Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Camila Ribeiro da Silva

Modelagem da Taxa de Analfabetismo no estado da Paraíba via Modelo de Regressão Beta.

João Pessoa, 28 de março de 2014

CAMILA RIBEIRO DA SILVA

MODELAGEM DA TAXA DE ANALFABETISMO DO ESTADO DA PARAÍBA VIA MODELO DE REGRESSÃO BETA.

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal da Paraíba, como requisito parcial para obtenção do Grau de Bacharel. Área de Concentração: Estatística Aplicada.

Orientadora: Profª. Drª. TATIENE CORREIA DE SOUZA.

João Pessoa, 28 de março de 2014

CAMILA RIBEIRO DA SILVA

MODELAGEM DA TAXA DE ANALFABETISMO DO ESTADO DA PARAÍBA VIA MODELO DE REGRESSÃO BETA

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal da Paraíba, como requisito parcial para obtenção do Grau de Bacharel. Área de Concentração: Estatística Aplicada.

Aprovada em 28 de março de 2014.

BANCA EXAMINADORA

Prof.ª Dra. TATIENE CORREIA DE SOUZA - Orientadora
UFPB

Prof.ª Dra. TARCIANA LIBERAL PEREIRA
UFPB

Prof. Dr. HEMÍLIO FERNANDES CAMPOS COÊLHO
UFPB

Este trabalho é carinhosamente dedicado

aos meus pais, Rizete e José.

Agradecimentos

Agradeço primeiramente à DEUS, companheiro inseparável, pelo dom da vida e por me abençoar em todos os momentos.

Aos meus pais, Rizete e José, pelo amor incondicional, carinho e dedicação, aos quais dedico todas as minhas conquistas.

Aos meus irmãos, Carol e Gabriel, e a minha prima Roberta, pelo carinho, incentivo e pelos momentos de descontração.

Ao meu namorado, Alisson, pelo carinho, amor, **paciência** inesgotável, dedicação e companheirismo, por ter sido apoio necessário para reerguer-me nos momentos difíceis e por estar comigo em todos os momentos.

A minha grande amiga Luzinete e ao seu esposo Cláudio, pelo estímulo, perseverança e confiança depositados em mim.

A dona Terezinha e ao senhor Azevedo, pelo carinho, apoio e confiança durante a minha trajetória.

A professora Tatiene, pela orientação constante, dedicação, confiança e pela convivência enriquecedora dos últimos meses. Obrigada pela oportunidade concedida, pela confiança em mim depositada e por todo aprendizado.

Ao Professor Hemílio, pela confiança e incentivo dedicado a mim em seus projetos. Por ter proporcionado a oportunidade de engrandecer meus conhecimentos ao longo desses anos.

Aos Professores, Tarciana e Hemílio, pelas contribuições para melhoria deste trabalho.

Aos meus amigos Pedro, Ramon, Marília e Geisislane, pelos momentos de alegria durante a graduação e por fazerem parte desta conquista. Amizades que vou levar para toda vida.
A todos os Professores do DE-UPPB, por contribuírem para minha formação acadêmica.

Aos demais colegas do Bacharelado em Estatística, Ianne, Jéssica, Camila, Marina, Aldine, Michele, Andreza, Alisson, Jodavid.

A todos os funcionários do Departamento de Estatística.

Ao CNPq, pelo apoio financeiro.

É muito melhor lançar-se em busca de conquistas grandiosas,
mesmo expondo-se ao fracasso,
do que alinhar-se com os pobres de espírito,
que nem gozam muito nem sofrem muito,
porque vivem numa penumbra cinzenta,
onde não conhecem nem vitória, nem derrota.
(Theodore Roosevelt)

Resumo

O analfabetismo se constitui em um dos mais 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. Se a educação sozinha não transforma a sociedade, sem ela tampouco a sociedade muda, defendeu FREIRE (1979). As taxas de analfabetismo no Brasil, normalmente tratadas dentro do universo de números e metas, deveriam, segundo especialistas em educação, ser também analisadas dentro da área de política social e econômica, já que a população considerada analfabeta é a mesma que sofre de outros problemas que afligem o país. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), atualmente, no Brasil são 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. No Estado da Paraíba, não saber ler nem escrever é a 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. Neste contexto, ajustamos um modelo de regressão beta FERRARI & CRIBARI-NETO (2004) com o intuito de explicar a taxa de analfabetismo no Estado da Paraíba. Diferentemente do modelo de regressão linear, o modelo de regressão beta possui aplicabilidade na modelagem de variáveis do tipo taxas ou proporções. O mesmo proporciona uma maior flexibilidade para a modelagem, fornecendo estimativas mais precisas, quando se trata de dados no intervalo (0,1).

Palavras-chave: Analfabetismo, educação, modelo de regressão beta.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 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 a educação sozinha não transforma a sociedade, sem ela tampouco a sociedade muda, defendeu FREIRE (1979). O fenômeno do analfabetismo está intrinsecamente ligado à desigualdade social, ambos reforçando-se mutuamente, embora não deixe de ser tomado como fruto da incompetência individual ou mal que assolou o país, como epidemia a ser erradicada, como frequentemente se ouve.

As taxas de analfabetismo no Brasil, normalmente tratadas dentro do universo de números e metas, deveriam, segundo especialistas em educação, serem também analisadas dentro da área de política social e econômica, já que a população considerada analfabeta é a mesma que sofre de outros problemas que afligem o país. 'Se você fizer o mapa do analfabetismo no Brasil, ele vai coincidir com o mapa da fome, com o do desemprego, e da alienação. Não raro esse analfabeto é o que fica doente, o que passa fome, o que vive de subemprego', afirma a pedagoga Silvia Colello, pesquisadora da Faculdade de Educação da Universidade de São Paulo (USP), na Revista eletrônica de jornalismo científico - (http://www.comciencia.br/comciencia).

Os últimos dados do Instituto Brasileiro de Geografia e Estatística (IBGE) sobre analfabetismo configuram um mapa de desigualdades que Alceu Ferraro, da Faculdade de Educação da Universidade Federal do Rio Grande do Sul (UFRGS), atribui à concen-

tração de terra, de renda e de oportunidades. Segundo Ferraro, que já foi membro do Comitê de Pesquisa do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), 'o país continua pagando o preço de dois fatores conjugados. Primeiro, do descaso secular do Estado, e, segundo, de um conjunto de fatores responsáveis pela enorme desigualdade social que tem, desde sempre, marcado a sociedade brasileira'.

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 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. Em uma reportagem a Revista eletrônica de jornalismo científico - (http://www.comciencia.br/comciencia), o diretor da Faculdade de Educação da Universidade Estadual de Campinas (Unicamp) e líder do Grupo de Pesquisa ALLE - Alfabetização, Leitura e Escrita, Sérgio da Silva Leite diz 'Este é um conceito muito discutível. Se utilizarmos um critério um pouco mais exigente, esses índices mudam e essa é uma das razões pelas quais o IBGE não muda esses conceitos, porque o que está em jogo é a própria imagem do país'.

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 Instituto Brasileiro de Geografia e Estatística (IBGE) que registrou em todo o país uma redução do número de analfabetos em 2009. A redução considerada pelo próprio IBGE como 'leve', produziu uma taxa de analfabetismo de 9,7% no ano passado, com o registro de pouco mais de 14 milhões de pessoas nestas condições em todo o país. O Nordeste concentra 52,7% do total de analfabetos do Brasil, segundo Instituto Brasileiro de Geografia e Estatística (IBGE). 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.

Acredita-se que o maior problema do abandono escolar começa no ensino médio, quando as pessoas começam a trabalhar. Ainda segundo a pesquisa (IPEA), do ponto de vista do ensino fundamental, na Paraíba e no Brasil como um todo há praticamente uma universalidade na cobertura de crianças estudando. Isto já é uma evolução muito grande, mas quando chega no ensino médio, a Paraíba ainda está abaixo da média nacional e regional de frequência escolar. A faixa etária do ensino médio vai geralmente dos 15 aos 18 anos, que é quando muitas pessoas acabam desistindo de estudar para começar a trabalhar.

A Unidade da Federação com a maior média de anos estudados é o Distrito Federal (9,6), seguida de São Paulo (8,5). Já os dois únicos estados que possuem média mais baixa que a Paraíba são Piauí (5,8) e Alagoas (5,7). Com relação ao índice de analfabetismo, os maiores percentuais foram registrados em Alagoas, onde 24,6% não sabem ler nem escrever, e no Piauí, que tem o índice de 23,4% de analfabetos entre as pessoas com 15 anos ou mais. No país, o índice de analfabetismo é de 9,7% e a média é de 7,5 anos de estudo.

Em função desta situação cada vez mais preocupante, 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, como descrito anteriormente. No âmbito federal, é possível citar alguns dos principais programas, entre eles, Educação de Jovens e Adultos (EJA), Brasil Alfabetizado, Programa de Alfabetização Solidária, Projovem Urbano, entre outros. No entanto, diante de tantas iniciativas, uma questão que surge a partir dessa discussão é por que o Brasil, com vários programas de erradicação do analfabetismo, não está conseguindo reduzir de forma significativa à taxa de analfabetismo nos últimos anos.

Um exemplo disso é o programa Brasil alfabetizado, voltado à alfabetização de jovens, adultos e idosos, que foi implantado desde 2003 buscando abolir o analfabetismo no Brasil, por meio de convênios com instituições alfabetizadoras, estaria tendo sua eficácia questionada após divulgação dos resultados da PNAD (Pesquisa Nacional de Amostra por Domicílio) (2005). Verificou-se que entre 2004 e 2005, a taxa de analfabetismo medida pela

PNAD divulgada anualmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE) caiu apenas 0,3%, passando de 11,4% para 11,1%. Assim como em 2004, as quedas mais significativas da taxa de analfabetismo ocorreram nas regiões Norte e Nordeste (1,1% e 0,6%, respectivamente). No sudeste a taxa de analfabetismo manteve-se a mesma, sendo de 6,6%. Neste sentido, a persistência da alta taxa de analfabetismo colocou em questionamento a eficácia do programa Brasil Alfabetizado em atingir as pessoas jovens e adultas consideradas analfabetas absoluto (aquelas pessoas que não conseguem escrever e ler um bilhete simples).

Além disso, o compromisso assumido por 164 países, entre eles o Brasil, de melhorar a qualidade da educação até 2015 não será atingido globalmente. A previsão está no 11° Relatório de Monitoramento Global de Educação para Todos divulgado pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO). O relatório registra, no entanto, que muitos países alcançaram avanços significativos. O compromisso Educação para Todos traz seis metas que integram o Acordo de Dacar (Senegal), assinado em 2000. Pelo acordo, até 2015, os países devem expandir cuidados na primeira infância e educação, universalizar o ensino primário, promover as competências de aprendizagem e de vida para jovens e adultos, reduzir o analfabetismo em 50%, alcançar a igualdade de gênero e melhorar a qualidade da educação.

