MODELAGEM DA TAXA DE ANALFABETISMO NO ESTADO DA PARAÍBA VIA MODELO DE REGRESSÃO BETA

Camila Ribeiro da SILVA[1]
Tatiene Correia de SOUZA[1]

▪ RESUMO: Este artigo tem como objetivo avaliar os fatores que influenciam a taxa de analfabetismo no Estado da Paraíba no ano de 2010. Utilizamos o modelo de regressão beta proposto por Ferrari & Cribari–Neto (2004) com a finalidade de modelar a taxa de analfabetismo nos municípios paraibanos. Adicionalmente, calculamos o impacto do gasto com assistencialismo do programa Bolsa Família na taxa de analfabetismo, e verificamos que é necessário um investimento per capita de aproximadamente R$ 411,00 reais, para que se busque uma redução na taxa de analfabetismo.

▪ PALAVRAS-CHAVE: Assistencialismo; modelo de regressão beta; taxa de analfabetismo.

1 Introdução

O analfabetismo se constitui em um dos fundamentais problemas da sociedade brasileira e, consequentemente, é um dos temas mais debatidos quando se discutem políticas sociais. A taxa de analfabetismo é um índice que há muito desafia os brasileiros, estando presente há muito tempo na sociedade. Os avanços tecnológicos, as mudanças pelas quais passaram o mundo e o Brasil em particular amenizaram esse problema, mas não conseguiram extraí-lo de uma vez por todas de nosso País. Se apenas a educação não transforma a sociedade, sem ela tampouco a sociedade muda, defendeu Freire (1979).

Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), atualmente, no Brasil existem aproximadamente 14 milhões de analfabetos. A maior parte se encontra na região Nordeste, em municípios com até 50 mil habitantes, na população com mais de 15 anos, entre negros e pardos e na zona rural. Os dados do censo

[1]Universidade Federal da Paraíba – UFPB, Departamento de Estatística, CEP: 58089–900, João Pessoa, PB, Brasil. E-mail: *camilaribeiroufpb@hotmail.com*, *tatiene@de.ufpb.br*

2010 mostram uma redução de 29% em relação aos números apresentados em 2000, mas ainda insatisfatória, especialmente, quando considerados os critérios utilizados pelo IBGE, que considera alfabetizada a pessoa capaz de ler e escrever um bilhete simples.

A taxa de analfabetismo na Região Nordeste, reconhecida historicamente por ter o maior número de iletrados do país, caiu de 22,4% (2004) para 18,7% (2009). A informação foi divulgada pelo IBGE que registrou em todo o país uma redução do número de analfabetos em 2009. O estudo realizado em 2011 aponta que 12,9 milhões de brasileiros com mais de 15 anos de idade não sabem ler nem escrever. Destes, 6,8 milhões estão na região Nordeste, que possui taxa de analfabetismo de 16,9%, quase o dobro da média nacional, de 8,6%.

No Estado da Paraíba, o analfabetismo é uma realidade vivida por cerca de 21,6% dos paraibanos com 15 anos ou mais, afirma o Instituto de Pesquisa Econômica Aplicada (IPEA). Segundo os dados, a Paraíba é o terceiro Estado do país com o maior índice de analfabetos, e ocupa a terceira posição, entre as unidades da Federação, com a menor média de anos de estudo, estando atrás dos Estados do Piauí (5,8) e Alagoas (5,7). A Unidade da Federação com a maior média de anos estudados é o Distrito Federal (9,6), seguida de São Paulo (8,5).

Diversos programas de combate ao analfabetismo têm sido implementados nos últimos anos, principalmente nos âmbitos federal e estadual. No entanto, as taxas de analfabetismo no Brasil, apesar de terem se reduzido nos últimos anos, ainda apresentam níveis elevados, principalmente nas regiões Norte e Nordeste. Além dos programas assistenciais voltados ao amparo educacional, há também os programas assistenciais de transferência de renda. No país, um dos principais programas de transferência de renda é o Bolsa Família (BF). Criado em 2003 durante o governo do presidente Lula, o programa Bolsa Família é reconhecido internacionalmente como o maior programa de transferência de renda do mundo, atendendo atualmente a 13,8 milhões de famílias. Segundo Tereza Campelo, ministra do Desenvolvimento Social e Combate à Fome, durante cerimônia de comemoração dos dez anos do programa, as ações do programa Bolsa Família têm gerado resultados positivos não só para a redução da extrema pobreza no Brasil, mas também para diversos setores estratégicos do governo, como saúde e educação(disponível em `http://www.sae.gov.br/site/?p=18894)`.

