Universidade Federal de Juiz de Fora
Departamento de Estatística
Curso de Estatística

Luís Gustavo Silva e Silva

Estimador regressão para dados assimétricos

Juiz de Fora
2010

Luís Gustavo Silva e Silva

Estimador regressão para dados assimétricos

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção parcial do grau de BACHAREL em Estatística.

Orientador: Clécio da Silva Ferreira

Doutor em Estatística - Universidade de São Paulo

Co-orientador: Marcel de Toledo Vieira

Doutor em Estatística - University of Southampton

Juiz de Fora

2010

Silva, Luís

Estimador regressão para dados assimétricos / Luís Silva - 2010

31.p

1.Amostragem 2. Modelos assimétricos.. I.Título.

CDU 536.21

Luís Gustavo Silva e Silva

Estimador regressão para dados assimétricos

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção parcial do grau de BACHAREL em Estatística.

Aprovado em 9 de dezembro de 2010

BANCA EXAMINADORA

Clécio da Silva Ferreira

Doutor em Estatística - Universidade de São Paulo

Marcel de Toledo Vieira

Doutor em Estatística - University of Southampton

Camila Borelli Zeller

Doutora em Estatística - Universidade Estadual de Campinas

A Deus. Aos meus pais e irmãos.

Aos amigos, pelo apoio e companheirismo.

Resumo

O presente trabalho tem como objetivo propor uma modificação na estrutura do estimador do tipo regressão, considerando o estimador do coeficiente linear da reta de regressão, ou seja, denotado por $\hat{\beta}$ sob assimetria ao invés de normalidade dos dados. Desta forma nomeamos este estimador proposto como estimador para dados assimétricos. A partir desta mudança iremos comparar os estimadores quanto ao viés e a variância dos mesmos, à variância de $\hat{\beta}$ estimados para cada estimador, considerando uma população com distribuição normal assimétrica. Para ilustrar a metodologia proposta, iremos considerar um estudo de simulação e uma aplicação aos dados reais.

Palavras-chaves: Distribuição normal assimétrica, estimador regressão.

Abstract

This paper aims at proposing a modification in the structure of the regression estimator, considering the estimator of the coefficient of linear regression line, ie, denoted by *hat beta* skewness rather than under normality. Thus we named this proposed estimator as an estimator for skewness data. From this change we will compare the estimators on the bias and variance of the same, the variance of *hat beta* estimates for each estimator, assuming a normal distribution skewed. To illustrate the proposed methodology, we consider a simulation study and application to actual data.

Keywords: Skew-normal distribution, regression estimator..

Agradecimentos

À Deus, pela proteção e paz de espírito concedida ao longo de toda a minha vida.

À minha família que sempre me incentivou para os estudos, ao carinho especial dos meu pais Dorcina e Wilson, que nunca mediram esforços para me ajudar ao longo destes anos. Ao Papai que a cada telefonema me incentiva aos estudos, mesmo com poucas palavras, mas no entanto sempre sábias. À Mamãe com sua doçura e abraços que confortam todo o cansaço. Pedro e Carol, não poderia ter irmãos melhores do que estes, obrigado por acreditarem nos meus sonhos.

A todos os meus tios e primos, em especial aos tios Zezinho e Geraldinho (*in memorian*) pelo incentivo incansável ao longo dos meus estudos.

Aos amigos de Governador Valadares, Bruno pelas frases inigualáveis, Lucas pela sinceridade e alegria, Marcelo pelas gargalhadas sem motivos, Marconi por sempre acreditar na minha caminhada na Estatística e por me incentivar a UFJF, Matheus pela confiança e amizade incondicional que sempre me motivou a realizar os meus sonhos e ao Roberto que foi um grande companheiro ao longo destes anos e pelas nossas conversas sem fim. A todos eles pela amizade sincera, aos momentos de muita alegria e por sempre me confortarem nos momentos difíceis.

Aos amigos de Juiz de Fora, em especial ao Samuel, homem de profunda piedade e discernimento espiritual. Bruno, pela serenidade e pelos passeios na roça. Iago, pela determinação e exemplo de que mesmo com as dificuldade podemos ser os melhores. A Laura, minha irmã de coração, pelas conversas em sua casa e pelo seu abraço aconchegante. E todos os outros amigos, Victor, Carol e Thiago.

À Priscila, pelo companheirismo, apoio e compreensão, que sem eles não poderia cumprir tal etapa, por nunca me deixar esmorecer e pelo carinho incondicional que me deu ao longo deste anos.

Aos meus amigos e orientadores Clécio e Marcel, pela confiança e pelo aprendizado ao longo deste trabalho.

Aos professores do departamento de Estatística, em especial a Professora

Camila, que mesmo não sendo a minha orientadora teve papel importantíssimo neste trabalho. Ao professor e amigo Joaquim, pelo incentivo à curiosidade acadêmica e aos ensinamentos. Ao professor Márcio, grande amigo e incentivador.

"...a saudade será uma companheira eterna, pois momentos bons serão sempre lembrados".

Silva e Silva

"Confia no Deus eterno de todo o seu coração e não se apóie na sua própria inteligência. Lembre-se de Deus em tudo o que fizer, e ele lhe mostrará o caminho certo." *(Prov. 3:5-6)*

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Em muitas situações práticas a distribuição normal e os modelos de regressão para dados com distribuição normal têm sido de grande utilidade. Entretanto, há indicações de que a suposição de normalidade não se aplica em certas situações, por exemplo, quando há falta de simetria dos dados. Desta forma propõe-se como alternativa a utilização de uma distribuição, de forma que se consiga modelar a assimetria dos dados e além disso, incluir a distribuição normal como um caso particular.

A distribuição normal assimétrica univariada surgiu independentemente em vários artigos estatísticos, entre os principais trabalhos pode-se destacar Roberts (1966), O'Hagan e Leonard (1976) e Aigner et al. (1977). No Capítulo 2 descrevemos a distribuição normal assimétrica introduzida por Azzalini (1985, 1986) nas formas padrão e de locação escala, destacando suas propriedades.