As metas do Acordo de Dacar, no entanto, continuarão a ser perseguidas depois de 2015 e serão definidos critérios claros e mensuráveis e objetivos específicos de financiamento para a educação. É vital que se coloque em prática uma sólida estrutura educacional global pós-2015 para solucionar problemas pendentes e, ao mesmo tempo, lidar com novos desafios, segundo o relatório. O pior resultado, de acordo com o documento, foi na meta de reduzir o analfabetismo entre adultos. Apenas 29% dos países conseguiram cumprir o compromisso.

Foi observado que a paridade de gênero no primeiro nível do ensino secundário teve o melhor resultado, em que 70% dos países participantes alcançaram a meta. Ainda segundo o relatório, à medida que nos aproximamos de 2015 e determinamos uma nova agenda a ser seguida, todos os governos devem investir na educação como um acelerador do desenvolvimento inclusivo. Segundo a Declaração Mundial sobre Educação para Todos (1990), mais de 960 milhões de adultos são analfabetos, sendo que mais de 1/3 dos adultos do mundo não têm acesso ao conhecimento impresso e às novas tecnologias que poderiam

melhorar a qualidade de vida e ajudá-los a adaptar-se às mudanças sociais e culturais.

Além dos programas assistenciais voltados ao amparo educacional, como os citados anteriormente, 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).

Erradicar o analfabetismo é uma meta válida, mas que traz consigo outro problema ainda maior, o da exclusão social, ligado a aspectos como a democratização dos bens culturais, o acesso à cultura, justiça, moradia e trabalho. Reduzir os índices de analfabetismo para valores próximos de zero é um compromisso assumido pelo Brasil em diversas ocasiões e documentos. 'O fim do analfabetismo em números, no entanto, pode não significar, em termos reais, uma mudança efetiva. O Brasil pode até cumprir essas metas de alfabetização, mas esses números nunca vão representar a real situação da exclusão educacional e do analfabetismo no país. Sempre por trás dos números estão ocultas as atrocidades praticadas com a educação em relação aos seus aspectos qualitativos. O qualitativo é sacrificado em prol do quantitativo para se cumprir metas, para mostrar números aos organismos internacionais que fornecem recursos para a melhoria da educação em países subdesenvolvidos como o Brasil', afirma o professor Marcos Augusto de Castro Perez em uma reportagem concedida ao Jornal da Universidade de Santa Cruz (disponível em http://www.uesc.br/jornal/2012/jornal_171.pdf).

Capítulo 2

Objetivos

Esta monografia tem por objetivo modelar a taxa de analfabetismo do Estado da Paraíba utilizando o modelo de regressão beta. Mais especificamente,

- Identificar a partir do modelo obtido os principais fatores que influenciam (o aumento ou diminuição da taxa) a taxa de analfabetismo.
- Adicionalmente, avaliar a influência positiva e negativa das variáveis selecionadas para explicar a taxa de analfabetismo;
- Avaliar o impacto do gasto com assistencialismo, Bolsa Família, na taxa de analfabetismo do Estado da Paraíba;
- Estimar qual o valor per capita que deveria ser gasto com os programas sociais para mudar o cenário da taxa de analfabetismo no Estado da Paraíba.

Capítulo 3

Referencial teórico

3.1 Análise de regressão

Os modelos estatísticos constituem-se em ferramentas úteis para resumir e interpretar dados. Em muitas áreas do conhecimento, empregam-se modelos de regressão com o objetivo de investigar e modelar a relação entre uma variável aleatória de interesse, denominada de variável resposta (Y), e um conjunto de variáveis explicativas ($X_1, \ldots, X_k$), as quais se acredita serem responsáveis pela variabilidade de Y. O modelo de regressão normal linear é um dos mais utilizados em análises empíricas. No entanto, tal modelo torna-se inapropriado em situações em que a variável resposta é restrita ao intervalo (0,1), como ocorre com taxas e proporções, pois, é possível que sejam gerados valores ajustados fora do suporte da resposta. Adicionalmente, este tipo de informação apresenta, geralmente, heteroscedasticidade e assimetria, o que pode resultar em conclusões incorretas, quando estas são resultado de inferências baseadas nas suposições do modelo linear clássico. Uma alternativa de solução, geralmente empregada, é transformar a variável resposta de tal forma que esta assuma valores em toda reta e em seguida, modelar a resposta transformada. No entanto, esse enfoque apresenta algumas desvantagens, como, por exemplo, o fato de que os parâmetros do modelo não podem ser facilmente interpretados em termos da resposta original, (ver PAOLINO (2001)), dependendo da transformação. Este modelo assume uma relação linear entre a variável resposta e as variáveis explicativas, ou seja,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i, i = 1, \ldots, n \tag{3.1}$$

em que, Y_i representa a variável resposta, $x_{1i}, \ldots, x_{ki}$ representam k variáveis explicativas

e ϵ o erro aleatório do modelo, que é assumido ter distribuição $\mathcal{N}(0, \sigma^2)$. De acordo com PAULA (2004), durante muitos anos os modelos normais lineares foram utilizados para descrever a maioria dos fenômenos aleatórios. Mesmo quando o fenômeno em estudo não apresentava uma resposta para qual fosse razoável a suposição de normalidade, algum tipo de transformação era considerada para alcançar tal suposição. Certamente, a família de transformações mais conhecida foi proposta por BOX & COX (1964), que transforma o valor observado y positivo em

$$z = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & , \quad \text{se } \lambda \neq 0 \\ \log(y) & , \quad \text{se } \lambda = 0, \end{cases}$$

sendo λ uma constante desconhecida. O objetivo da transformação de BOX & COX, quando aplicada a um determinado conjunto de valores observados, é tornar a variável transformada aproximadamente normal, linear e com variância constante em relação a um conjunto de variáveis explicativas $x_1, \ldots, x_k$. No entanto, raramente isso ocorre para um único valor de λ (BOX & DRAPER, 1987). O avanço da tecnologia, acompanhado com a produção acelerada de informações, impulsionou a necessidade de ajustes de modelos que retratem de forma mais precisa a realidade. As suposições como normalidade e variância constante para o erro aleatório, e a adoção de uma forma funcional linear entre regressando e regressores, tornam o modelo normal linear inadequado para modelagem de diversos fenômenos práticos, gerando assim resultados irrealistas. Diante desta necessidade, é essencial o desenvolvimento de modelos de regressão mais flexíveis e menos restritivos, permitindo, portanto, a construção de modelos mais próximos da realidade. Neste contexto, diversas técnicas de modelagem de regressão univariadas foram desenvolvidas, objetivando superar as principais limitações dos modelos lineares. Dentre essas técnicas, os Modelos Lineares Generalizados (Generalized Linear Models - GLM) merecem destaque. Neste modelo, é assumida uma distribuição de probabilidade pertencente à família exponencial para a variável resposta y, em que a média μ de y é modelada como uma função das variáveis explicativas e a variância de y, dada por $V(y) = \phi\nu(\mu)$, depende do parâmetro de dispersão ϕ, suposto constante, e da média μ, através da função de variância $\nu(\mu)$. Além disso, para distribuições pertencentes à família exponencial a assimetria e curtose de y são, em geral, funções de μ e ϕ. Desse modo, nos modelos lineares generalizados, a variância, a assimetria e a curtose não são modeladas explicitamente em função das variáveis explicativas, mas implicitamente, por estarem em função de μ. Além

dos modelos descritos anteriormente, ainda existem os Modelos Aditivos Generalizados (Generalized Additive Models - GAM), que ocupam um lugar de destaque na literatura de NELDER & WEDDERBURN (1972) e HASTIE & TIBSHIRANI (1990). Além disso, ainda temos os Modelos Lineares Mistos (efeitos aleatórios), Modelos Lineares Generalizados Mistos (Generalized Linear Mixed Models - GLMM), entre outros, que surgiram como alternativas para modelagem de dados. No entanto, esses modelos não apresentam uma estrutura que permita modelar uma variável que está restrita no intervalo unitário, ou seja, $0 < y < 1$.

Devido às limitações existentes nos modelos abordados anteriormente para modelagem de variáveis no intervalo (0,1), sejam relacionadas às suposições restritivas ou a outros aspectos, FERRARI & CRIBARI-NETO (2004) propuseram um modelo para o ajuste de variáveis que assumem continuamente valores no intervalo unitário padrão, denominado modelo de regressão beta, que será tratado com mais detalhes nas seções subsequentes.

3.1.1 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. Um modelo de regressão beta foi proposto por FERRARI & CRIBARI-NETO (2004) como uma forma de suprir algumas das limitações associadas aos modelos tradicionais (Modelos de Regressão Linear), principalmente, no que se refere a estrutura da variável reposta. Os autores fazem uso de uma parametrização alternativa que permite a modelagem da média da resposta envolvendo um parâmetro de dispersão. Desta forma, através de uma função de ligação, a média fica relacionada a um preditor linear, de forma semelhante ao que se observa nos modelos lineares generalizados (ver MCCULLAGH & NELDER, 1989). O parâmetro de dispersão indexado no modelo, em certas situações, pode variar ao longo das observações (SMITHSON & VERKUILEN, 2006; ESPINHEIRA et al. 2008a, 2008b). Podendo existir, portanto, uma extensão do modelo que considera um submodelo para a média e um

submodelo para a precisão, dando mais flexibilidade ao ajuste dos dados (SIMAS et al., 2010). Dessa forma, o modelo de regressão beta torna-se mais adequado para modelagem da variável resposta, quando esta se trata de uma proporção ou taxa, em que se verifica simultaneamente a presença de heteroscedasticidade e assimetria.

3.1.2 Definição

A distribuição beta é muito flexível para a modelagem de taxas e proporções, e sua densidade pode ter formas muito diferentes, dependendo dos dois parâmetros que indexam a distribuição. A função de densidade da distribuição beta é dada por

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, 0 < y < 1$$

sendo $p > 0$, $q > 0$ e $\Gamma(p)$ é a função gama avaliada no ponto p, ou seja

$$\Gamma(p) = \int_0^\infty y^{p-1}\mathrm{e}^{-y}dy.$$

A média e a variância de y são respectivamente,

$$\mathrm{E}(y) = \frac{p}{p+q} \quad \text{e} \quad \mathrm{var}(y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

No entanto, é comum em análise de regressão modelar a média da variável resposta e também definir o modelo de tal modo que ele tenha um parâmetro de dispersão (ou precisão). Para obter uma estrutura de regressão para a média da variável resposta com um parâmetro de dispersão, FERRARI & CRIBARI-NETO (2004) utilizaram uma reparametrização da densidade beta. Fazendo $\mu = \dfrac{p}{p+q}$ e $\phi = p + q$, ou seja, $p = \mu\phi$ e $q = (1-\mu)\phi$, temos

$$\mathrm{E}(y) = \mu \quad \text{e} \quad \mathrm{var}(y) = \frac{V(\mu)}{1+\phi},$$

em que $V(\mu) = \mu(1-\mu)$, μ é a média da variável resposta e ϕ pode ser interpretado como o parâmetro de precisão, no sentindo que para μ fixo, quanto maior o valor de ϕ, menor a variância de y.