O nosso objetivo no presente artigo é explicar a taxa de analfabetismo no Estado da Paraíba a partir de variáveis geográficas e socioeconômicas. Para tanto, se faz necessário o uso de modelos adequados para situações em que a variável resposta esta restrita ao intervalo $(0, 1)$. O modelo de regressão beta foi proposto por Ferrari & Cribari–Neto (2004) como uma forma de suprir algumas das limitações associadas aos modelos tradicionais (modelo de regressão linear), principalmente, no que se refere a estrutura da variável reposta. A classe de modelos de regressão beta tem como objetivo permitir a modelagem de respostas que pertencem ao intervalo (0,1), por meio de uma estrutura de regressão que contém uma função de ligação, covariáveis e parâmetros desconhecidos. Adicionalmente, estimamos o impacto do gasto com o programa Bolsa Família no Estado da Paraíba sobre a taxa de

analfabetismo. O nosso interesse aqui reside em estimar qual a quantia que deveria ser destinada ao assistencialismo para que o cenário da taxa de analfabetismo fosse revertido no Estado da Paraíba. Vale ressaltar aqui que os resultados e conclusões apresentados neste artigo foram baseados na taxa de analfabetismo dos 223 municípios do Estado da Paraíba no ano de 2010.

Além desta introdução, este artigo está dividido em cinco seções. Na Seção 2, apresentamos o modelo de regressão beta, bem como algumas medidas de avaliação da qualidade do ajuste. Na Seção 3, descrevemos os dados utilizados. Na Seção 4, especificamos o modelo de regressão beta considerado, bem como as inferências associadas. Por fim, a seção 5 conclui o artigo.

2 O modelo de regressão beta

A classe de modelos de regressão beta tem como objetivo permitir a modelagem de respostas que pertencem ao intervalo $(0, 1)$, por meio de uma estrutura de regressão que contém uma função de ligação, covariáveis e parâmetros desconhecidos. Muitos estudos, em diferentes áreas do conhecimento, como em Brehm & Gates (1993), Hancox et al. (2010), Kieschnick & Mccullough (2003), Smithson & Verkuilen (2006), utilizam regressão beta ou outras abordagens para examinar como um conjunto de covariáveis se relaciona com alguma percentagem ou proporção. Em tais modelos, assume-se que a resposta média é relacionada com um preditor linear por meio de uma função de ligação. O preditor linear envolve covariáveis e parâmetros de regressão desconhecidos. Estes modelos também são indexados por um parâmetro de dispersão, que em certas situações pode variar ao longo das observações (Smithson & Verkuilen, 2006; Espinheira et al. 2008a, 2008b; Cribari–Neto & Souza, 2012; Souza & Cribari–Neto, 2013).

Ferrari & Cribari-–Neto (2004) propuseram uma parametrização alternativa para a densidade beta que permite a modelagem da média da resposta através de uma estrutura de regressão e que envolve também um parâmetro de precisão. A função de densidade beta nessa reparametrização tem a forma

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \qquad (1)$$

em que $0 < \mu < 1$ e $\phi > 0$. Aqui, $\mathrm{E}(y) = \mu$ e $\mathrm{var}(y) = \frac{V(\mu)}{1+\phi}$, sendo $V(\mu) = \mu(1-\mu)$ a função de variância, μ é a média da variável resposta e ϕ pode ser interpretado como o parâmetro de precisão.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, em que cada y_i, $i = 1, \ldots, n$, segue a densidade da Equação (1) com média μ_i e parâmetro de precisão ϕ_i sendo desconhecidos, o modelo de regressão beta assume que a média satisfaz a seguinte relação funcional

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = \eta_i,$$

em que $\beta = (\beta_1, \ldots, \beta_k)^\top$ é um vetor de parâmetros de regressão desconhecidos ($\boldsymbol{\beta} \in \mathbb{R}^k$), $x_{i1}, \ldots, x_{ik}$ são observações de k covariáveis ($k < n$), η_i é o preditor linear e $g(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável, com domínio em (0,1) e imagem em $\mathbb{R}$, denominada de função de ligação. Além disso, podemos considerar ainda que o parâmetro de dispersão ϕ_i varia ao longo das observações (Simas et al., 2010). Desse modo, podemos admitir que o parâmetro de precisão é dado por

$$h(\phi_i) = \sum_{j=1}^{q} z_{ij}\gamma_j = \vartheta_i,$$

em que, $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ é um vetor de parâmetros desconhecido ($\gamma \in \mathbb{R}^q$), $z_{i1}, \ldots, z_{iq}$ são observações de q covariáveis ($q < n$), assumidas fixas e conhecidas e $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável. Existem várias possíveis escolhas para a função de ligação $g(\cdot)$ e $h(\cdot)$. Entre elas, podemos utilizar a especificação logito

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right),$$

ou a função probito

$$g(\mu) = \Phi^{-1}(\mu),$$

em que $\Phi(\cdot)$ é a função acumulada da distribuição normal padrão, entre outras. Para maiores detalhes sobre funções de ligação, ver McCullagh & Nelder (1989).