Os estimadores do tipo regressão, são estimadores que utilizam variáveis auxiliares em sua estrutura com intuito de melhorar a precisão de suas estimativas. Estas variáveis auxiliares devem ter uma relação linear para com a variável de interesse e ser conhecida para toda população. No Capítulo 3, apresentaremos as propriedades do estimador regressão quando estamos interessados em estimar a média de uma variável de interesse.

O objetivo deste trabalho é propor uma modificação na estrutura do estimador do tipo regressão, considerando o estimador do coeficiente linear da reta de regressão, ou seja, denotado por $\hat{\beta}$ sob assimetria ao invés de normalidade dos dados. Desta forma nomeamos este estimador proposto como estimador para dados assimétricos. A partir desta mudança iremos comparar os estimadores quanto ao viés e a variância dos mesmos, à variância de $\hat{\beta}$ estimados para cada estimador, considerando uma população com distribuição normal assimétrica. Para ilustrar a metodologia proposta, iremos considerar um estudo de simulação e uma aplicação aos dados reais.

2 Modelo Normal Assimétrico

2.1 Distribuição Normal Assimétrica Padrão

Definição 2.1.1. Uma variável aleatória Z tem distribuição normal assimétrica padrão se sua função densidade de probabilidade é dada por

$$f_Z(z) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R} \tag{2.1}$$

onde $\phi(\cdot)$ e $\Phi(\cdot)$ são as funções densidade de probabilidade e distribuição de uma normal padrão, respectivamente (veja Azzalini, 1985).

O parâmetro λ caracteriza a forma da distribuição e também é denominado parâmetro de assimetria, que para valores negativos de λ indicam assimetria negativa e para valores positivos de λ indicam assimetria positiva. Se $\lambda = 0$, a densidade acima coincide com a densidade da distribuição normal padrão e portanto é simétrica. Será utilizada a seguinte notação: $Z \sim NA(\lambda)$.

Através de (2.1) podemos ver que a função distribuição da normal assimétrica pode ser facilmente obtida se tivermos acesso a um programa que calcule a distribuição acumulada de uma normal univariada padrão.

Apresentamos algumas propriedades interessantes da densidade em (2.1):

1. Se $Z \sim NA(\lambda)$, então $-Z \sim NA(-\lambda)$;

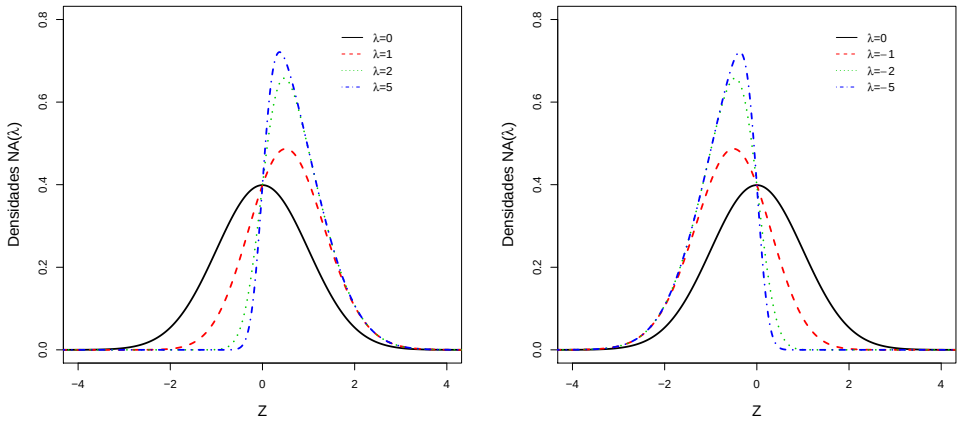
2. Se $Z \sim NA(\lambda)$, então $Z^2 \sim \mathcal{X}^2_1$;

3. (Representação estocástica de Henze, 1986) Se $U, V \sim N(0,1)$, independentes, então

$$\frac{\lambda}{\sqrt{1+\lambda^2}}|U| + \frac{1}{\sqrt{1+\lambda^2}}V \sim NA(\lambda). \tag{2.2}$$

Esta última propriedade é útil para gerar amostras da distribuição normal assimétrica a partir da normal padrão. Outras propriedades podem serem encontradas em Ferreira (2008).

A partir da representação estocástica da distribuição normal assimétrica podemos derivar importantes medidas (Ferreira, 2008), como média e variância da distribuição



Função densidade $NA(\lambda)$, para $\lambda \geq 0$

Função densidade $NA(\lambda)$, para $\lambda \leq 0$

Figura 2.1: Função densidade $NA(\lambda)$, para diferentes valores de λ.

$NA(\lambda)$ que são dadas por

$$E[Z] = c\rho \quad \text{e} \quad Var(Z) = 1 - c^2\rho^2, \tag{2.3}$$

com $c = \sqrt{\frac{2}{\pi}}$.

O modelo (2.1) é estendido introduzindo parâmetros de locação $\mu \in \mathbb{R}$ e de escala $\sigma > 0$. Neste caso, será utilizada a notação $Y \sim NA(\mu, \sigma^2, \lambda)$.

2.1.1 Distribuição Normal Assimétrica de Locação-Escala

Definição 2.1.2. Uma variável aleatória Y tem distribuição normal assimétrica com parâmetros de posição μ e de escala σ se sua função densidade de probabilidade é da forma

$$f_Y(y) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right), \quad y \in \mathbb{R}. \tag{2.4}$$

É fácil verificar que se $Z \sim NA(\lambda)$ e $Y = \mu + \sigma Z$, então $Y \sim NA(\mu, \sigma^2, \lambda)$ (veja Azzalini, 1985).

A média e a variância de uma variável aleatória $Z \sim NA(\mu, \sigma^2, \lambda)$ são dadas por

$$E[Y] = \mu + c\sigma\rho \quad \text{e} \quad Var[Y] = \sigma^2\left(1 - c^2\rho^2\right). \tag{2.5}$$

Mais informações sobre propriedades e inferência estatística em modelos normais assimétricos podem ser encontradas em Rodríguez (2005), Gómez (2005), Lin et al. (2007) e Ferreira (2008).

