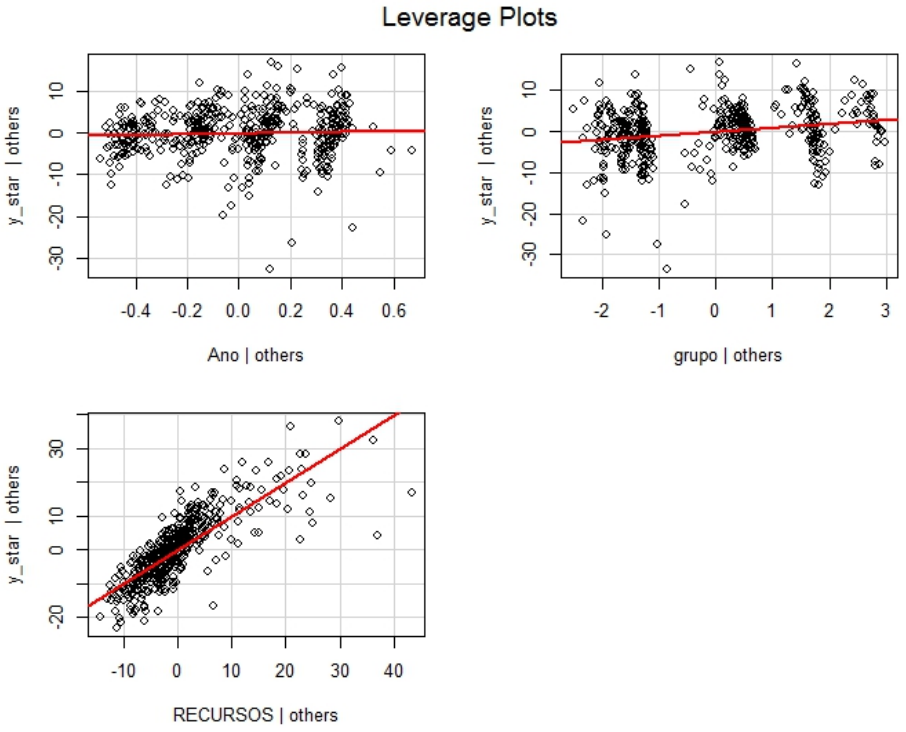

Figura 9 – Gráfico de Leverage para cada efeito do modelo de regressão múltipla com a variável transformada.