O logaritmo da função de verossimilhança de n observações independentes é

$$\ell(\beta, \gamma) = \sum_{i=1}^{n} \ell_i(\mu_i, \phi_i),$$

em que

$$\begin{aligned} \ell_i(\mu_i, \phi_i) &= \log\Gamma(\phi_i) - \log\Gamma(\mu_i\phi_i) - \log\Gamma((1-\mu_i)\phi_i) \\ &+ (\mu_i\phi_i - 1)\log y_i + \{(1-\mu_i)\phi_i - 1\}\log(1-y_i). \end{aligned}$$

Como os estimadores de máxima verossimilhança de β e γ não possuem forma fechada, eles precisam ser obtidos numericamente maximizando a função de log-verossimilhança através de algum algoritmo de maximização não-linear.

Sob certas condições de regularidade, temos que, para tamanhos de amostras grandes, a distribuição conjunta de $\hat{\beta}$ e $\hat{\gamma}$ é aproximadamente normal $(k+q)$ multivariada,

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \sim \mathcal{N}_{k+q}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, K^{*-1}\right),$$

aproximadamente, sendo $\hat{\beta}$ e $\hat{\gamma}$ os estimadores de máxima verossimilhança de β e γ, respectivamente.

2.1 Algumas medidas de diagnóstico e adequabilidade para modelos de regressão beta.

Uma etapa importante na análise de um ajuste de regressão é a verificação de possíveis afastamentos das suposições feitas para o modelo, bem como a existência de observações extremas que podem causar desvios nos resultados do ajuste. Desse modo, uma etapa imprescindível na análise de regressão é a validação do modelo ajustado.

Espinheira et al. (2008a) propuseram diferentes tipos de resíduos para o modelo de regressão beta com dispersão constante e constataram que o mesmo tipicamente apresenta desempenho superior ao resíduo proposto por Ferrari & Cribari–Neto (2004), especialmente no sentido de identificar observações influentes para as estimativas das médias. Ferrari et al. (2011) apresentaram uma modificação para o resíduo ponderado padronizado 2 considerando o modelo de regressão beta com dispersão variável. O resíduo proposto por esses autores, r_{pp}, é dado por

$$r_i^{pp} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\nu_i(1-h_{ii}^*)}},$$

em que $y_i^* = \log\left\{\frac{y_i}{(1-y_i)}\right\}$, $\mu_i^* = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i)$, em que $\psi(\cdot)$ é a função digama. Adicionalmente, h_{ii}^* o i-ésimo elemento de $H^* = (W\Phi)^{1/2}X(X^\top\Phi W X)^{-1}X^\top(\Phi W)^{1/2}$, em que X é uma matriz $n \times k$ de covariáveis $(k < n)$, $\Phi = \text{diag}\{\phi_1, \ldots, \phi_n\}$ e $W = \text{diag}\{w_1, \ldots, w_n\}$, com $w_i = \phi_i\nu_i\frac{1}{g'(\mu_i)^2}$, em que $\nu_i = \psi'(\mu_i\phi_i) + \psi'((1-\mu_i)\phi_i)$.

Além da análise dos resíduos, uma medida global da qualidade do ajuste proposta por Ferrari & Cribari–Neto (2004) pode ser obtida através do cálculo do pseudo-R^2, definido como o quadrado do coeficiente de correlação entre $\hat{\eta}$ e $g(y)$. Na evolução dos métodos de diagnóstico uma etapa que se mostrou relevante foi a detecção de observações que exercem efeito desproporcional no ajuste, podendo interferir inclusive em resultados inferenciais. Neste contexto, encontram-se a distância de Cook e alavancagem generalizada (Rocha & Simas, 2011). A distância de Cook (1977) visa medir o impacto de uma observação particular nas estimativas dos coeficientes da regressão a partir de sua exclusão do conjunto de dados. A distância de Cook para o modelo de regressão beta apresentada em Espinheira et al. (2008a), é dada por:

$$C_i = \frac{h_{ii}^*(r_i^{pp})^2}{(1-h_{ii}^*)^2}.$$

3 Descrição dos dados

Uma breve descrição das variáveis consideradas neste estudo está apresentada na Tabela 1. As fontes dos dados são o Atlas do Desenvolvimento Humano no Brasil (2013) e o Ministério do Desenvolvimento social e Combate à Fome.