2.2 Modelo de regressão com assimetria

Segundo Ferreira (2008), um conjunto de n observações independentes, denotadas por $Y_1, \ldots, Y_n$, onde $Y_i \sim NA(\mu_i, \sigma^2, \lambda)$, $i = 1, \ldots, n$. Associado com a observação i, considere um vetor $p \times 1$ de covariáveis $\mathbf{x}_i$, através do qual especifica-se o preditor linear $\mu_i = \mathbf{x}_i^\top \beta$, onde β é um vetor p-dimensional de coeficientes de regressão desconhecidos. Assim, relacionando os dois conjuntos de variáveis, tem-se o modelo

$$\begin{aligned} y_i &= \beta_0 + \textstyle\sum_{k=1}^{p} x_{ik}\beta_k + \varepsilon_i, \quad i = 1, \ldots, n, \\ &= \mathbf{x}_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim NA(0, \sigma^2, \lambda). \end{aligned} \tag{2.6}$$

Note que $E[\varepsilon_i] = \sqrt{\frac{2}{\pi}}\frac{\sigma\lambda}{\sqrt{1+\lambda^2}} \neq 0$, para $\lambda \neq 0$. Em termos de previsão para Y, geralmente considera-se $\widehat{Y}_i = \widehat{\beta}_0 + \sum_{k=1}^{p} x_{ik}\widehat{\beta}_k$ como preditor de $Y_i|\mathbf{x}_i$. Uma forma de corrigir esta distorção é considerar $\widehat{Y}_i = \widehat{\beta}_0 + \sum_{k=1}^{p} x_{ik}\widehat{\beta}_k + \sqrt{\frac{2}{\pi}}\frac{\widehat{\sigma}\widehat{\lambda}}{\sqrt{1+\widehat{\lambda}^2}}$ como preditor de $Y_i|\mathbf{x}_i$. Outra possibilidade é considerar o modelo centrado (ver Freitas, 2005).

Pode-se verificar que a função log-verossimilhança para $\theta = (\beta^\top, \sigma^2, \lambda)^\top$ para uma amostra de n observações $(y_1, \ldots, y_n)$ é dada por, ver Ferreira (2008)

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{m} \log\left[2\phi(y_i|\mathbf{x}_i^\top\beta, \sigma^2)\Phi\left(\frac{\lambda(y_i - \mathbf{x}_i^\top\beta)}{\sigma^2}\right)\right] \\ &= \sum_{i=1}^{m} \log\left[2\int_0^{+\infty} \phi(y_i|\mathbf{x}_i^\top\beta, \sigma^2)\phi(t_i|\lambda(y_i - \mathbf{x}_i^\top\beta), \sigma^2)dt_i\right]. \end{aligned} \tag{2.7}$$

2.3 Estimação de Máxima Verossimilhança via Algoritmo EM

Se maximizarmos diretamente a função acima para encontrarmos as estimativas de máxima verossimilhança dos parâmetros pode ser complicado, devido à presença de integrais na expressão 2.7. Uma alternativa para a solução deste problema é utilizar um procedimento de estimação usando algoritmo EM, uma ferramenta usual para estimação de máxima verossimilhança para modelos com dados incompletos.Mais especificamente, seja

$\mathbf{y}$ o conjunto de dados observados e $\mathbf{t}$ denotando o conjunto de dados faltantes. O dado completo $\mathbf{y}_c = (\mathbf{y}, \mathbf{t})$ é $\mathbf{y}$ aumentado com $\mathbf{s}$. Denota-se por $\ell_c(\theta|\mathbf{y}_c)$, $\theta \in \Theta$, a função log-verossimilhança dos dados completos e por $Q(\theta|\widehat{\theta}) = E[\ell_c(\theta|\mathbf{y}_c)|\mathbf{y}, \widehat{\theta}]$, o valor esperado desta função. Cada iteração do algoritmo EM envolve dois passos, um passo E e um passo M, definidos como:

- Passo E: Calcule $Q(\theta|\theta^{(k)})$ como uma função de θ;
- Passo M: Encontre $\theta^{(k+1)}$ tal que $Q(\theta^{(k+1)}|\theta^{(r)}) = \max_{\theta\in\Theta} Q(\theta|\theta^{(k)})$.

Utilizando a representação estocástica de Henze (1986) (Propriedade 3), o modelo de regressão (2.6) acima pode ser escrito como

$$\begin{aligned} Y_i|T_{=}t_i &\overset{ind}{\sim} N\left(\mathbf{x}_i^\top\beta + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}t_i, \frac{\sigma^2}{1+\lambda^2}\right), \\ T_i &\overset{iid}{\sim} NT(0,1), i = 1, \ldots, n \end{aligned} \tag{2.8}$$

denotando por $NT(\mu, \sigma^2)$ a distribuição normal truncada à esquerda de zero (Johnson et al., 1994), com parâmetro de locação μ e de escala σ^2.

De $\ell(\theta)$ em (2.7), a distribuição conjunta de y_i e t_i é dada por

$$f(y_i, t_i) = \phi(y_i|\mathbf{x}_i^\top\beta, \sigma^2)\phi(t_i|\lambda(y_i - \mathbf{x}_i^\top\beta), \sigma^2)\mathbf{I}(t_i > 0).$$

Seja $\mathbf{y} = (y_1, \ldots, y_n)^\top$ e $\mathbf{t} = (t_1, \ldots, t_n)^\top$ e tratando $\mathbf{t}$ como dado faltante, segue que a função log-verossimilhança completa associada com $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{t}^\top)^\top$ é dada por

$$\begin{aligned} \ell_c(\theta|\mathbf{y}_c) &= \sum_{i=1}^{n} \log f(y_i, t_i) \\ &\propto -n\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\beta)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[t_i - \lambda(y_i - \mathbf{x}_i^\top\beta)]^2 \\ &= -n\log\sigma^2 - \frac{1}{2\sigma^2}\mathbf{t}^{2\top}\mathbf{1}_n + \frac{\lambda}{\sigma^2}\mathbf{t}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1+\lambda^2}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \end{aligned} \tag{2.9}$$

onde $\mathbf{t^2} = (t_1^2, \ldots, t_n^2)^\top$, $\mathbf{1}_n$ é um vetor de 1's de tamanho n e $\mathbf{X} = (\mathbf{x}_1 \ldots, \mathbf{x}_n)^\top$ é a matriz de planejamento, de dimensão $n \times p$.