Utilizando essa reparametrização, a densidade da distribuição beta pode ser escrita como

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \tag{3.2}$$

em que $0 < \mu < 1$ e $\phi > 0$. Como dito anteriormente, a distribuição beta é bastante flexível para modelar proporções, dependendo dos dois valores dos parâmetros que a indexam, a densidade assume formas bem variadas, acomodando distribuições simétricas, assimétricas, formas de J e de U. A partir da Figura 3.1 é possível observar algumas formas da densidade da distribuição beta para diferentes valores de (μ, ϕ).

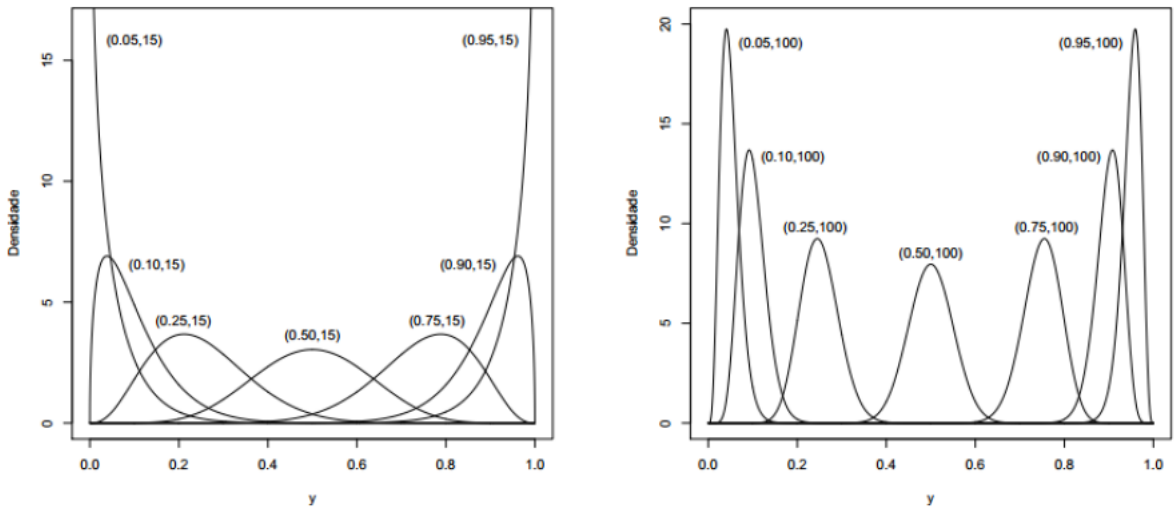

Figura 3.1: Densidades da distribuição beta para diferentes valores de (μ, ϕ).

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, em que cada y_i, $i = 1, \ldots, n$, segue a densidade da Equação (3.2) com média μ_i e parâmetro de precisão desconhecido ϕ, o modelo de regressão beta com dispersão fixa assume que a média satisfaz a seguinte relação funcional

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = \eta_i, \tag{3.3}$$

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.

3.1.3 Funções de ligação

Existem algumas possíveis escolhas para a função de ligação $g(\cdot)$. 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, ou ainda a função complemento log-log

$$g(\mu) = -\log(-\log(\mu)),$$

têm-se ainda, as funções de ligação cloglog ($\log\{-\log(1-\mu_i)\}$) e Cauchy ($tan(\pi(\mu-0.5))$). Para maiores detalhes sobre estas funções de ligação, ver MCCULLAGH & NELDER (1989).

3.1.4 Estimação

PAOLINO (2001) mostrou que uma aproximação normal padrão, para os casos em que a resposta está restrita ao intervalo unitário (0,1), e a estimação dos parâmetros via mínimos quadrados, podem, geralmente, conduzir a uma avaliação dos efeitos das covariáveis bastante imprecisa. E, por esta razão, PAOLINO (2001) propôs a distribuição Beta como melhor alternativa para modelar situações em que se envolvem proporções, além de propor o método de máxima verossimilhança, para estimação dos parâmetros do modelo.

Desse modo, a estimação dos parâmetros do modelo de regressão beta é feita através do método de máxima verossimilhança. O log da densidade apresentada na Equação (3.2) é dada por

$$\begin{aligned}\log f(y;\mu,\phi) &= \log\Gamma(\phi) - \log\Gamma(\mu\phi) - \log\Gamma((1-\mu)\phi) \\ &+ (\mu\phi)\log(y) + [(1-\mu)\phi - 1]\log(1-y_i)\end{aligned}$$

e a função de log-verossimilhança é da forma

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

em que,

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

Fazendo, $y_i^* = \log\left\{\frac{y_i}{(1-y_i)}\right\}$ e $\mu_i^* = \psi(\mu_i\phi) - \psi((1-\mu_i)\phi)$, em que $\psi(\cdot)$ é a função digama, i.e., $\psi(z) = \partial \log\Gamma(z)/\partial z$ para $z > 0$, a função escore obtida através da diferenciação da função log-verossimilhança com relação aos parâmetros é dada por $(U_\beta(\beta,\phi)^\top$, $U_\phi(\beta,\phi)^\top)$, em que

$$U_\beta(\beta, \phi) = \phi X^\top T(y_i^* - \mu_i^*),$$

com X sendo uma matriz $n \times k$ cuja t-ésima linha é $x_i^\top$, $i = 1, \ldots, n$, $T = \text{diag}[1/g'(\mu_i)]$ e

$$U_\phi(\beta, \phi) = \sum_{i=1}^{n} [\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi((1-\mu_i)\phi) + \psi(\phi)].$$

Seja $\zeta = (\beta^\top, \phi)^\top$ os estimadores de máxima verossimilhança de β e ϕ que são obtidos como solução do sistema de equações não-lineares $U(\zeta) = 0$. Como estes estimadores não possuem forma fechada, eles precisam ser obtidos numericamente maximizando a função log-verossimilhança através de algum algoritmo de otimização não-linear. Entre os principais algoritmos pode-se citar o de Newton-Raphson, Escore de Fisher, BFGS, entre outros.

Agora seja $c = (c_1, \ldots, c_n)^\top$, com $c_i = \phi[\psi'(\mu_i\phi)\mu_i - \psi'((1-\mu_i)\phi)(1-\mu_i)]$, em que $\psi'(\cdot)$ é a função trigama, $D = \text{diag}\{d_1, \ldots, d_n\}$, com $d_i = \psi'(\mu_i\phi)\mu_i^2 + \psi'((1-\mu_i)\phi)(1-\mu_i)^2 - \psi'(\phi)$ e $W = \text{diag}(w_1, \ldots, w_n)$ com

$$w_i = \phi\{\psi'(\mu_i\phi) + \psi'((1-\mu_i)\phi)\}\frac{1}{\{g'(\mu_i)\}^2}.$$

A matriz de informação de Fisher é dada por

$$K = K(\beta, \phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix},$$

em que $K_{\beta\beta} = \phi X^\top W X$, $K_{\beta\phi} = K_{\phi\beta} = X^\top T c$, e $K_{\phi\phi} = \mathrm{tr}(D)$.

FERRARI & CRIBARI-NETO (2004) ressaltam que, diferentemente do que acontece nos modelos lineares generalizados (MCCULLAGH & NELDER, 1989), no modelo de regressão beta, os parâmetros β e ϕ não são ortogonais. Os autores argumentam ainda que sob as condições de regularidade, quando o tamanho da amostra é grande,

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

aproximadamente, em que $\hat{\beta}$ e $\hat{\phi}$ são os estimadores de máxima verossimilhança de β e ϕ, respectivamente. Adicionalmente, usando expressões padrões para inversas de matrizes particionadas (RAO,1973, p.33), FERRARI & CRIBARI-NETO obtêm a inversa da matriz de informação de Fisher dada por

$$K^{-1} = K^{-1}(\beta, \phi) = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix},$$

em que,

$$K^{\beta\beta} = \frac{1}{\phi}(X^\top W X)^{-1}\left[I_k + \frac{X^\top T c c^\top T^\top X (X^\top W X)^{-1}}{\gamma\phi}\right],$$

com $\gamma = \mathrm{tr}(D) - \phi^{-1} c^\top T^\top X (X^\top W X)^{-1} X^\top T c$,

$$K^{\beta\phi} = (K^{\phi\beta})^\top = -\frac{1}{\gamma\phi}(X^\top W X)^{-1} X^\top T c,$$

e $K^{\phi\phi} = \gamma^{-1}$. Aqui I_k é a matriz identidade de ordem k.

A seguir será apresentado brevemente uma descrição dos principais algoritmos e a forma que eles desenvolvem o processo de estimação.

Processo iterativo de Newton-Rhapson

Seja $\theta = (\beta, \phi)^\top$ o vetor de parâmetros. $U(\theta) = (U_\beta(\beta, \phi)^\top; U_\phi(\beta, \phi))^\top$, o vetor de funções escore de dimensão $(k+1) \times 1$. Para a obtenção do estimador de máxima verossimilhança do vetor θ expandimos a função escore $U(\theta)$ em torno de um valor inicial $\theta^{(0)}$, de modo que

$$U(\theta) \simeq U\left(\theta^{(0)}\right) + U'\left(\theta^{(0)}\right)\left(\theta - \theta^{(0)}\right),$$

em que $U^{'}(\theta)$ denota a derivada de 1ª ordem de $U(\theta)$ com respeito a $\theta^{\top}$ Fazendo $U(\theta) = 0$ e repetindo o processo acima, chegamos ao processo iterativo

$$\theta^{(m+1)} = \theta^{(m)} + \{-U^{'}(\theta^{(m)})\}^{-1} U(\theta^{(m)}), \quad m = 0, 1, \ldots \tag{3.4}$$

O aspecto mais trabalhoso desse processo iterativo é a inversão da matriz $U^{'}(\theta)$.

Processo iterativo de Escore de Fisher

A expressão (3.4) apresenta uma forma alternativa equivalente assintoticamente, uma vez que, pela lei dos grandes números $U^{'}(\theta)$, converge para a matriz K quando $n \longrightarrow \infty$. Assim, substituindo a informação observada em (3.4) pela esperada, obtemos a seguinte aproximação

$$\theta^{(m+1)} = \theta^{(m)} + \{-k^{(m)}\}^{-1} U\left(\theta^{(m)}\right), \quad m = 0, 1, \ldots .$$

Esse procedimento iterativo é denominado Escore de Fisher. Da mesma forma que o método de Newton-Rhapson, o aspecto mais trabalhoso é a inversão da matriz K. Além dos dois métodos citados, têm-se ainda o método BFGS, que utiliza o mesmo princípio do método de Newton-Rhapson, diferenciando-se pelo fato de utilizar uma sequência de matrizes simétricas e positivas definidas $B^{(m)}$ no lugar da matriz $U'\left(\theta^{(m)}\right)^{-1}$.