Tabela 1 - Descrição das variáveis utilizadas.

Variável	Definição
Taxa	Taxa de analfabetismo (18 anos ou mais de idade).
MI	Mortalidade infantil: Número de crianças que não deverão sobreviver ao primeiro ano de vida em cada 1.000 crianças nascidas vivas.
Renda	Renda per capita: Razão entre o somatório da renda de todos os indivíduos residentes em domicílios particulares permanentes e o número total desses indivíduos.
Gini	Índice de Gini: Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita.
AE	Percentual da população em domicílios com banheiro e água encanada.
Lixo	Percentual da população em domicílios com coleta de lixo.
Densidade	Percentual da população em domicílios com densidade populacional maior que 2.
PP	Percentual de pobres.
PR	População rural (mil habitantes).
PU	População urbana (mil habitantes).
Gasto	Gasto com assistencialismo per capita: Razão entre o gasto (em reais) com o programa de transferência de renda (Bolsa família) e a população do município.

Na Tabela 2 apresentamos algumas estatísticas descritivas: mínimo, máximo, média, mediana e desvio padrão das variáveis utilizadas. Algumas conclusões podem ser extraídas da Tabela 1. O valor médio da taxa de analfabetismo foi de 0,319 com um desvio padrão de 0,068. Os valores extremos verificados foram respectivamente, 0,085 referente à capital João Pessoa e 0,461 referente ao município de Pedro Régis.

O valor médio para variável mortalidade infantil (MI) foi de 26,69, ou seja, em média, aproximadamente 27 crianças deverão não sobreviver até um ano de idade a cada mil nascidas vivas. Com relação a renda per capita registramos o valor médio igual a R$ 277,35, sendo os valores mínimo e máximo iguais a R$ 166,28 e R$ 1.036,21, respectivamente, conforme apresentado na Tabela 2.

Tabela 2 - Estatísticas descritivas das variáveis utilizadas.

Variável	mínimo	máximo	média	mediana	desvio padrão
Taxa	0,085	0,461	0,319	0,324	0,068
MI	15,270	44,000	26,689	26,000	5,608
Renda	166,28	1036,21	277,35	263,84	92,08
Gini	0,400	0,700	0,500	0,500	0,043
AE	3,260	97,150	64,268	67,230	17,731
Lixo	61,370	100,000	94,689	97,000	6,507
Densidade	12,220	50,850	31,171	31,170	6,353
PP	11,590	60,980	39,113	39,340	7,932
PR	8000	18813	4160,76	3175	3285,73
PU	4730	720785	12729,5	3818	5508,5
Gasto	88,22	224,17	172,39	172,31	22,85

Considerando a variável *Gini* temos que o valor médio foi igual a 0,5, um indicativo de que os municípios do Estado da Paraíba apresentam desigualdade com relação a distribuição de renda. Verificamos que, em média, 64,27% dos municípios da amostra possuem domicílios com banheiro e água encanada, enquanto para a variável *Densidade* a média foi igual a 31,17%. Com relação ao percentual de pobres (PP) nos municípios, 39,11% dos municípios apresentam população que se enquadra nesta situação (ver Tabela 2).

A população rural (PR) média observada nos municípios foi igual a 4160 habitantes, enquanto a população urbana (PU) a média foi de 12.729,5 habitantes. O gasto médio com assistencialismo (*Gasto*) nos municípios da Paraíba foi de R$ 172,39, com mínimo e máximo de R$ 88,22 e R$ 224,17, respectivamente. Quanto a variável percentual da população com coleta de lixo (*Lixo*), observamos que em média 94,69% dos municípios observados possuem o serviço de coleta de lixo (ver Tabela 2) .

4 Especificação do modelo

Nesta seção apresentamos uma modelagem empírica relacionada à taxa de analfabetismo no Estado da Paraíba. Como a variável resposta taxa de analfabetismo é restrita ao intervalo $(0, 1)$ e exibe assimetria, utilizaremos a classe de modelos de regressão beta proposta por Ferrari & Cribari–Neto (2004) que assume para a variável resposta distribuição beta, denotada por $B(\mu_i, \phi)$, além de uma relação não linear entre a média da variável resposta e as variáveis explicativas. O procedimento computacional foi desenvolvido utilizando o pacote

`betareg` (Cribari–Neto & Zeleis, 2010) do *software* estatístico `R` . Na seleção dos modelos utilizamos o critério de seleção de modelos AIC (Akaike's information criterion), que foi proposto por Akaike (1974), o BIC e o pseudo-R^2.