De (2.7), tem-se que $T_i|y_i \sim NT(\lambda(y_i - \mathbf{x}_i^\top\beta), \sigma^2)$. Seja $\widehat{t}_i = \mathrm{E}[T_i|\theta = \widehat{\theta}, y_i]$ e $\widehat{t^2}_i = \mathrm{E}[T_i^2|\theta = \widehat{\theta}, y_i]$. Então, usando os momentos da distribuição normal truncada (Lachos, 2004), tem-se que

$$\widehat{t}_i = \widehat{\lambda}\widehat{\eta}_i + \widehat{\sigma}W_{\Phi_1}\left(\frac{\widehat{\lambda}\widehat{\eta}_i}{\widehat{\sigma}}\right) \text{ e } \widehat{t^2}_i = \widehat{\lambda}^2\widehat{\eta}_i^2 + \widehat{\sigma}^2 + \widehat{\lambda}\widehat{\sigma}\widehat{\eta}_i W_{\Phi_1}\left(\frac{\widehat{\lambda}\widehat{\eta}_i}{\widehat{\sigma}}\right), \tag{2.10}$$

onde $W_{\Phi_1}(u) = \phi_1(u)/\Phi_1(u)$ e $\widehat{\eta}_i = (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})$, $i = 1, \ldots, n$.

Denote por $\theta^{(k)} = (\beta^{(k)^\top}, \sigma^{2(k)}, \lambda^{(k)})^\top$ a estimativa de θ para a k-ésima iteração. Segue que a esperança com respeito a $\mathbf{t}$, condicionada em $\mathbf{y}$, da função log-verossimilhança completa (Passo E), tem a forma

$$\begin{aligned} Q(\theta|\widehat{\theta}^{(k)}) &= \mathrm{E}[\ell_c(\theta|\mathbf{y}_c)|\mathbf{y}, \widehat{\theta}^{(k)}] \\ &= -n\log\sigma^{2(k)} - \frac{1}{2\sigma^{2(k)}}\sum_{i=1}^{n}\widehat{t^2}_i^{(k)} + \frac{\lambda^{(k)}}{\sigma^{2(k)}}\sum_{i=1}^{n}\widehat{t}_i^{(k)}(y_i - \mathbf{x}_i^\top\beta^{(k)}) \\ &\quad -\frac{1+\lambda^{(k)^2}}{2\sigma^{2(k)}}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\beta^{(k)})^2. \end{aligned} \tag{2.11}$$

Portanto, tem-se o seguinte algoritmo EM:
Passo E: Dado $\theta = \widehat{\theta}^{(k)}$, calcule $\widehat{t}_i^{(k)}$ e $\widehat{t^2}_i^{(k)}$, para $i = 1, \ldots, n$, usando (2.10).
Passo M: Atualize $\widehat{\theta}^{(k+1)}$ maximizando $Q(\theta|\widehat{\theta}^{(k)})$ em θ, que leva às seguintes soluções analíticas:

$$\begin{aligned} \widehat{\beta}^{(k+1)} &= [\mathbf{X}^\top\mathbf{X}]^{-1}\mathbf{X}^\top\mathbf{y} - \frac{\lambda^{(k)}}{1+\lambda^{(k)^2}}[\mathbf{X}^\top\mathbf{X}]^{-1}\mathbf{X}^\top\widehat{\mathbf{t}}^{(k)}, \\ \widehat{\sigma^2}^{(k+1)} &= \frac{1}{2n}\left[\widehat{\mathbf{t}^2}^{(k)^\top}\mathbf{1}_n - 2\lambda^{(k)}\widehat{\mathbf{t}}^{(k)^\top}(\mathbf{y} - \mathbf{X}\beta^{(k)}) + (1+\lambda^{(k)^2})Q(\beta^{(k)})\right], \\ \widehat{\lambda}^{(k+1)} &= \frac{\widehat{\mathbf{t}}^{(k)^\top}(\mathbf{y} - \mathbf{X}\beta^{(k)})}{Q(\beta^{(k)})}, \end{aligned} \tag{2.12}$$

onde $Q(\beta^{(k)}) = (\mathbf{y} - \mathbf{X}\beta^{(k)})^\top(\mathbf{y} - \mathbf{X}\beta^{(k)})$.

Claramente, se $\lambda = 0$, $\widehat{\beta} = [\mathbf{X}^\top\mathbf{X}]^{-1}\mathbf{X}^\top\mathbf{y}$ é o EMV de β do modelo normal simétrico. Por outro lado, $\lambda = 0$ implica $\widehat{\mathbf{t^2}}_i = \widehat{\sigma^2}$, resultando na equação $\widehat{\sigma^2} = \frac{1}{2n}\left[n\widehat{\sigma^2} + Q(\widehat{\beta})\right]$, implicando $\widehat{\sigma^2} = \frac{Q(\widehat{\beta})}{n}$, coincidindo com o EMV de σ^2 do modelo normal simétrico.

São usados como valores iniciais para θ no algoritmo os estimadores de momentos (Rodríguez, 2005). O EMV de θ através do algoritmo EM é resultado encontrado em Ferreira (2008).

3 Estimador do tipo regressão

A utilização de informações auxiliares para melhorar a precisão das estimativas é uma das área de pesquisa no âmbito da teoria de amostragem. O estimador de regressão introduzido neste capítulo é um tipo de estimador que faz o uso eficiente de informações auxiliares sobre a população afim de melhorar as estimativas. (Sarndal, Swenson & Wretman, 1992).

Segundo Ferraz & Vieira (2009)

> Estimadores assistidos por modelos lineares pertencem à classe geral dos estimadores do tipo regressão. Os estimadores de regressão tiveram sua origem de trabalhos de Hansen, na década de 1940. Novos avanços desta metodologia ocorreram na década de 1980, quando foi investigada por Sarndal, Swenson & Wretman(1992).

A utilização deste estimador é dada em situações que o elemento i da população finita U, tem-se associado o par (X_i, Y_i), $i = 1, ..., N$ obedecendo uma relação linear, ou seja,

$$Y_i = \alpha + \beta X_i + e_i \tag{3.1}$$

onde e_i é o desvio em torno da reta, $i = 1, ..., N$.