Os três processos iterativos são sensíveis à estimativa inicial $\theta^{(0)}$. FERRARI & CRIBARI-NETO(2004) sugerem utilizar como uma estimativa inicial de β a estimativa de mínimos quadrados ordinários desse vetor de parâmetros, obtida de uma regressão linear da resposta transformada em $g(y_1), \ldots, g(y_n)$ em X, isto é,$\left(X^{\top}X\right)^{-1} X^{\top}z$, em que $z = (g(y_1), \ldots, g(y_n))^{\top}$. Em relação ao parâmetro de precisão, os mesmos autores sugerem um valor inicial para ϕ baseado no fato de que $Var(Y_k) = \dfrac{\mu_k(1-\mu_k)}{1+\phi}$. De forma alternativa podemos escrever

$$\phi = \left[\mu_k \frac{(1-\mu_k)}{Var(Y_k)}\right] - 1.$$

3.2 Testes de Hipóteses

Em geral, após a estimação dos parâmetros populacionais, realizam-se testes afim de determinar se as hipóteses feitas sobre estes parâmetros são suportadas por evidências obtidas a partir de dados amostrais. Neste contexto, testes baseados na função de verossimilhança são amplamente empregados devido a suas propriedades de otimalidade. Os procedimentos mais frequentemente utilizados são os testes da razão de verossimilhança, escore e Wald, que são assintoticamente equivalentes sob a hipótese nula.

Estes testes, porém, são realizados com base em valores críticos obtidos a partir de aproximações assintóticas, o que, em geral, causa considerável distorção de tamanho em amostras finitas. Tal distorção pode ser reduzida através de refinamentos assintóticos, produzidos por ajustes realizados sobre a função de verossimilhança ou, alternativamente, por emprego de testes baseados em *bootstrap*.

No caso dos modelos de regressão beta, as estatísticas dos testes da razão de verossimilhança, escore e Wald foram apresentadas por FERRARI & CRIBARI-NETO (2004). Um teste de erro de especificação, baseado no teste RESET (RAMSEY,1969), para o modelo com dispersão constante foi desenvolvido por CRIBARI-NETO & LIMA (2007). FERRARI & PINHEIRO (2011) visando realizar inferências mais precisas em amostras finitas, derivaram o ajuste de SKOVGAARD (2001) para esta classe de modelos. As autoras concluíram que as estatísticas ajustadas propostas, sob a hipótese nula, têm distribuição mais próxima da distribuição qui-quadrado de referência que a estatística original.

Dessa forma, os métodos de inferência para o modelo de regressão beta baseiam-se, fundamentalmente, no método da máxima verossimilhança. Em que, são consideradas usualmente três estatísticas para testar hipóteses relativas aos parâmetros $\beta's$. São elas a razão de verossimilhanças, escore e Wald.

3.2.1 Teste da Razão de Verossimilhança

É possível realizar testes assintóticos para fazer inferências sobre os parâmetros desconhecidos. Considere, por exemplo, o teste da hipótese nula $\mathcal{H}_0 : \beta = \beta^{(0)}$ *versus* a hipótese alternativa $\mathcal{H}_1 : \beta \neq \beta^{(0)}$, em que $\beta = (\beta_1, \ldots, \beta_m)^\top$ e $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_m^{(0)})^\top$ para $m < k$ e $\beta^{(0)}$ dado. A estatística da razão de verossimilhanças é

$$\varpi_1 = 2\left\{l\left(\hat{\beta},\hat{\phi}\right) - l\left(\tilde{\beta},\tilde{\phi}\right)\right\},$$

sendo, $l\left(\hat{\beta},\hat{\phi}\right)$ o logaritmo natural da função de verossimilhança $\left(\tilde{\beta},\tilde{\phi}\right)^{\top}$ o estimador de máxima verossimilhança restrito de (β,ϕ) obtido pela imposição da hipótese nula. Sob condições gerais de regularidade e sob $\mathcal{H}_0$, $\varpi_1 \xrightarrow{D} \chi^2_m$. Dessa forma o teste pode ser avaliado usando valores críticos aproximados da distribuição χ^2_m, onde m é a dimensão do espaço paramétrico sob a hipótese $\mathcal{H}_0$.

3.2.2 Teste Escore

Para descrever o teste escore de $\mathcal{H}_0 : \beta = \beta^{(0)}$, considere o vetor $U_{1\beta}$ um vetor coluna m dimensional contendo os primeiros m elementos da função escore de β e $K_{11}^{\beta\beta}$ a matriz $m \times m$ formada das m primeiras linhas e das m primeiras colunas da matriz K^{-1}. Pode-se mostrar que $U_{1\beta} = \phi X_1^k k(y^* - \mu^*)$, em que X é particionada como $[X_1, X_2]$ seguindo a partição de β. A estatística escore de RAO pode ser escrita com

$$\varpi_2 = \tilde{U}_{1\beta}^T K_{11}^{\beta\beta} U_{1\beta},$$

em que o til indica que as quantidades estão sendo avaliadas no estimador de máxima verossimilhança restrito. Sob condições gerais de regularidade e sob $\mathcal{H}_0$, $\varpi_2 \xrightarrow{D} \chi^2_m$, em que m é a dimensão do espaço paramétrico sob a hipótese $\mathcal{H}_0$.

3.2.3 Teste de Wald

Da mesma forma, podemos utilizar o teste de Wald para realizar inferências assintóticas acerca do vetor de parâmetros β. A estatística do teste $\mathcal{H}_0 : \beta = \beta^{(0)}$ versus $\mathcal{H}_1 : \beta \neq \beta^{(0)}$ é dada por

$$\varpi_3 = \left(\hat{\beta} - \beta^{(0)}\right)^T \left(\hat{K}_{11}^{\beta\beta}\right)^{-1} \left(\hat{\beta} - \beta^{(0)}\right),$$

em que $\hat{K}_{11}^{\beta\beta}$ é igual a $K_{11}^{\beta\beta}$ avaliado no estimador de máxima verossimilhança sem restrição, e $\hat{\beta}_k$ é o estimador de máxima verossimilhança de β_k. Sob condições gerais de regularidade

e sob $\mathcal{H}_0$, $\varpi_3 \xrightarrow{D} \chi^2_m$, em que m é a dimensão do espaço paramétrico sob a hipótese $\mathcal{H}_0$. Em particular, para testar a significância do k-ésimo parâmetro regressor β_k, $k = 1, 2, \ldots, n$, podemos utilizar a raiz quadrada sinalizada da estatística de Wald, isto é,

$$W = \frac{\hat{\beta}}{sd\left(\hat{\beta}\right)},$$

em que, $sd\left(\hat{\beta}\right)$ é o erro padrão assintótico do estimador de máxima verossimilhança de β, obtido da inversa da matriz de informação de Fisher avaliada nos estimadores de máxima verossimilhança.

3.2.4 Teste de especificação

Depois de especificada e estimados os coeficientes da regressão, a próxima etapa consiste na realização de testes que permitam avaliar a especificação do modelo. Pois, ao postularmos uma estrutura paramétrica de regressão não é possível saber de fato se o modelo considerado retrata adequadamente a realidade do fenômeno em estudo. Caso a especificação utilizada seja errônea, inferências imprecisas podem ocorrer no que se refere à estimação dos parâmetros, intervalos de confiança e testes de hipóteses. Testes de especificação em modelos de regressão formam uma importante área de pesquisa. RAMSEY (1969) introduziu o teste de erro de especificação (RESET) em análise de regressão linear para detectar forma funcional inapropriada e variáveis omitidas. O teste foi desenvolvido comparando-se a distribuição dos resíduos sob a hipótese de que a especificação do modelo é correta com a distribuição dos resíduos produzidos sob a hipótese alternativa de que existe um erro de especificação. Sob a hipótese nula de ausência de erro de especificação existirá um estimador eficiente, consistente e assintoticamente normal. Contudo, sob a hipótese alternativa de má especificação, esse estimador será viesado e inconsistente (HAUSMAN,1978).

RAMSEY & SCHMIDT (1976) mostraram que o teste RESET baseado no uso dos resíduos de mínimos quadrados é equivalente ao teste originalmente proposto por RAMSEY (1969). Dessa forma, o procedimento do teste se reduz a incluir uma forma não-linear ao modelo, através de potências de variáveis adicionais, chamadas de variáveis de teste, e por meio de valores críticos da distribuição F, testar a exclusão de tais variáveis. A intuição por trás do teste é que se essas variáveis de teste têm algum poder em explicar

a variável dependente, então o modelo está mal especificado. Alguns autores estudaram as propriedades do teste RESET. RAMSEY & GILBERT (1972) sugeriram usar como variáveis de teste a segunda, terceira e quarta potências do valor ajustado. THURSBY & SCHMIT (1977) recomendaram usar a segunda, terceira e quarta potências das variáveis independentes. SHUKUR & EDGERTON (2002) generalizaram o teste RESET para cobrir sistemas de equações. CRIBARI-NETO & LIMA (2007) propuseram um teste de erro de especificação para modelos de regressão beta. O teste é útil para identificar não-linearidade negligenciada e função de ligação incorretamente especificada.

3.3 Modelo de regressão beta com dispersão variável

O modelo de regressão beta proposto por FERRAI & CRIBARI-NETO (2004) assume que o parâmetro de precisão é uma constante na função de variância. Isto implica dizer que a dispersão (inverso da precisão) é constante para todas as observações. No entanto, as perdas de eficiência em usar modelos com dispersão constante, quando na verdade a dispersão é variável, podem ser substancial. De fato, a estimação eficiente dos parâmetros em uma regressão depende da modelagem correta da dispersão. Muitos autores têm considerado a modelagem da dispersão para dados normais. No contexto de modelos lineares generalizados, SMYTH & VERBYLA (1999) definem os chamados modelos lineares duplos que permitem que a média e a variância sejam modeladas simultaneamente. Neste contexto, será apresentado um modelo de regressão beta em que o parâmetro de dispersão varia com as observações, havendo assim uma estrutura heteroscedástica. Note-se que, mesmo que ϕ seja constante ao longo das observações as variâncias de $y_1, \ldots, y_n$ não serão constantes, pois dependerão das médias desconhecidas, estas variando de acordo com a estrutura de regressão. Assim, o conceito de heteroscedasticidade no presente contexto difere daquele empregado em modelos lineares normais de regressão, em que, sob homoscedasticidade as variâncias condicionais são constantes. De fato, segundo HOUAISS (2001), homoscedasticidade é a propriedade de apresentar a mesma variância ou dispersão. Em modelos normais, a medida de dispersão tipicamente utilizada é a variância, logo as duas medidas não se confundem. No entanto, nos modelos da família exponencial, homoscedasticidade significa que o parâmetro de dispersão é o mesmo para todas as observações.