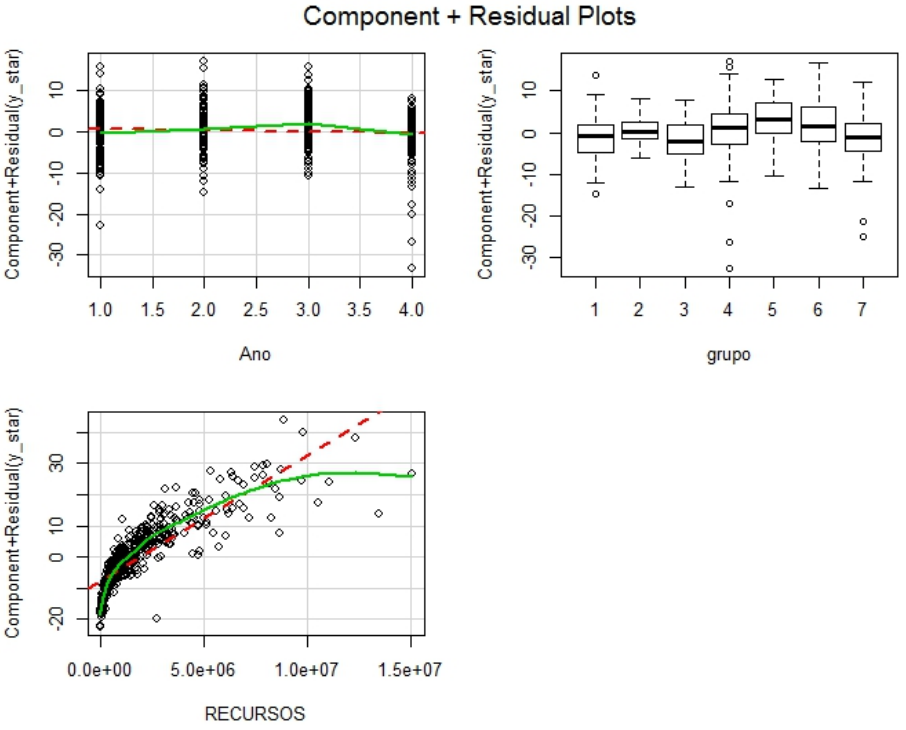

Figura 10 – Gráficos da variável adicionada para cada efeito do modelo de regressão múltipla com a variável transformada.

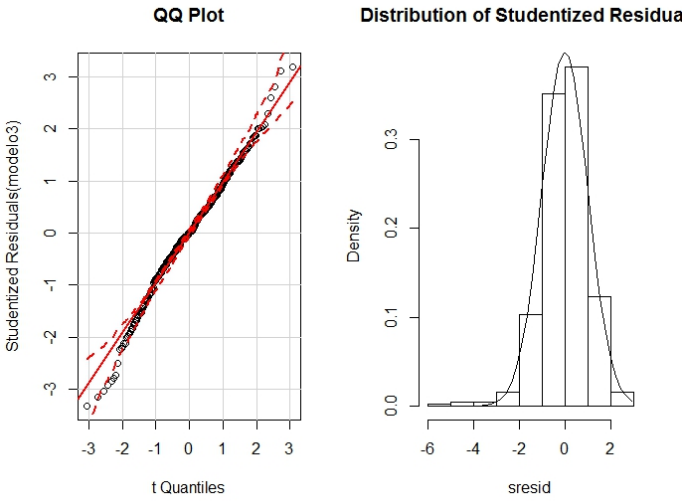

Figura 11 – Gráfico de probabilidade normal com envelope simulado e histograma dos resíduos estudentizados para a variável transformada e com 11 observações retiradas.

O modelo final ajustado foi o que se apresenta na Tabela 5, em que todos efeitos empregados no modelo de regressão múltipla foram significativos a exceção dos grupos 3, 4 e 7. Pode-se afirmar que ano a região do velho chico apresentaram efeito negativo para o Número de contratos as demais regiões houve tendência crescente para a quantidade de contratos realizados.

Tabela 5 – Estimativas dos parâmetros, erros-padrão e Valor t do modelo para a variável transformada com a retirada de 11 observações.

	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
(Intercepto)	13,4685	0,7164	18,80	0,0000
Ano	-0,4600	0,2194	-2,10	0,0366
grupo2	1,8542	0,7361	2,52	0,0121
grupo3	-0,5921	1,1273	-0,53	0,5997
grupo4	1,1091	0,8073	1,37	0,1702
grupo5	2,7001	0,9509	2,84	0,0047
grupo6	2,2077	0,7421	2,97	0,0031
grupo7	-0,0775	0,8045	-0,10	0,9233
RECURSOS	0,0000	0,0000	36,50	0,000