Na Figura 1 estão apresentados o histograma e o box-plot da variável taxa de analfabetismo. É possível observar que a distribuição dessa variável é assimétrica, portanto, é necessário considerar um modelo adequado, i.e., um modelo que capture essa assimetria. Assim sendo, utilizamos o modelo de regressão beta.

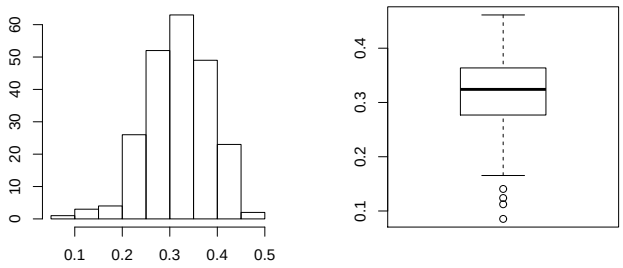

Figura 1 - Histograma e gráfico *box-plot* da taxa de analfabetismo no Estado da Paraíba.

Inicialmente, selecionamos o modelo mais adequado para explicar a taxa de analfabetismo. Para tanto, considerou-se algumas possibilidades de modelos, com o objetivo de verificar qual a função de ligação mais adequada para a modelagem. Após a seleção, verificamos que o modelo mais adequado para explicar a taxa de analfabetismo foi o modelo com função de ligação logit.

Com o objetivo de reduzir possíveis erros de especificação do modelo estimado, realizou-se os testes da razão de verossimilhança e Wald, a fim de verificar a suposição de dispersão constante. Na Tabela 3 são sumarizados os resultados desses testes.

Tabela 3 - Teste da Razão de verossimilhança e Wald.

Teste	Estatística	p-valor
Razão de verossimilhança	7,6269	0,0221
Wald	2401,8	$< 1 \times 10^{-10}$

Observa-se que a suposição de dispersão constante é rejeitada ao nível de significância de 5% para ambos os testes, indicando que o parâmetro ϕ deve

ser modelado explicitamente através das covariáveis. Vale mencionar que para a realização dos testes foi considerado o seguinte modelo para a estrutura de regressão da precisão:

$$\phi_i = \exp\left(\gamma_1 + \gamma_2 PP_i + \gamma_3 PR_i\right), i = 1, \ldots, 223.$$

Dessa forma, realizou-se uma modelagem do parâmetro de precisão, cujas estimativas dos parâmetros, erros-padrão e p-valores do modelo final são apresentados na Tabela 4. Vale mencionar que apenas a função de ligação log foi considerada para modelagem do parâmetro de precisão. A partir do ajuste observou-se que as variáveis PP e PR foram significativas ao nível de 5% para o modelo da precisão.

Tabela 4 - Estimativas pontuais e p-valores.

Covariável	$\hat{\beta}$	p-valor	$\hat{\gamma}$	p-valor
Intercepto	$-1,8113$	$< 1 \times 10^{-5}$	5,2580	$< 1 \times 10^{-10}$
Renda	$-0,0026$	$< 1 \times 10^{-6}$	–	–
Gini	2,1158	$< 1 \times 10^{-7}$	–	–
Densidade	0,0075	0,0027	–	–
Gasto	0,0027	0,0004	–	–
PP	–	–	$-0,0234$	0,0499
PR	–	–	7×10^{-5}	0,0250

Através da análise dos coeficientes estimados para o modelo selecionado (ver Tabela 4) é possível verificar que as covariáveis *Gini*, *Densidade* e *Gasto*, influenciam positivamente a taxa de analfabetismo. O sinal positivo do coeficiente da variável *Gini* (índice de Gini) indica que um acréscimo no índice de Gini do município corresponde a um aumento na taxa de analfabetismo, quando as demais variáveis são mantidas constantes. Isto porque o mesmo mede a desigualdade de renda e disparidades sociais. De forma similar, o sinal positivo do coeficiente da variável *Densidade* (percentual da população em domicílios com densidade maior que 2) implica que municípios com alta densidade apresentam alta taxa de analfabetismo.