Ressaltamos que alguns autores já apresentaram modelos de regressão baseados na distribuição beta em que a média e a dispersão são modeladas simultaneamente. A estimação dos parâmetros dos modelos da média e da dispersão também é feita através do método de máxima verossimilhança. SMITHSON & VERKUILEN (2006) apresentam, adicionalmente, testes de hipóteses tanto para o modelo da média quanto para o modelo da precisão. Desse modo, a seguir será apresentado uma extensão do modelo proposto por FERRARI & CRIBARI NETO (2004), para situações em que o parâmetro de precisão ϕ não é constante para todas as observações. Sendo, a precisão modelada em termos de covariáveis e de parâmetros desconhecidos da mesma forma que a média condicional. Como também, expressões para a função escore, matriz de informação de Fisher e para sua inversa. Adicionalmente, serão apresentados testes para hipótese nula de homoscedasticidade.

3.3.1 Definição e estimação do modelo

A partir de agora será assumido que $y_1, \ldots, y_n$ são variáveis aleatórias independentes, segue a densidade em (3.2) com média μ_i definida em (3.3) e precisão ϕ_i, admitimos ainda que

$$h(\phi_i) = \sum_{j=1}^{q} z_{ij}\gamma_j = \vartheta_i, \tag{3.5}$$

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. O logaritmo da função de verossimilhança é

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

com

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

Logo, para $t = 1, \ldots, k$ a função escore para β_t é dada por

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta_t} = \sum_{i=1}^{n} \frac{\partial \ell_i(\mu_i\phi_i)}{\partial \mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_t},$$

com $d\mu_i/d\eta_i = 1/g'(\mu_i)$ e

$$\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i} = \phi_i\left[\log\frac{y_i}{1-y_i} - \{\psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i)\}\right].$$

Definimos

$$\mu_i^* = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i)$$

e, consequentemente,

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta_t} = \sum_{i=1}^{n} \phi_i(y_i^* - \mu_i^*)\frac{1}{g'(\mu_i)}x_{it},$$

em que y_i^* é o logito de y_i, como definido anteriormente. Podemos escrever o vetor $U_\beta(\beta,\gamma)$, função escore de $\beta = (\beta_1,\ldots,\beta_k)^\top$, através da seguinte expressão matricial

$$U_\beta(\beta,\gamma) = X^\top \Phi T(y^* - \mu^*),$$

em que, $\Phi = \text{diag}\{\phi_1,\ldots,\phi_n\}$ e $\mu^* = (\mu_1^* - \mu_n^*)^\top$.

No contexto em que ϕ varia ao longo das observações, a densidade beta apresentada em (3.2) escrita na forma da família exponencial biparamétrica canônica é dada por

$$f(y_i,\mu_i,\phi_i) = \exp\tau_1 T_1 + \tau_2 T_2 - \mathcal{A}(\tau)(1/y_i(1-y_i)),$$

em que, $\tau = (\tau_1,\tau_2) = (\mu_i\phi_i,\phi_i)$, $(T_1T_2) = (\log y_i/(1-y_i), \log(1-y_i))$ e

$$\mathcal{A}(\tau) = -\log\Gamma(\phi_i) + \log\Gamma(\mu_i\phi_i) + \log\Gamma((1-\mu_i)\phi_i).$$

Assim, segue que

$$E(T_1) = E(y_i^*) = \partial(\tau)/\partial\tau_1 = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i) = \mu_i^* \tag{3.6}$$

e

$$E(T_2) = E(\log(1-y_i)) = \partial(\tau)/\partial\tau_2 = -\psi(\phi_i) + \psi((1-\mu_i)\phi_i). \tag{3.7}$$

Note que (3.6) é equivalente a $E(\partial\ell_i(\mu_i,\phi_i)/\partial\mu_i)=0$, ou ainda, $E(U_\beta(\beta,\gamma))=0$. Consideremos agora as derivadas do logaritmo da função de verossimilhança em relação aos parâmetros que modelam a precisão, γ_j, $j=1,\ldots,q$. Temos que

$$\frac{\partial\ell(\beta,\gamma)}{\partial\gamma_j}=\sum_{i=1}^{n}\frac{\partial\ell_i(\mu_i,\phi_i)}{\partial\phi_i}\frac{d\phi_i}{d\vartheta_i}\frac{\partial\vartheta_i}{\partial\gamma_j},$$

sendo que, $\partial\phi_i/\partial\vartheta_i=1/h'(\phi_i)$. Também,

$$\begin{aligned}\frac{\partial\ell_i(\mu_i,\phi_i)}{\partial\phi_i} &= \mu_i\left[\log\frac{y_i}{1-y_i}-(\psi(\mu_i\phi_i)-\psi(1-\mu_i)\phi_i))\right]\\ &+ \log(1-y_i)-\psi((1-\mu_i)\phi_i)+\psi(\phi_i).\end{aligned}$$

Utilizando as definições de y_i^* e de μ_i^* dadas anteriormente, chegamos a

$$\frac{\partial\ell_i(\mu_i,\phi_i)}{\partial\phi_i}=\mu_i(y_i^*-\mu_i^*)+\log(1-y_i)-\psi((1-\mu_i\phi_i))+\psi(\phi_i).$$

Assim, a função escore para cada um dos parâmetros γ_j é dada por

$$\frac{\partial\ell(\beta,\gamma)}{\partial\gamma_j}=\sum_{i=1}^{n}\left[\mu_i(y_i^*-\mu_i^*)+\log(1-y_i)-\psi((1-\mu_i)\phi_i)+\psi(\phi_i)\right]\frac{1}{h'(\phi_i)}z_{ij},$$

que pode ser expressa por

$$\frac{\partial\ell(\beta,\gamma)}{\partial\gamma_j}=\sum_{i=1}^{n}a_i\frac{1}{h'(\phi_i)}z_{ij},$$

em que $a_i=\mu_i(y_i^*-\mu_i^*)+\log(1-y_i)-\psi((1-\mu_i)\phi_i)+\psi(\phi_i)$. A função escore para $\gamma=(\gamma_1,\ldots,\gamma_q)^\top$ pode ser expressa em forma matricial como

$$U_\gamma(\beta,\gamma)=Z^\top Ha,$$

em que Z é uma matriz $n\times q$ cuja t-ésima linha é $z_i^\top$, $H=\mathrm{diag}\{1/h'(\phi_1),\ldots,1/h'(\phi_n)\}$, e $a=(a_1,\ldots,a_n)^\top$. Note que (3.7) é equivalente a $E(a)=0$, ou ainda, $E(U_\gamma(\beta,\gamma))=0$.

Consideremos $W=\mathrm{diag}\{w_1,\ldots,w_n\}$, $C=\mathrm{diag}\{c_1,\ldots,c_n\}$,com

$$c_i=\phi_i\left[\psi'(\mu_i\phi_i)\mu_i-\psi'((1-\mu_i)\phi_i)(1-\mu_i)\right],$$

e $D^* = \text{diag}(d_1^*, \ldots, d_n^*)$, com

$$d_i^* = \left[\psi'(\mu_i\phi_i)\mu_i^2 + \psi'((1-\mu_i)\phi_i)(1-\mu_i^2) - \psi'(\phi_i)\right] \frac{1}{\{h'(\phi_i)\}^2}.$$

A matriz informação de Fisher é dada por,

$$K^* = K^*(\beta, \gamma) = \begin{pmatrix} K^*_{\beta\beta} & K^*_{\beta\gamma} \\ K^*_{\gamma\beta} & K^*_{\gamma\gamma} \end{pmatrix}$$

em que $K^*_{\beta\beta} = X^\top \Phi W X$, $K^*_{\beta\gamma} = {K^*_{\gamma\beta}}^\top = X^\top C T H Z$ e $K^*_{\gamma\gamma} = Z^\top D^* Z$.

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

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

em que

$$K^{*-1} = K^{*-1}(\beta, \gamma) = \begin{pmatrix} K_*^{\beta\beta} & K_*^{\beta\gamma} \\ K_*^{\gamma\beta} & K_*^{\gamma\gamma} \end{pmatrix},$$

com

$$K_*^{\beta\beta} = \left(X^\top \Phi W X - X^\top C T H Z (Z^\top D^* Z)^{-1} Z^\top H T C^\top X\right)^{-1},$$

$$K_*^{\beta\gamma} = \left(K_*^{\gamma\beta}\right)^\top = -K_*^{\beta\beta} X^\top C T H Z (Z^\top D^* Z)^{-1}$$

e

$$K_*^{\gamma\gamma} = \left(Z^\top D^* Z\right)^{-1} \{I_q + (Z^\top H T C^\top X) K_*^{\beta\beta} X^\top C T H Z (Z^\top D^* Z)^{-1}\}.$$

em que I_q é a matriz identidade de ordem q.

3.3.2 Testes da razão de verossimilhanças e Wald para verificar Dispersão constante

Considere a hipótese nula de homoscedasticidade, $\mathcal{H}_0 : \phi_1 = \ldots \phi_n = \phi$. Testar esta hipótese equivale a testar

$$\mathcal{H}_0 : \gamma_{(q-1)} = 0,$$

em que $\gamma_{(q-1)} = (\gamma_2, \ldots, \gamma_q)^\top$, no modelo definido em (3.5) com $z_{i1} = 1$ para $i = 1, \ldots, n$. Nesse contexto a estatística da razão de verossimilhança (RV) é

$$RV = 2\{\ell(\hat{\beta}, \hat{\gamma}) - \ell(\tilde{\beta}, \tilde{\gamma})\},$$

em que $\ell(\beta, \gamma)$ é o logaritmo da função de verossimilhança e $(\tilde{\beta}^\top, \tilde{\gamma}^\top)^\top$ é o estimador de máxima verossimilhança restrito de $(\beta^\top, \gamma^\top)^\top$, obtido pela imposição da hipótese nula.

A estatística de Wald para testar a hipótese acima é dada por

$$W = \hat{\gamma}_{(q-1)}^\top \left(\hat{K}^{\gamma\gamma}_{*(q-1)(q-1)}\right)^{-1} \hat{\gamma}_{(q-1)},$$

em que $\left(\hat{K}^{\gamma\gamma}_{*(q-1)(q-1)}\right)^{-1}$ é igual a $\left(K^{\gamma\gamma}_{*(q-1)(q-1)}\right)^{-1}$, avaliado no estimador de máxima verossimilhança irrestrito e $\hat{\gamma}_{(q-1)}$ é o estimador de máxima verossimilhança de $\gamma_{(q-1)}$.

Sob condições usuais de regularidade e sob $\mathcal{H}_0$, RV e W convergem em distribuição para $\chi^2_{(q-1)}$. Assim, os testes acima podem ser realizados usando valores críticos aproximados obtidos de quantis da distribuição $\chi^2_{(q-1)}$.

3.4 Técnicas de diagnóstico

Uma etapa importante na análise de um ajuste de regressão é a verificação de possíveis afastamentos das suposições feitas para o modelo, principalmente na parte aleatória (y_k), bem como a existência de observações extremas que podem causar desvios nos resultados do ajuste. Sabemos que todos os modelos são inevitavelmente simplificações, aproximações da realidade, e desse modo, uma etapa imprescindível da análise de regressão é a validação do modelo, no sentido de avaliar a qualidade desta aproximação.