A introdução da variável auxiliar X tem como intuito melhorar as estimativas de parâmetros como média ou total populacional. Na teoria de regressão assume-se que as quantidades X_i, $i = 1, ..., N$ são conhecidas, ou seja, conhecemos de antemão a média populacional μ_X, total populacional T_X e o desvio padrão populacional σ_X, ver Pessoa & Costa (2009).

Para uma amostra s de tamanho n, produzindo médias amostrais $\bar{y}$ e $\bar{x}$, o estimador regressão da média é dado por

$$\bar{Y}_{Reg} = \bar{y} + \hat{\beta}(\mu_X - \bar{x}), \tag{3.2}$$

onde $\hat{\beta}$ é o estimador de β, o coeficiente - do modelo linear que descreve a relação entre a variável de interesse y e a auxiliar x, segundo Bolfarine & Bussab (2005). Quando temos mais de uma variável auxiliar disponível podemos acomodá-las na forma geral do

estimador regressão, como segue:

$$\bar{y}_{\text{Reg}} = \bar{y} + \sum_{j=1}^{p} \hat{\beta}_j \left(\mu_{X_j} - \bar{x}_j\right). \tag{3.3}$$

Observando as formas 3.2 e 3.3 percebemos que uma maneira de interpretarmos o estimador regressão é que ele corresponde ao estimador natural da média populacional mais um termo de correção.

Algumas propriedades como média e variância dos estimadores do tipo regressão para Amostras Aleatórias Simples (AAS) e Amostra Aleatória Simples com Reposição (ASSc).

Teorema 1. *Seja $\bar{y}_{Reg}$ definido com $b = b_0$, fixo e conhecido. Então, para o plano AAS temos que, $\bar{y}_{Reg}$ é um estimador não viesado de μ_Y, isto é, $E[\bar{y}_{Reg}] = \mu_Y$.*

Teorema 2. *Com relação à AASc, tem-se que $Var[\bar{y}_{Reg}] = \frac{1}{n}\left(\sigma_Y^2 - 2b_0\sigma_{XY} + b_0^2\sigma_X^2\right)$*

Corolário 1. *Um estimador não viciado para $V_{Reg} = Var[\bar{y}_{Reg}]$ com b_0 fixado é dado por $\hat{V}_{Reg} = [\bar{y}_{Reg}] = \frac{1}{n}\left(\sigma_y^2 - 2b_0\hat{\sigma}_{xy} + b_0^2\hat{\sigma}_x^2\right)$. Sendo $\hat{\sigma}_y^2$, $\hat{\sigma}_{xy}^2$ e $\hat{\sigma}_x^2$ a variância da variável de interesse Y, covariância de x e y e a variância da variável auxiliar x, respectivamente.*

As provas dos teoremas 1, 2 e do corolário 1 podem ser encontradas em Bolfarine & Bussab (2005).

4 Aplicação

4.1 Estudo de Simulação

Nesta seção apresentaremos um estudo de simulação para verificarmos o comportamento dos estimadores regressão assimétrico, comum e o estimadro simples, quanto à variância, o coeficiente de variação (CV) e a média estimada por eles.

Para o nosso estudo de simulação geramos a variável auxiliar $X \sim NA(0, 1, \lambda)$ com 10000 observações, onde assumimos os seguintes valores para o parâmetro de assimetria $\lambda = 0, 5, 10$, a fim de comparar os resultados destas três populações. Geramos em seguida uma outra variável que correlaciona linearmente com X, ou seja, $Y = \beta_0 + \beta_1 X + \epsilon$, sendo $\epsilon \sim N(0, 1)$ e os parâmetros β_0 e β_1 fixos com os valores 10 e 5 respectivamente. A partir desta população selecionamos amostras com $n = 500, 1000$ e 5000 a partir do método de amostragem aleatória simples sem reposição.

Considerando $\lambda = 0$ e $n = 500$, percebemos que as estimativas apresentadas na Tabela 4.1 não apresentam muita diferença entre si e em relação ao verdadeiro valor da média 4.9819. Observamos também que para este caso a média estimada pelo estimador regressão comum (ERC)e o estimador regressão para dados assimétricos (ERA) foi a mesma, resultado este já esperado, pois para $\lambda = 0$ não temos nenhuma assimetria, como pode ser verificado visualmente pela Figura 4.1. Analisando a variância dos estimadores verificamos que o estimador simples teve a maior variância, enquanto que os dois estimadores regressão obtiveram a mesma variância e naturalmente menor que o estimador simples. Portando podemos afirmar que os estimadores regressão são mais eficientes que o estimador simples quando $\lambda = 0$. Os resultados encontrados para $\lambda = 5$ e $\lambda = 10$ são similares para ao caso citado anteriormente e como pode ser visto na Tabela 4.1. A média populacional para $\lambda = 5$ e $\lambda = 10$ são 12.8414 e 12.9542, respectivamente. Os resultados para λ negativo são análogos ao caso positivo.

A notação utilizada para identificar os estimadores foram adotas da seguinte forma: MS para o estimador da média simples, ERC para o estimador regressão comum (simétrico) e ERA para o estimador regressão para dados assimétricos. Na Tabela 4.1

adotamos M como a média das médias estimadas pelos estimadores em questão, V como a variância das estimativas das médias e CV como o coeficiente de variação dos estimadores.

Tabela 4.1: Comparações dos estimadores para diferentes λ's e $n = 500$.

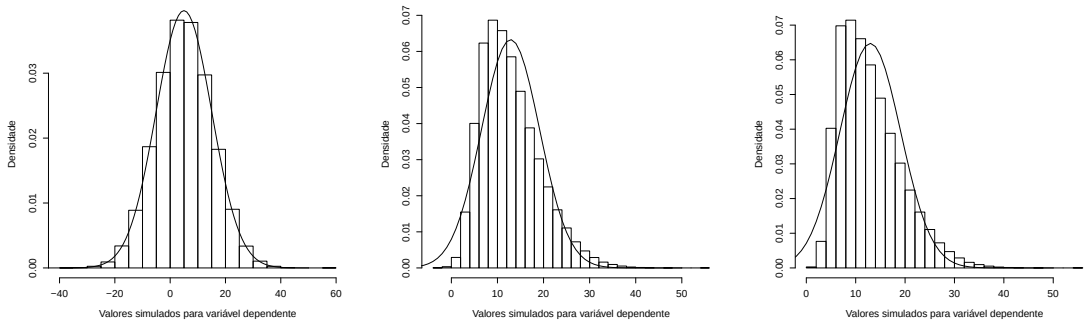
	$\lambda = 0$			$\lambda = 5$			$\lambda = 10$		
Estimadores	M	V	CV	M	V	CV	M	V	CV
MS	4.9712	0.2005	0.0901	12.8315	0.0771	0.0216	12.9465	0.0746	0.0211
ERC	4.9718	0.0023	0.0096	12.8377	0.0022	0.0036	12.9505	0.0022	0.0036
ERA	4.9718	0.0023	0.0096	12.8377	0.0022	0.0036	12.9505	0.0022	0.0036