Para a variável *Gasto* (gasto com assistencialismo), o sinal positivo do parâmetro indica que um incremento nessa variável implica no aumento da taxa de analfabetismo, mantendo-se as demais covariáveis fixas. Resultado um tanto curioso, pois espera-se que quanto maior a renda de uma família ou indivíduo mais elevado seria seu grau de instrução. No entanto, não foi isso que observamos no resultado em pauta. Segundo o Censo 2010 (IBGE) há uma forte correlação negativa entre analfabetismo e renda: quanto menor o rendimento médio do município, maior

a sua taxa de analfabetismo, e vice-versa. Contudo, não há relação direta de causa e efeito. Os dados mostram que não basta transferir dinheiro para as famílias mais pobres para reduzir o analfabetismo. Conforme os dados (Censo 2010), programas como o Bolsa Família, conseguiram promover a redução da taxa de analfabetismo da população em idade escolar (de 10 a 14 anos), porque as famílias com filhos nessa faixa etária têm que comprovar que as crianças estão frequentando a escola para receber o benefício. Em contrapartida, o efeito da transferência de renda é nulo sobre a melhoria educacional da população acima de 15 anos, que não tem a obrigação legal de ir à escola.

Por outro lado, a covariável *Renda* (renda per capita), exerce efeito negativo na taxa de analfabetismo, isto é, municípios com maior renda per capita tendem a apresentar uma menor taxa de analfabetismo.

Considerando a estrutura de regressão para precisão, temos que à medida que a covariável PP (percentual de pobres) aumenta, a precisão diminui, ou seja, os municípios que apresentam maior percentual de pobres tendem a apresentar respostas menos precisas. Em contrapartida, a covariável PR (população rural) exerce um efeito positivo na taxa de analfabetismo, ou seja, podemos dizer que quanto maior a população rural do município mais precisas serão as respostas.

Com o intuito de verificar possíveis afastamentos das suposições feitas para o modelo, a Figura 2 apresenta os gráficos dos resíduos ponderados padronizados 2 versus os índices das observações e também o gráfico da probabilidade normal com envelopes simulados. O modelo de regressão parece estar bem ajustado, dado a distribuição dos resíduos encontram-se dentro dos limites $(-3, 3)$, além disso, os resíduos permanecem dentro das bandas de confiança dos envelopes simulados, indicando que não há indícios de afastamento da suposição de que o modelo de regressão beta é adequado para os dados.

Com o intuito de avaliar se o modelo ajustado está corretamente especificado, realizou-se o teste Reset, considerando a segunda potência do valor ajustado. O p-valor do teste foi 0,8417, ao nível de significância de 5% não rejeitamos a hipótese nula de que o modelo encontra-se bem especificado. O pseudo-R^2 do modelo final foi igual 0,5732, indicando que aproximadamente 57,32% da variabilidade da variável resposta pode ser atribuída as covariáveis apresentadas.

Com o objetivo de complementar a análise de resíduos realizada anteriormente, construímos os gráficos da distância de Cook versus os valores preditos e da alavancagem generalizada versus valores preditos, que estão apresentados na Figura 3. Notamos que no Gráfico da distância de Cook a observação 40, que corresponde ao município de Cabedelo, encontra-se destacada das demais. No Gráfico da alavancagem generalizada, duas observações se destacaram em relação às demais, são elas: 50 e 94, que correspondem aos municípios de Campina Grande e João Pessoa, respectivamente. Salientando que esses dois municípios são alguns dos que apresentaram uma das menores taxas de analfabetismo na amostra. Excluímos individualmente as observações 40, 50 e 94, contudo as variações percentuais nas estimativas dos parâmetros foram relativamente pequenas.

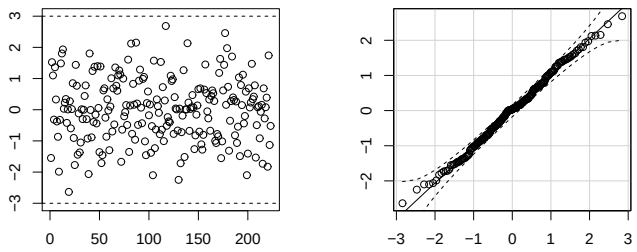

Figura 2 - Gráfico dos resíduos ordinários padronizados 2 versus índices das observações e gráfico da probabilidade normal com envelopes simulados.

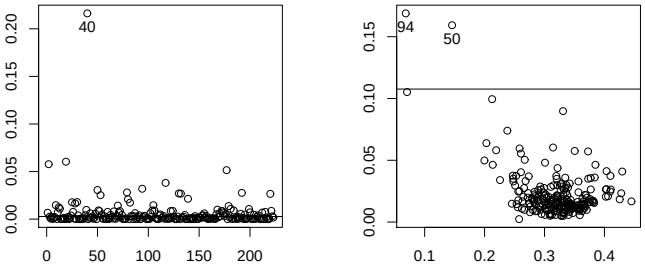

Figura 3 - Gráfico da distância de Cook e da alavancagem generalizada.