Em uma direção, o interesse recai em avaliar possíveis afastamentos das suposições admitidas para o modelo, entre as quais está a distribuição de probabilidade para os dados. Em outra direção, o interesse recai em investigar a robustez do modelo sob pequenas pertubações nas formulações iniciais, no sentido de avaliar a estabilidade dos resultados inferenciais. O modelo é considerado não robusto se pequenas pertubações na sua constituição original implicam em resultados significativamente distintos.

Uma medida global de qualidade do ajuste pode ser obtida através do cálculo do pseudo-R^2 definido como o quadrado do coeficiente de correlação amostral entre $\hat{\eta}$ e $g(y)$.

Em que $0 \leq R^2 \leq 1$, e quanto mais próximo de 1 for seu valor, melhor a qualidade do ajuste. Além do pseudo-R^2, também foram considerados os critérios de informação de Akaike (AIC) e Bayesiano(BIC) na seleção de modelos, para mais detalhes AKAIKE (1974). A seguir serão apresentadas outras medidas de diagnóstico.

3.4.1 Resíduos

Podemos definir como resíduo uma medida que objetiva identificar discrepâncias entre o modelo ajustado e os dados. Assim, é compreensível que a maioria dos resíduos esteja baseada na diferença $y_i - \widehat{E(y_i)}$. Neste sentido foi definido o primeiro resíduo ordinário para o modelo de regressão beta por FERRARI & CRIBARI-NETO (2004), dado por

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var}(y_i)}}, \tag{3.8}$$

em que $\mu_i = g^{-1}\left(x_i^\top \hat{\beta}\right)$ e $\widehat{\text{var}}(y_i) = \mu_i(1-\mu_i)/(1+\hat{\phi})$.

Os resíduos construídos a partir da função desvio (MCCULLAGH & NELDER, 1989) se baseiam na distância para cada observação entre o máximo do logaritmo da função de verossimilhança do modelo saturado e máximo do logaritmo da função de verossimilhança do modelo em investigação.

Como para valores grandes de ϕ, $\tilde{\mu}_i \approx y_i$, em que $\tilde{\mu}_i$ é a estimativa de μ do modelo saturado, FERRARI & CRIBARI-NETO (2004) propõem um resíduo componente desvio para o modelo de regressão beta dado por

$$r_i^d = sinal(y_i - \hat{\mu}_i)\left[2\left(\ell_i\left(\tilde{\mu}_i, \tilde{\phi}\right) - \ell_i\left(\hat{\mu}_i, \hat{\phi}\right)\right)\right]^{1/2}, \tag{3.9}$$

em que $\mu_i = g^{-1}\left(x_i^\top \hat{\beta}\right)$ e $\widehat{\text{var}}(y_i) = \mu_i(1-\mu_i)/(1+\hat{\phi})$.

Podemos notar que a i-ésima observação contribui com uma quantidade $(r_i^d)^2$ ao desvio, logo, uma observação com um valor absoluto grande de r_i^d pode ser vista como discrepante.

Outro resíduo utilizado para verificar a qualidade do ajuste proposto por ESPINHEIRA (2007), é o resíduo ponderado padronizado 2 dado por

$$r_i^{pp} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\nu_i(1 - h_{ii}^*)}}, \tag{3.10}$$

sendo h_{ii}^* o i-ésimo elemento de $H^* = \hat{W}^{1/2}\hat{X}(X^\top \hat{W} X)^{-1} X^\top \hat{W}^{1/2}$.

3.4.2 Análise de influência

Um modelo ajustado é uma representação de aspectos essenciais dos dados. No entanto, aspectos importantes de um modelo podem ser dominados por uma única observação. Assim, 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 (1977) e as matrizes de alavanca.

A terminologia pontos de alavanca deve-se ao fato de tais pontos exercerem uma influência desproporcional no próprio valor ajustado. Em modelos lineares normais a medida de alavancagem está associada à matriz de projeção da solução de mínimos quadrados da regressão linear de y versus X, dada por $H = X(X^\top X)^{-1}X^\top$ (HOAGLIN & WELSCH, 1978). De uma forma geral, como pontuaram alguns autores, tais como YOSHIZOE (1991), St. LAURENT & COOK (1992), entre outros, uma medida de alavancagem deve refletir mais diretamente a influência de y_i no próprio valor ajustado. Sob este ponto de vista, WEI et al. (1998) propõem uma matriz de alavanca generalizada que pode ser obtida para uma classe mais ampla que a dos estimadores de mínimos quadrados, tais como os estimadores de máxima verossimilhança, estimadores de método dos momentos e até estimadores obtidos sob o enfoque bayesiano.

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. Apesar de ter sido desenvolvida originalmente para modelos normais lineares, aproximações para a distância de Cook têm sido utilizadas em diversas classes de modelos. Algumas referências são: PREGIBON (1981), COOK & WEISBERG (1982), ATKINSON (1985), CORDEIRO & PAULA (1992). A abordagem de deleção individual de casos, em que se baseia a distância de Cook, é um exemplo de uma análise de influência global.

Apesar de tipicamente a detecção de observações (casos) influentes se basear em deleção, esta é apenas uma das muitas maneiras de perturbar a formulação dos dados para

acessar influência. Pequenas modificações dos valores de uma covariável, por exemplo, podem assinalar estruturas relevantes nos dados que normalmente não seriam detectadas por deleção. Uma análise de influência mais adequada deve considerar pequenas perturbações em diferentes elementos dos dados como, por exemplo, as covariáveis, o vetor de respostas ou a dispersão assumida. Este tipo de diagnóstico pode ser obtido utilizando o método de influência local, mais detalhes COOK (1986).

Como visto anteriormente, a distância de Cook é uma medida da influência de cada observação sobre as estimativas dos parâmetros de regressão, dada por

$$k^{-1}(\hat{\beta} - \hat{\beta_{(i)}})^\top X^\top W X(\hat{\beta} - \hat{\beta_{(i)}}),$$

em que, $\hat{\beta}_{(i)}$ é o parâmetro estimado sem a i-ésima observação. Ela mede a distância ao quadrado entre $\hat{\beta}$ e $\hat{\beta}_{(i)}$. Para evitar o ajuste do modelo $n+1$ vezes, pode-se utilizar uma aproximação à distância de Cook dada por

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

É comum construir o gráfico de C_i versus as observações. Outras medidas de diagnóstico podem ser consideradas, como medidas de influência local (ESPINHEIRA, 2007).

Para estender o conceito de alavanca, originalmente definido no modelo de regressão normal linear, para modelos mais gerais, WEI et al. (1998) buscaram captar o sentido essencial deste termo em Estatística. Baseados no ponto de vista de que uma medida de alavancagem deve refletir diretamente a influência de y_i no próprio valor ajustado, os autores propõem a matriz de alavancagem generalizada *(generalized leverage)* para estimadores de θ, dada por

$$GL_{(\theta)} = \frac{\partial \hat{y}}{\partial y^\top},$$

em que $\hat{y}$ é um estimador de y. Os autores mostram que a matriz de alavanca generalizada para $\hat{\theta}$, estimador de máxima verossimilhança de θ, tal que $E(y) = \mu = \mu(\theta)$, é dada por

$$GL_\theta = D_\theta \left(-\frac{\partial^2 \ell}{\partial \theta \partial \theta^\top} \right)^{-1} \frac{\partial^2 \ell}{\partial \theta \partial y^\top}$$

avaliada em $\hat{\theta}$, em que $D_\theta = \partial\mu/\partial\theta^\top$.

Utilizando este resultado FERRARI & CRIBARI-NETO (2004) obtêm para o modelo de regressão a matriz de alavanca generalizada para $\hat{\beta}$ e $\hat{\phi}$ dada por

$$GL(\beta,\phi) = GL(\beta) + \frac{1}{\gamma\phi} TX(X^\top QX)^{-1}X^\top\{Tf(f^\top TX(X^\top QX)^{-1})X^\top TM - b^\top)\},$$

com $GL(\beta) = TX(X^\top QX)^{-1}X^\top TM$ e $Q = \text{diag}(q_1,\ldots,q_n)$, em que

$$q_i = \left[\phi\{\psi^{'}(\mu_{i\phi}) + \psi^{'}((1-\mu_i)\phi)\} + (y_i^* - \mu_i^*)\frac{g^{''}(\mu_i)}{g^{'}(\mu_i)}\right]\frac{1}{\{g^{'}(\mu_i)\}^2}, \quad i = 1,\ldots,n.$$

Outras quantidades são dadas por $M = \text{diag}(m_1,\ldots,m_n)$ com $m_i = 1/y_i(1-y_i)$, $f = (f_1,\ldots,f_n)^\top$ com $f_i = \{c_i - (y_i^* - \mu_i^*)\}$ e $b = (b_1,\ldots,b_n)^\top$ com $b_i = -(y_i - \mu_i)/y_i(1-y_i)$, $i = 1,\ldots,n$. Quando ϕ é grande, $GL(\beta,\phi) \approx GL(\beta)$.

3.4.3 Gráfico de probabilidade meio-normal com envelopes

Como a distribuição dos resíduos não é conhecida, gráficos de probabilidade meio-normal com envelope simulado são ferramentas de disgnóstico muito úteis. A ideia é acrescentar ao gráfico meio - normal usual um envelope simulado que pode ser usado para decidir se as respostas observadas são consistentes com o modelo ajustado.

Capítulo 4

Materiais

4.1 Descrição da amostra

Os dados utilizados para modelagem da taxa de analfabetismo do Estado da Paraíba foram obtidos através do Atlas do Desenvolvimento Humano no Brasil 2013, disponível no site `http://www.pnud.org.br`. Ele conta com mais de 180 indicadores de população, educação, habitação, saúde, trabalho, renda e vulnerabilidade.

Concebido como uma ferramenta simples e amigável de disponibilização de informações, o Atlas Brasil 2013 facilita o manuseio de dados e estimula análises. A ferramenta oferece um panorama do desenvolvimento humano dos municípios e as desigualdades entre eles em vários aspectos. Sua relevância vem justamente da capacidade de fornecer informações sobre a unidade político-administrativa mais próxima do cotidiano dos cidadãos: o município.

Além das informações socioeconômicas do município, outra importante informação sobre o município foi o gasto do governo federal com assistencialismo nos municípios paraibanos. Essa informação foi extratída do Ministério do desenvolvimento social e combate à fome, disponível no site `http://www.mds.gov.br`.