A Figura 4.1 ilustra a distribuição dos dados para cada amostra de tamanho 500 selecionada de sua respectiva população, para cada histograma foi sobreposto a curva da densidade da distribuição normal. Note que para valores de λ maior que zero temos uma assimetria positiva e quando $\lambda = 0$ temos que os dados se aproximam muito bem de uma distribuição normal.

Outra medida adotada a fim de comparar os estimadores regressão é variância do coeficiente linear ($\hat{\beta}$) presente em ambos modelos. A variância encontrada para o coeficiente linear do estimador regressão comum foi 0.0019 quando $\lambda = 0$, este foi o mesmo valor encontrado para a variância do estimador regressão para dados assimétricos, ou seja, para este caso ambos estimadores possuem a mesma precisão. O resultado para $\lambda = 5$ foram similares ao caso em que $\lambda = 0$, porém a variância do coeficiente linear de ambos estimadores regressão foi 0.0052 e para o caso em $\lambda = 10$ a variância de β foi 0.0054. Portando para todos os caso analisados o estimador regressão comum foi robusto a assimetria.

Os resultados obtidos da simulação para $n = 1000$ e considerando os diferentes valores de λ's encontra-se na tabela 4.2. Observa-se que mesmo com o aumento do tamanho amostral as estimativas da média são as mesmas para ambos estimadores regressão, diferenciando apenas das estimativas para $n = 500$. Para $\lambda = 0$ o verdadeiro valor da média é igual a 4.9819, portanto as estimativas dos três estimadores se comportam bem, porém quando analisado a variância destes, percebemos que os estimadores regressão são mais eficientes que o estimador simples, este resultado pode ser encontrado na literatura, Ferraz & Vieira (2009). Os resultados encontrados para $\lambda \neq 0$ são análogos aos resultados de $\lambda = 0$, diferenciando apenas as estimativas da média por serem de populações diferentes.

Figura 4.1: Histogramas da variável dependente para diferentes λ's e $n = 500$



Variável dependente com $\lambda = 0$ | Variável dependente com $\lambda = 5$ | Variável dependente com $\lambda = 10$

A variância dos coeficiente linear de ambos estimadores regressão quando $\lambda = 0$ foi 0.0009, e para $\lambda = 5$ esta variância foi 0.0027 e por fim para $\lambda = 10$ encontramos a variância igual a 0.0029. Note que houve um razoável crescimento da variância a medida que aumenta a assimetria, porém não encontramos nenhuma evidência de diferença entre os estimadores regressão.

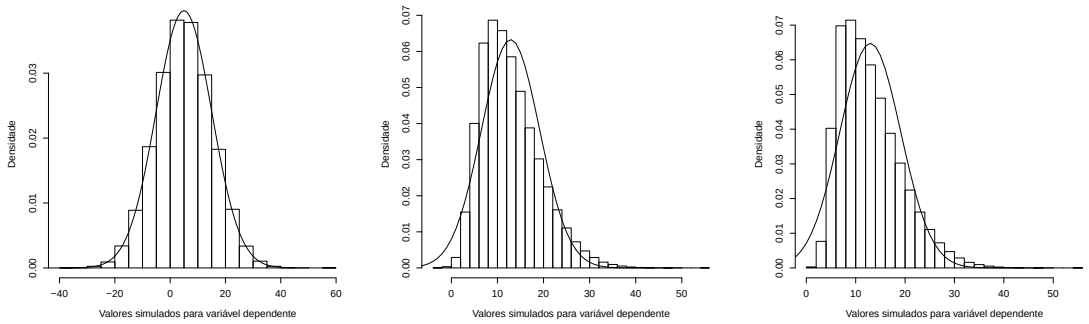
Tabela 4.2: Comparações dos estimadores para diferentes λ's e $n = 1000$.

	$\lambda = 0$			$\lambda = 5$			$\lambda = 10$		
Estimadores	M	V	CV	M	V	CV	M	V	CV
MS	4.9835	0.0947	0.0617	12.8391	0.0400	0.0156	12.9520	0.0384	0.0151
ERC	4.9819	0.0009	0.0061	12.8413	0.0009	0.0024	12.9541	0.0009	0.0024
ERA	4.9819	0.0009	0.0061	12.8413	0.0009	0.0024	12.9541	0.0009	0.0024

Comparando a variância dos estimadores para os diferentes n percebemos que a variância destes estimadores decrescem a medida que n cresce, este resultado já é esperado, pois para amostras maiores esperamos que a precisão melhore, portando foi exatamente isso que ocorreu. Fazendo a comparação entre os estimadores regressão, notamos mais uma vez que não houve diferença entre eles. Portanto afirmamos novamente que não temos evidências que o estimador regressão comum não seja robusto à assimetria.

Quanto a variância dos coeficientes linear para $n = 5000$ não encontramos diferenças entre os estimadores regressão, mesmo quando aumentamos o valor de λ. Os resultados encontrados para variância de β foram 0.0002, 0.0005 e 0.0005, para $\lambda = 0, 5, 10$ respectivamente.