5 Conclusão

Neste trabalho ajustou-se o modelo de Regressão linear múltipla ao conjunto de dados do Pronaf - DAP. As variáveis analisadas foram consideradas os totais de Número de contratos em função dos recursos alocados para cada região do Estado da Bahia e dos anos que foram feitas as alocações de recursos e as regiões (grupos) do Estado. Após o ajuste com todas as variáveis, utilizamos ferramentas para analisar o ajuste e adequação do modelo de regressão linear múltipla. Foi aplicado o teste de Shapiro-Wilks para os resíduos do modelo sem transformação e após a transformação Box-Cox (transformação raiz quadrada) e este rejeitou a hipótese de normalidade dos mesmos ambas as vezes, porém o teste de Durbin-Watson não rejeitou a hipótese nula ao se fazer a transformação, indicando que a transformação de Box-cox corrigiu uma das suposições do modelo de regressão, realizamos uma análise de resíduos graficamente e por meio dela é possível verificar que houve melhora no ajuste da distribuição aos dados e que existe alguns pontos identificados que podem ser importantes e interferir na análise, foram retirados esses pontos considerados *outliers* e influentes na qualidade do ajuste do modelo. O modelo final foi ajustado sem 11 observações e a retirada das observações melhorou significamente o ajuste, reduzindo o erro padrão e o p-valor relacionado ao parâmetro de cada variável.

Referências

AKAIKE, H.Information theory as an extension of the maximum likelihood principe. In: INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY, 2.,Budapest 1973.Proceedings. Budapest, akadêmia Kiadó, 1973.p 267-281.

BOX, G.E.P.; COX, D.R.; An Analysis of Transformations. Journal of the Royal Statistical Society, London, v.26, nº2, 211 - 252, 1964.

BOX, George EP et al. Time series analysis: forecasting and control. John Wiley & Sons, 2015. APA

CHARNET, R.; FREIRE, C.A.L.; CHARNET, E.M.R.; BONVINO, H. ANáLISE DE MODELOS DE REGRESSÃO LINEAR - Com aplicacões. 2º ed. Campinas, SP:UNICAMP, 2008.

DELEEUW, Jan. Information theory and extension of the maximum likelihood principe by hirotogu Akaike. **Departament of Statistics, UCLA**, 1992.

Duan, N."Smearing Estimate: A Nonparametric Retronsformation Method"JASA, 78, 605-610, 1983.

DRAPER, N.R.; SMITH, H.; Applied regression analysis. 3º ed. New York, New York: John Wiley & Sons, 1998. 706 p.

FERREIRA, D.F.; Estatística Básica. 2º ed. rev. Lavras, MG: UFLA, 2009. 664 p.

GUJARATI, D. N; DAWN,C.P. Econometria Básica. AMGH, Editora, 2011.

HOFFMAN, A. ANÁLISE DE REGRESSÃO - Uma Introdução à Econometria. 4º ed. São Paulo, SP: Hucitec, 2006. 378 p.

MILLER, D.M.; Reducing Transformation Bias in Curve Fitting. The American Statistician, v.38, nº2, 124 - 126, 1984.

MONTGOMERY, D.C.; PECK, E.A.; VINING, G.G.; Introduction to Linear Regression Analysis. 3º ed. New York, New York: John Wiley & Sons, 2003. 641 p.

MONTGOMERY, D. C.; RUNGER, G.C. (2003) **Estatística aplicada a probabilidade para engenharia** 2ªed. Rio de janeiro: LTC-Livros técnicos e ciêntificos S.A 463p.

PAULA, Gilberto Alvarenga. **Modelos de Regressão: com apoio computacional**. São Paulo: IME-USP,2004.

QUEIROZ, M.P.F.; Testes de Hipóteses em Regressão Beta Baseados em Verossimilhança Perfilada Ajustada e em Bootstrap. (Dissertação mestrado), Universidade Federal de

Pernambuco, PE, 2011.

RENCHER, A. C.; SCHAALJE, G.B. **Linear models in statistics**.John Wiley Sons, 2007.

RODRIGUES, S.A.; DINIZ, C.A.R. Modelo de regressão heterocedástico. **Revista de Matemática e Estatística**. v. 24, n. 2, p.133-146, 2006.

SHAPIRO, S. S, WILK, M. B. An analyses of variance teste for normality (cpmplete samples) **Biometria**, p.591-611,1965.

TAYLOR, J.M.G; The Retransformed Mean after a Fitted Power Transformatio, JAVA, 81, 114 - 118, 1986.