A amostra foi composta pelos 223 municípios do estado da Paraíba e os dados coletados referem-se ao ano de 2010. As variáveis explicativas e a nomenclatura adotada nessa monografia são

- Taxa de analfabetismo (18 anos ou mais de idade) - $Taxa$;
- Mortalidade infantil: Número de crianças que não deverão sobreviver ao primeiro

ano de vida em cada 1.000 crianças nascidas vivas - MI;

- Renda per capita : Razão entre o somatório da renda de todos os indivíduos residentes em domicílios particulares permanentes e o número total desses indivíduos - *Renda*;
- Índice de Gini: Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita. Esse índice varia de 0 (quando não hà desigualdade) a 1 (desigualdade máxima) - *Gini*;
- Percentual da população em domicílios com banheiro e água encanada - AE;
- Percentual da população em domicílios com coleta de lixo - *Lixo*;
- Percentual da população em domicílios com densidade populacional maior que 2 - *Densidade*;
- Percentual de pobres - PP;
- População rural - PR;
- População urbana - PU;
- 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 - *Gasto*.

A variável resposta foi definida como a taxa de analfabetismo, enquanto as demais variáveis descritas acima foram consideradas como variáveis explicativas, que podem influenciar a variável resposta.

4.2 Aspectos computacionais

Todos os resultados gráficos e numéricos inerentes a análise de regressão (estimação dos parâmetros, testes de hipóteses, análise de diagnóstico) apresentados nesta monografia foram obtidos utilizando o ambiente de programação, análise de dados e gráficos R em sua versão 3.0.2 para sistema operacional Microsoft Windows, que se encontra disponível gratuitamente através do site `http://www.R-project.org`. O `R` foi criado por Ross Ihaka e Robert Gentleman na Universidade de Auckland com o objetivo de produzir um

ambiente de programação parecido com S, uma linguagem desenvolvida no AT&T Bell Laboratories, cuja versão comercial é o S-Plus, tendo as vantagens de ser de livre distribuição e possuir código fonte aberto. Maiores detalhes sobre o R podem ser encontrados em CRIBARI NETO & ZARKOS (1999).

O procedimento computacional para obtenção das estimativas de máxima verossimilhança dos parâmetros foi desenvolvido utilizando o pacote *betareg*, que consiste em um conjunto de rotinas voltadas para a construção de modelos de regressão beta. Além das rotinas para otenção das estimativas dos parâmetros, o pacote contém um conjunto de gráficos úteis para a análise de adequação do modelo. Pra maiores detalhes consultar (CRIBARI-NETO & ZELEIS, 2010).

O pacote GAMLSS implementado no software R (STASINOPOULOS et al., 2008) permite o ajuste de diversos modelos da classe em estudo. No pacote GAMLSS já estão implementadas o ajuste com várias distribuições de probabilidade. Sendo incluídas no pacote tanto distribuições simples, que pertencem à família exponencial, quanto várias distribuições que envolvem 3 ou quatro parâmetros. O pacote permite ainda selecionar modelos a partir de procedimentos automáticos que utilizam o critério de informação de Akaike generalizado (AKAIKE, 1983) e construir gráficos de diagnóstico utilizando resíduos quantílicos (DUNN & SMYTH, 1996).

Esta monografia foi digitada utilizando o sistema de tipografia LaTeX desenvolvido por Leslie Lamport em 1985, que consiste em uma série de macros ou rotinas do sistema TeX (criado por Donald Knuth na Universidade de Stanford) que facilitam o desenvolvimento da edição do texto. Detalhes sobre o sistema de tipografia LaTeX podem ser encontrados em LAMPORT (1994) ou através do site `http://www.tex.ac.uk/CTAN/latex`.

Capítulo 5

Resultados e Discussões

5.1 Análise exploratória

Para analisar a taxa de analfabetismo do Estado da Paraíba, torna-se imprescindível realizar inicialmente uma análise exploratória dos dados, com o objetivo de identificar as principais particularidades desta taxa para o estado.

Na Tabela 5.1, encontram-se algumas estatísticas descritivas da variável resposta taxa de analfabetismo. Para os 223 municípios do estado da Paraíba, o valor médio da taxa de analfabetismo foi de 0,3186 com um desvio padrão de 0,0675. O coeficiente de variação foi de aproximadamente 21%, indicando uma baixa variabilidade no valor da taxa de analfabetismo nos municípios observados. Os valores extremos verificados foram respectivamente, 0,0854 referente à capital João Pessoa e 0,4613 referente ao município de Pedro Régis. Quanto à simetria e curtose, verifica-se que a variável é assimétrica negativa, e possui um achatamento leptocúrtico, indicando que a distribuição normal não é adequada para modelar esses dados, o que é esperado, já que dados restritos ao intervalo unitário tendem a ser assimétricos.

Tabela 5.1: Estatísticas descritivas da variável taxa de analfabetismo.

Estatística	Estimativa
Mínimo	0,0854
Máximo	0,4613
Média	0,3186
Mediana	0,3241
Desvio Padrão	0,0675
Assimetria	$-0,4021$
Curtose	0,3039

A Figura 5.1 apresenta o histograma da variável resposta $Taxa$. Verifica-se que essa variável é de natureza contínua e seu domínio é estritamente positivo. Além disso, é possível observar uma assimetria negativa, também verificada pelo coeficiente de assimetria anteriormente.

Através do box-plot da variável $Taxa$, Figura 5.2, é possível identificar a existência de observações atípicas. Do total da amostra, apenas 4 municípios apresentaram valor atípico para esta variável, ou seja, os municípios que apresentaram menor taxa de analfabetismo. Essas observações, referem-se aos municípios de João Pessoa, Cabedelo, Campina Grande e Várzea, que apresentaram as seguintes porcentagens: 8,5%, 11,2%, 12,4% e 14,1%, respectivamente.

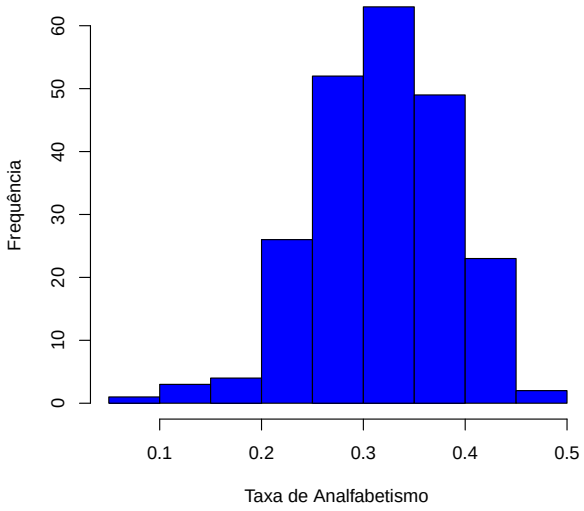

Figura 5.1: Histograma da variável taxa de analfabetismo.

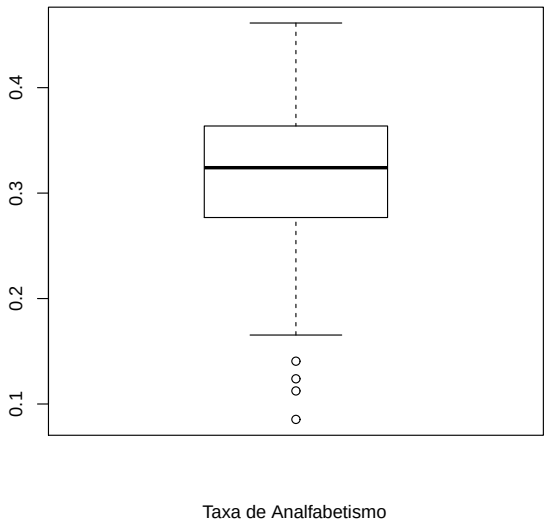

Figura 5.2: Box-plot da variável taxa de analfabetismo.

Além da variável resposta, também verificou-se as principais características das covariáveis presentes no estudo. 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, sendo o desvio padrão igual a 5,61. A covariável Renda per capita ($Renda$) registrou um valor médio igual a R$ 277,35, sendo os valores mínimo e máximo iguais a R$ 166,28 e R$ 1.036,21, respectivamente. O desvio padrão para esta variável foi igual a 92,07. Para a variável Índice de Gini ($Gini$), o valor médio foi igual a 0,50, com desvio padrão de 0,04. Os extremos observados para está variável, foram 0,40 e 0,70. Com relação a variável percentual da população em domicílios com banheiro e água encanada (AE), observou-se que, em média, 64,27% dos municípios da amostra possuem domicílios com banheiro e água encanada. Quanto 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 encontra nesta situação. A população rural (PR) média observada nos municípios foi igual a 4.160 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.

Para verificar a relação existente entre a variável resposta e as variáveis explicativas descritas na seção anterior, calcularam-se as correlações de Pearson entre essas variáveis (Tabela5.2). Verifica-se que nenhuma das covariáveis apresentaram correlações superiores a 0,80 entre si, o que pode ser um indicativo da não existência de multicolinearidade entre as variáveis. Entretanto, observa-se que as covariáveis $Renda$ e PP apresentaram uma correlação negativa elevada $(-0,7531)$.

Tabela 5.2: Matriz de correlação para as variáveis observadas.

	Taxa	*MI*	*Renda*	*Gini*	*AE*	*Lixo*	*Densidade*	*PP*	*PR*	*PU*	*Gasto*
Taxa	1,0000	0,3554	$-0,6028$	0,0027	$-0,4046$	0,3907	0,6761	$-0,0274$	$-0,3688$	0,5061	$-0,2372$
MI	0,3554	1,0000	$-0,4271$	$-0,0505$	$-0,3806$	0,2137	0,4906	$-0,0853$	$-0,2185$	0,2931	$-0,1226$
Renda	$-0,6028$	$-0,4271$	1,0000	0,4305	0,4708	$-0,2958$	$-0,7531$	0,1057	0,6581	$-0,5485$	0,1597
Gini	0,0027	$-0,0505$	0,4305	1,0000	0,0586	0,0164	0,1092	0,2316	0,2663	$-0,1031$	$-0,0838$
AE	$-0,4046$	$-0,3806$	0,4708	0,0586	1,0000	$-0,1551$	$-0,5676$	0,0361	0,2533	$-0,3309$	0,0818
Lixo	0,3907	0,2137	$-0,2958$	0,0164	$-0,1551$	1,0000	0,3820	0,0508	$-0,1124$	0,2618	$-0,2052$
Densidade	0,6761	0,4906	$-0,7531$	0,1092	$-0,5676$	0,3820	1,0000	$-0,0438$	$-0,3946$	0,4888	$-0,2373$
PP	$-0,0274$	$-0,0853$	0,1057	0,2316	0,0361	0,0508	$-0,0438$	1,0000	0,1905	$-0,2325$	$-0,0593$
PR	$-0,3688$	$-0,2185$	0,6581	0,2663	0,2533	$-0,1124$	$-0,3946$	0,1905	1,0000	$-0,4133$	0,0648
PU	0,5061	0,2931	$-0,5485$	$-0,1031$	$-0,3309$	0,2618	0,4889	$-0,2325$	$-0,4133$	1,0000	$-0,1416$
Gasto	$-0,2372$	$-0,1226$	0,1597	$-0,0838$	0,0818	$-0,2052$	$-0,2373$	$-0,0593$	0,0648	$-0,1416$	1,0000.