Figura 4.2: Histogramas da variável dependente para diferentes λ's e $n = 1000$

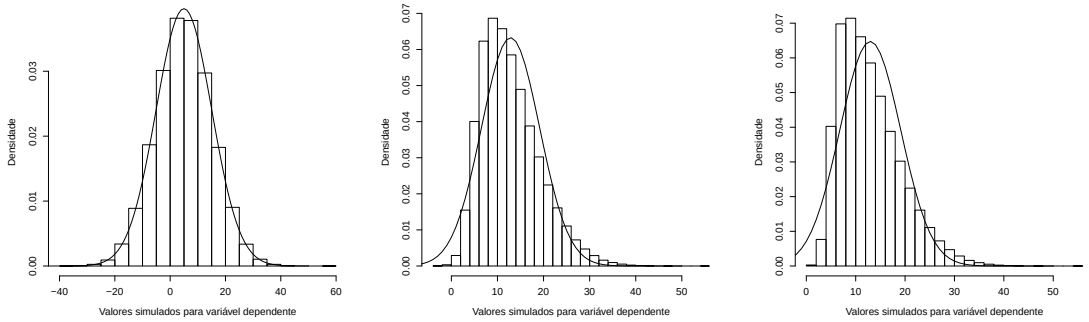


Variável dependente com $\lambda = 0$ | Variável dependente com $\lambda = 5$ | Variável dependente com $\lambda = 10$

Tabela 4.3: Comparações dos estimadores para diferentes λ's e $n = 5000$.

	$\lambda = 0$			$\lambda = 5$			$\lambda = 10$		
Estimadores	M	V	CV	M	V	CV	M	V	CV
MS	4.9777	0.0192	0.0279	12.8376	0.0074	0.0067	12.9499	0.0070	0.0065
ERC	4.9811	0.0002	0.0029	12.8406	0.0002	0.0011	12.9533	0.0002	0.0011
ERA	4.9811	0.0002	0.0029	12.8406	0.0002	0.0011	12.9533	0.0002	0.0011

Figura 4.3: Histogramas da variável dependente para diferentes λ's e $n = 5000$



Variável dependente com $\lambda = 0$ | Variável dependente com $\lambda = 5$ | Variável dependente com $\lambda = 10$

Na Tabela 4.4 apresentamos as médias dos coeficientes lineares estimados pelo método comum e pelo método que considera à assimetria dos dados. Como podemos ver as estimativas foram as mesmas quando consideramos quatro casas decimais. Dessa forma, em média a estrutura do estimador regressão para dados assimétricos não irá diferenciar

muito da estrutura do estimador regressão comum.

Tabela 4.4: Média dos coeficientes lineares dos estimadores regressão para diferentes valores de n e λ.

	$n = 500$			$n = 1000$			$n = 5000$		
	$\lambda = 0$	$\lambda = 5$	$\lambda = 10$	$\lambda = 0$	$\lambda = 5$	$\lambda = 10$	$\lambda = 0$	$\lambda = 5$	$\lambda = 10$
ERC	10.004	9.9991	9.9991	10.0037	9.9975	9.9973	10.0018	9.9977	9.9976
ERA	10.004	9.9992	9.9992	10.0037	9.9975	9.9973	10.0018	9.9977	9.9976

4.2 Aplicação aos dados reais

Nesta seção, utilizaremos os dados do Programa de Avaliação do Ciclo Básico de Alfabetização (Proalfa) que tem como objetivo principal mensurar o desempenho em Língua Portuguesa de crianças em fase de alfabetização no estado de Minas Gerais.(Vieira & Souza, 2008)

O Proalfa é aplicado aos alunos das fases I, II e III do ensino fundamental matriculados em escolas das redes estadual e municipal, em Minas Gerais. Esta pesquisa tem como objetivo avaliar de forma censitária alunos matriculados em escolas públicas na fase II do ciclo inicial de alfabetização do ensino fundamental da rede estadual ou 2ª série do ensino fundamental das redes municipais. Já as fases I e III são avaliadas de maneira amostral. O Programa faz parte do Sistema Mineiro de Avaliação da Educação Pública (Simave) e foi desenvolvido por meio da parceria entre a Secretaria de Estado da Educação (SEE), o Centro de Políticas Públicas e Avaliação da Educação (Caed), da Universidade Federal de Juiz de Fora (UFJF), orgão este que cedeu os dados para análise, e o Centro de Alfabetização, Leitura e Escrita (Ceale), da Universidade Federal de Minas Gerais (UFMG). Para o nosso estudo estaremos interessados nos alunos da fase II do ano de 2008, onde teremos informações para toda a população.

Na tabela 4.5, apresentamos algumas estatísticas descritivas da proficiência dos alunos em Língua Portuguesa por rede de ensino e gênero, e por ela podemos verificar que a média das escolas estaduais é superior a média das escolas municipais, o mesmo ocorre para o gênero, onde a crianças do gênero feminino tem a proficiência média superior às crianças do gênero masculino.

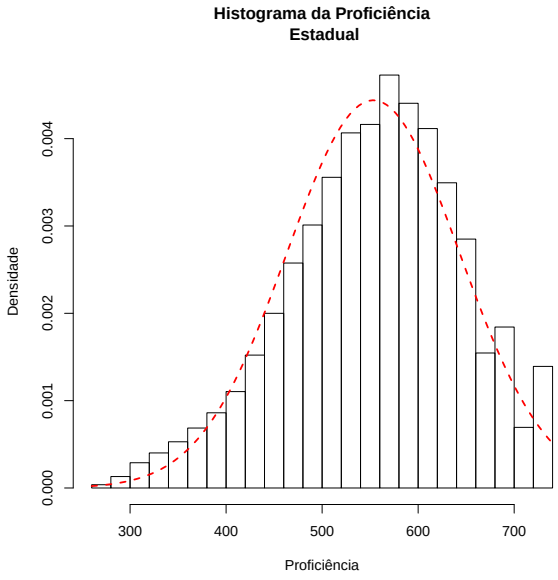
A Figura 4.4 representa a distribuição da proficiência em Língua Portuguesa

Tabela 4.5: Análise descritiva da proficiência dos alunos em Língua Portuguesa das escolas estaduais

	Média	Mediana	Desvio padrão
Estadual	553.25	560.30	89.86
Municipal	517.11	519.03	91.88
Masculino	522.61	525.52	94.12
Feminino	541.23	545.07	90.31
Total	531.54	535.07	92.78

dos alunos de escolas estaduais. Pelo histograma podemos notar que há uma certa assimetria negativa na distribuição dos dados, e com o auxílio da curva normal com média e variância das proficiências, notamos o deslocamento da distribuição. A partir destas evidências avaliaremos o desempenho dos três estimadores sobre os dados da proficiência dos alunos em Língua Portuguesa, considerando como variável auxiliar o número de vezes em que o aluno repetiu a série. A escolha desta variável auxiliar se deu ao fato desta está correlacionada com a proficiência do aluno em Língua Portuguesa. Esta correlação foi verificada pela Correlação de Pearson que teve p-valor<0.001.