Como visto anteriormente, verificamos que a estimativa do parâmetro da variável gasto com assistencialismo (*Gasto*) foi positivo, indicando uma contribuição positiva para a variável resposta. No entanto, torna-se curioso o fato dessa covariável apresentar este tipo de comportamento, pois, os programas de transferência de renda são considerados importantes mecanismos para o enfrentamento da pobreza e como possibilidade de dinamização das relações sociais, principalmente nos pequenos municípios do país. Partindo desse ponto de vista, foi realizada uma investigação mais detalhada da variável gasto com assistencialismo. O objetivo seria investigar o comportamento dessa variável considerando um novo cenário, ou melhor, verificar o impacto estimado na taxa de analfabetismo a medida que o gasto com assistencialismo aumenta. Então, para calcular o impacto da covariável *Gasto* sobre a taxa de analfabetismo, precisamos levar em consideração tal transformação, ou seja

$$\frac{\partial E(y_i)}{\partial Gasto_i} = \frac{\partial \mu_i}{\partial Gasto_i}. \tag{2}$$

Considerando a função de ligação logit, a expressão (2), com as covariáveis selecionadas para o modelo é dada por

$$\frac{\partial E(y_i)}{\partial Gasto_i} = \beta_5 \frac{\exp(\beta_1 + \beta_2 Renda_i + \beta_3 Gini_i + \beta_4 Densidade_i + \beta_5 Gasto_i)}{(1 + \exp(\beta_1 + \beta_2 Renda_i + \beta_3 Gini_i + \beta_4 Densidade_i + \beta_5 Gasto_i))^2}.$$

Para mensurar este impacto, consideramos algumas situações diferentes, são elas: os valores gerados para a variável gasto com assistencialismo (*Gasto*) variaram de R$ 85,00 a R$ 600,00, a covariável *Renda* foi fixada no primeiro, segundo e terceiro quartis, enquanto que as outras covariáveis são fixadas na mediana.

Desse modo, foi possível obter uma maior compreensão dessa variável em relação aos dados observados. Com base na Figura 4 é possível extrair algumas conclusões. Verifica-se que a contribuição dessa variável é crescente para valores inferiores a R$ 411,00 reais.

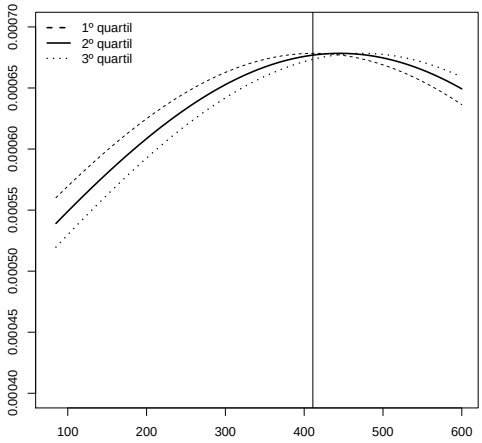

Figura 4 - Gráfico do gasto com assistencialismo per capita versus impacto estimado na taxa de analfabetismo.

Por outro lado, para valores superiores a esse valor estimado, observa-se que o impacto causado na variável resposta pelo gasto é decrescente. Vale ressaltar que este impacto é maior para a população com renda per capita de até R$ 230,12 (1º quartil). O que faz todo sentido, uma vez que, as famílias com menor renda tendem a apresentar uma necessidade maior desse benefício, em comparação aos outros

grupos (2º quartil e 3º quartil). Em relação ao grupo da população que possui uma renda per capita igual ou superior a R$ 293,82 (3º quartil), o impacto causado é menor. Isto porque esse grupo da população, geralmente, é o que apresenta melhores condições de vida (em termos de saúde, educação, etc.), e que não depende, de forma direta, desse tipo de assistência.

5 Conclusões

Neste artigo consideramos a classe de modelos de regressão beta para modelar a taxa de analfabetismo do Estado da Paraíba. A partir do modelo verificou-se que as variáveis índice de Gini, percentual da população com densidade maior que 2 e gasto com assistencialismo contribuem de forma positiva para o aumento da taxa, ou seja, a medida que essas variáveis aumentam, há um aumento na taxa de analfabetismo, quando as demais covariáveis são mantidas constantes. Por outro lado, a variável renda per capita, apresenta relação inversamente proporcional com a taxa de analfabetismo. Desse modo, um acréscimo no valor dessa variável corresponde a uma redução na taxa de analfabetismo.

Com base nos sinais das estimativas dos parâmetros, observou-se uma contribuição positiva da variável gasto com assistencialismo (*Gasto*), indicando que um aumento dessa variável representa um acréscimo no valor da taxa de analfabetismo.