Figura 4.4: Histograma da proficiência dos alunos em Língua Portuguesa



Utilizando os dados dos alunos das escolas estaduais selecionamos 1000 amostras segundo o método de Amostragem Aleatória Simples sem Reposição de tamanho 500 e calculamos as estimativas da média para cada amostra utilizando o estimador simples, ERC e ERA, e logo adiante calculamos a média e a variância das médias estimadas.

Tabela 4.6: Comparações dos estimadores para valores diferentes de n.

n	Estimadores	M	V	CV
	MS	553.1195	16.3015	0.0073
500	ERC	553.1479	16.1897	0.0073
	ERA	553.1461	16.2086	0.0073
	MS	553.0806	7.9888	0.0051
1000	ERC	553.0812	7.9703	0.0051
	ERA	553.0819	7.9717	0.0051
	MS	553.2590	1.4678	0.0022
5000	ERC	553.2670	1.4503	0.0022
	ERA	553.2660	1.4500	0.0022

Pela tabela 4.2 verificamos que com o aumento do tamanho amostral as estimativas de todos os estimadores aproxima cada vez mais do valor real (553,25), resultado este já esperado. Quanto à variância dos estimadores regressão podemos afirmar que foram relativamente próximas quando $n = 500$, porém, quando comparada com o estimador simples notamos uma diferença pequena entre eles. Para os tamanhos amostrais 1000 e 5000 as variância dos estimadores foram muito próximas.

Analisando a variância dos $\hat{\beta}$'s dos estimadores regressão, para $n = 500$, notamos que há um ganho de precisão quando utilizamos o estimador regressão para dados assimétricos, pois a variância do coeficiente linear calculada é 44.5719, enquanto que para o estimador regressão comum foi 46.9047, portanto há um ganho de aproximadamente 5% quando utilizamos o estimador regressão para dados assimétricos.

Para $n = 1000$, a variância dos $\hat{\beta}$'s se aproximam um pouco, mas ainda sim há um ganho de aproximadamente 3% quando usamos o estimador regressão para dados assimétricos. A variância dos $\hat{\beta}$'s para o ERC e ERA foram 22.0556 e 21.4307, rescpectivamente. Já para $n = 5000$ o ganho foi de aproximadamente 4% quando utilizamos o ERA. As estimativas da variância dos $\hat{\beta}$'s foram 4.4632 e 4.2779 para ERC e ERA, respectivamente.

5 Conclusão

Neste trabalho, apresentamos uma nova alternativa para os estimadores do tipo regressão e o nomeamos de estimador regressão para dados assimétricos, onde este considera em sua estrutura o coeficiente linear baseado em dados assimétricos normal assimétrico.

Os resultados do estudo de simulação mostraram que para valores diferentes de n e diferentes λ's não houve ganho de precisão nas estimativas quando utilizado o estimador proposto, porém no estudo empírico obtivemos ganhos de precisão. Este resultado contraditório pode ser devido às simulações, pois nestas tanto a variável dependente quanto a variável auxiliar são assimétricas, portanto o estimador regressão comum incorpora de certa forma a assimetria em suas estimativas, enquanto que os dados reais à assimetria é incorporada apenas na variável dependente. Uma hipótese para os resultados parecidos dos estimadores regressão é que a principal diferença do modelo regressão assimétrico para o modelo de regressão comum está no intercepto, e no estimador do tipo regressão consideramos apenas o coeficiente linear dos modelos, e estes coeficientes são relativamente parecidos.

Referências Bibliográficas

[1] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal Statistics*, **12**, 171-178.

[2] Bazán, J.L.G. (2005). Uma família de modelos de resposta ao ítem normal assimétricas. Tese de doutorado. IME - Universidade de São Paulo.

[3] Bolfarine, H., Bussab, W. O. (2005). *Elementos de amostragem*. São Paulo: Blucher, 2005. (ABE - Projeto Fisher), 145-156.

[4] Ferraz, C., Vieira, M. D. T. (2009). *Amostragem assistida por modelos lineares*. 8^o Encontro Mineiro de Estatística - 10 e 11 de setembro - Juiz de Fora - MG.

[5] Ferreira, C. D. S., *Inferência e Diagnóstico em Modelos Assimétricos*. Tese de doutorado, Departamento de Estatística, IME-USP. São Paulo.

[6] Freitas, L.A. (2005). *Modelo de regressão com erros normais assimétricos: uma abordagem bayesiana*. Dissertação de Mestrado, Departamento de Estatística, Universidade Federal de São Carlos. São Carlos.

[7] Gómez, H.W. (2005). *Extensiones Asimétricas de Distribuciones Simétricas: Propiedades e Inferência*. Tese de Doutorado. Facultad de Matemáticas, Pontifícia Universidad Católica de Chile. Santiago, Chile.

[8] Henze, N. (1986). A probabilistic representation of the "skew-normal" distribution. *Scandinavian Journal of Statistics*, **13**, 271-275.

[9] Lachos, V.H. (2004). *Modelos lineares mistos assimétricos*. Tese de Doutorado, Departamento de Estatística, IME-USP. São Paulo.

[10] Lin, T.I., Lee, J.C. e Yen, S.Y. (2007). Finite mixture modeling using the skew normal distribution. *Statistica Sinica*, **17**, 909-927.

[11] Pessoa, D. G. C., Costa, A. W. N. D. (2009). *Experimentos com amostragem e estimação usando o R*. II Escola de Amostragem e Metodologia de Pesquisa - 29 de setembro - Natal - RN.

[12] Rodríguez, C.L.B. (2005). *Inferência bayesiana no modelo normal assimétrico*. Dissertação de mestrado. Departamento de Estatística, IME-USP. São Paulo.

[13] Vieira, M. D. T; Souza, A. C. Plano Amostral da Pesquisa do PROALFA de 2008. Relatório Técnico. Juiz de Fora: Departamento de Estatística, UFJF, 2008.