Com intuito de investigar o resultado referente a relação entre o gasto com o assistencialismo e a taxa de analfabetismo, estimamos o impacto do gasto com o assistencialismo sobre a taxa de analfabetismo. Os resultados revelaram que para reverter o cenário da taxa de analfabetismo no Estado da Paraíba, ou seja, diminuir a taxa do analfabetismo considerando o gasto com o assistencialismo e as demais covariáveis constantes, é necessário que o gasto per capita com o assistencialismo seja de aproximadamente R$ 411,00 reais. Salientando que, essa redução é mais expressiva no grupo da população com renda per capita de até R$ 230,12.

O impacto estimado na taxa de analfabetismo levando em consideração o programa bolsa família, proporcionou a obtenção de resultados relevantes. Apesar do impacto ser positivo, em alguns momentos, deve-se analisar outros fatores que possam estar contribuindo para a evasão escolar, tais como, dificuldade de acesso a escola, falta de interesse por partes dos alunos, entre outros.

De modo geral, o ajuste obtido a partir de modelos de regressão beta, configurou-se como uma ferramenta poderosa para a estimação da taxa de analfabetismo no Estado da Paraíba. A classe de modelos utilizada permitiu tornar mais precisas as estimativas, uma vez que é possível captar de forma concisa as expectativas a priori das relações entre as variáveis explicativas e a variável resposta.

SILVA, C. R.; SOUZA, T. C. Modeling the rate of illiteracy in the state of Paraiba via Beta regression model. *Rev. Bras. Biom.,* São Paulo, v.32, n.3, p.345-359, 2014.

▪ *ABSTRACT: This article aims to evaluate the factors that influence the rate of illiteracy in the State of Paraíba in 2010. We used the regression model beta proposed by Ferrari & Cribari-Neto (2004) in order to model the rate of illiteracy in paraibanos municipalities. Additionally, we calculated the impact spending on assistance programs on illiteracy rate. We concluded that is necessary an investment of approximately R$ 411,00 to reduction in the rate of illiteracy.*

▪ *KEYWORDS: Assistance programs; beta regression model; illiteracy rate.*

Referências

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v.19, p.716–723, 1974.

BREHM, J.; GATES, S. Donut shops and speed traps: Evaluating models of supervision on police behavior. *American Journal of Political Science*, v.37, n.2, p.555–581, 1993.

COOK, R. D. Detection of influential observations in linear regression. *Technometrics*, v.19, p.15–18, 1977.

CRIBARI–NETO, F.; ZEILEIS, A. Beta Regression in R. *Journal Of Statistical Software*, v.34, p.1–24, 2010.

CRIBARI–NETO, F.; SOUZA, T. C. Testing inference in variable dispersion beta regressions. *Journal of Statistical Computation and Simulation*, v.82, p.1827–1843, 2012.

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI–NETO, F. On beta regression residuals. *Journal of Applied Statistics*, v.35, p.407–419, 2008a.

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI–NETO, F.Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, v.52, p.4417–4431, 2008b.

FERRARI, S. L. P.; CRIBARI–NETO, F. Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, v.31, p.799-815, 2004.

FERRARI, S. L. P.; Espinheira, P. L.; CRIBARI–NETO, F. Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, v.65, p.337–351, 2011.

FREIRE. P. *Educação e mudança*. 24 ed. Rio de Janeiro: Paz e Terra, 1979. 79p.

HANCOX,D.; HOSKIN, C. J.; WILSON, R. S. Evening up the score: Sexual selection favours both alternatives in the colour-polymorphic ornate rainbowfish. *Animal Behaviour*, v.80, n.5, p.845–851, 2010.

KIESCHNICK, R. ; McCULLOUGH, B. D. Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, v.3, p.193–213, 2003.

McCULLAGH, P.; NELDER, J. A. *Generalized Linear Models*, 2nd ed. London: Chapman and Hall. 1989. 511p.

ROCHA, A.; SIMAS, A. Influence diagnostics in a general class of beta regression models. *TEST*, v.20, p.95–119, 2011.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, v.54, p.348–366, 2010.

SMITHSON, M.; VERKUILEN, J. A better lemon-squeezer? Maximum likelihood regression with beta–distribuited dependent variables. *Psychological Methods*, v.11, p.54–71, 2006.

SOUZA, T. C.; CRIBARI–NETO, F. Uma estimativa do impacto eleitoral do Programa Bolsa-Família. *Revista Brasileira de Biometria*, v.31, p.79–103, 2013.

Recebido em 07.05.2014.

Aprovado após revisão em 18.08.2014.