No entanto, para analisar de maneira mais concisa as relações entre as variáveis explicativas e a variável de interesse, é importante a utilização de métodos mais robustos. A seguir, apresentamos os resultados da aplicação do modelo de regressão beta na modelagem da taxa de analfabetismo para o estado da Paraíba.

5.2 Modelo de regressão beta

5.2.1 Ajuste do modelo de regressão beta para explicar a taxa de analfabetismo

Para verificar de forma mais precisa as relações existentes entre a variável resposta e as variáveis explicativas e a magnitude dessas relações faz-se necessário à utilização da análise de regressão. Nesse contexto, o modelo de regressão normal linear ocupa lugar de destaque na literatura, assumindo para a variável resposta uma distribuição normal, além de uma relação linear entre a média da variável resposta e as covariáveis. No entanto, o modelo normal torna-se inadequado para modelagem de dados restritos ao intervalo unitário, já que estes além de serem assimétricos possuem um padrão típico de heterocedasticidade, violando, portanto, as suposições básicas do modelo de regressão linear. Dessa forma, torna-se necessário modelos apropriados para dados restritos a um determinado intervalo (a, b) de forma a acomodar adequadamente as principais características inerentes a esses tipos de dados. 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.

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 uma exaustiva seleção de modelos, verificou-se que o modelo mais adequado para explicar a taxa de analfabetismo do Estado da Paraíba, foi o modelo com função de ligação logit. Os valores dos critérios de informação AIC (Critério de informação de Akaike), BIC (Critério de informação Bayesiano) e pseudo-R^2 foram $-723,1041$, $-702,6611$ e $0,5727$, respectivamente. Neste modelo, apenas as variáveis *Renda*, *Gini*,

Densidade e *Gasto* foram significativas ao nível de 5% para a média da variável resposta.

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 5.3 são sumarizados os resultados desses testes. 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

$$\hat{\phi}_i = \exp\left(\gamma_1 + \gamma_2 PP + \gamma_3 PR\right).$$

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 5.4. 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 5.3: Teste da Razão de Verossimilhança e Wald.

Teste	Estatística	p-valor
razão de verossimilhança	7,6269	0,0221
Wald	2401,8	0,0000

Tabela 5.4: Estimativas dos parâmetros.

Parâmetro μ			
	Estimativa	Erro-padrão	p-valor
Intercepto	$-1,81132$	0,22650	0,0000
Renda	$-0,00260$	0,00024	0,0000
Gini	2,11577	0,38006	0,0000
Densidade	0,00751	0,00250	0,0027
Gasto	0,00271	0,00077	0,0004
Parâmetro ϕ			
Intercepto	$5,25800$	0,49560	0,0000
PP	$-0,02340$	0,01192	0,0499
PR	$0,00007$	0,00003	0,0250

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. Apesar do coeficiente de determinação não ter sido tão expressivo, vale mencionar que são inúmeros os fatores/ou variáveis que podem influenciar em tal taxa.

Através da análise dos coeficientes estimados para o modelo selecionado 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 > 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.

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 também aumenta. Por exemplo, os municípios que apresentam maior percentual de pobres tendem a apresentar respostas mais precisas. Em contrapartida, a covariável PR (população rural) exerce um efeito negativo na taxa de analfabetismo, ou seja, podemos dizer que quanto maior a população rural do município menos precisas serão as respostas.

Após a estimação do modelo, torna-se fundamental a verificação das suposições intrísecas ao modelo de regressão. Neste sentido, foi realizada uma análise de diagnóstico com base no resíduo padronizado 2, proposto e estudado por ESPINHEIRA (2007) a fim de identificar possíveis desvios das suposições do modelo de regressão beta. Na Figura 5.3

são apresentados os gráficos dos resíduos versus os índices das observações (a) e o gráfico dos resíduos versus o preditor linear (b). Com base nos gráficos é possível verificar que o modelo encontra-se bem ajustado, dado que a distribuição dos resíduos (a) encontram-se dentro dos limites $(-3, 3)$, e os pontos estão dispersos de forma aleatória em torno do zero. Além disso, a partir do gráfico (b) é possível verificar que duas observações apresentaram comportamento atípico em relação as demais observações, essas referem-se aos municípios de Cabedelo e Jacaraú. No entanto, não há fortes indícios de violação de que a função de ligação utilizada é adequada.

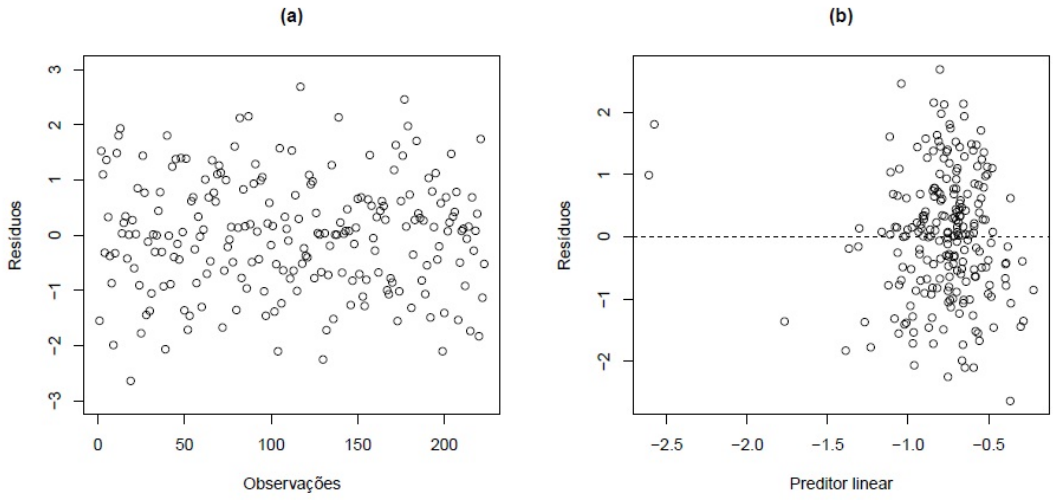

Figura 5.3: (a) Resíduos versus índices das observações; (b) Resíduos versus preditor linear.

Para a verificar à adequação da distribuição beta aos dados foi construído o gráfico dos quantis da distribuição normal padrão versus resíduos, com envelope simulado (Figura 5.4). Nota-se que todos os pontos encontram-se dentro do envelope simulado. Dessa forma podemos afirmar que não há indícios de afastamento da suposição de que o modelo de regressão beta fornece uma boa representação para os dados. Logo, a distribuição $\mathrm{B}(\mu_i, \phi_i)$ configura-se como adequada para modelar a taxa de analfabetismo do estado da Paraíba.

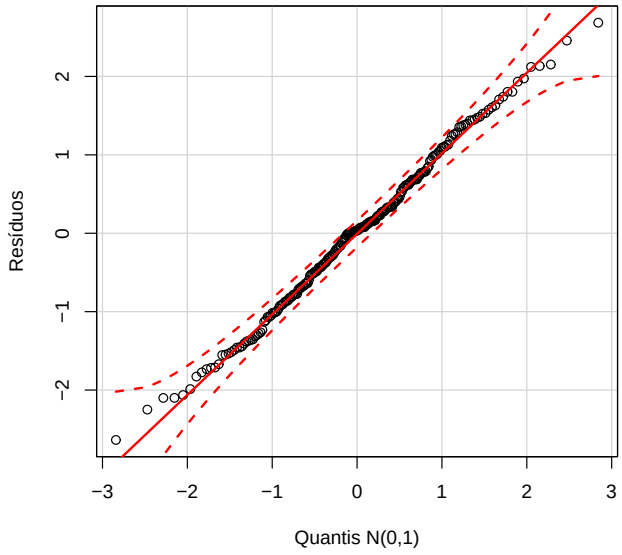

Figura 5.4: Gráfico da distribuição normal padrão versus resíduos.

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 5.5. Desse modo, é possível verificar que no gráfico da distância de Cook (a) a observação 40, que corresponde ao município de Cabedelo, encontra-se destacada das demais, identificando-se como um ponto de influência. Por outro lado, o gráfico (b), que refere-se aos possíveis pontos de alavanca, aponta duas observações se destacam em relação às demais, sendo elas a observação 50 e 94, que correspondem aos municípios de Campina Grande e João Pessoa, respectivamente. Vale salientar que esses dois municípios são alguns dos que apresentaram uma das menores taxas de analfabetismo na amostra.

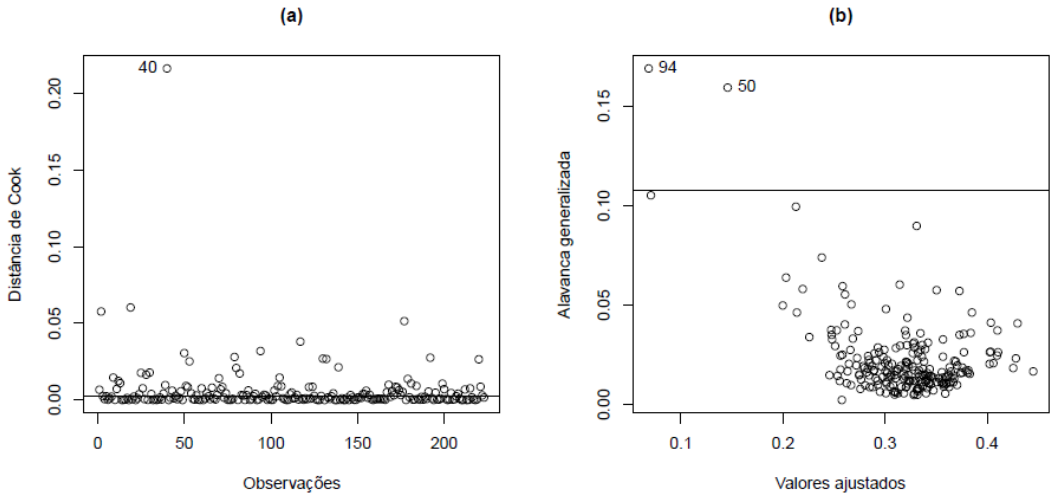

Figura 5.5: (a) Distância de Cook versus ordem das observações; (b) Alavancagem generalizada versus preditor linear.

Para analisar a influência das observações 40, 50 e 94 nas estimativas dos parâmetros, ajustou-se novamente o modelo descrito anteriormente sem essas observações e foram calculadas as variações percentuais em relação às estimativas obtidas com os dados completos. Os resultados encontram-se na Tabela a seguir (5.5). Verificamos que apesar da retirada das observações não se verifica mudanças substanciais nos resultados inferenciais, o que evidencia a robustez do modelo.

Tabela 5.5: Variação percentual das estimativas dos parâmetros sem as observações 40, 50 e 94 que correspondem aos municípios de Cabedelo, Campina Grande e João Pessoa, respectivamente.

Parâmetro μ				
	Variação% (40)	Variação% (50)	Variação% (94)	Variação% (50 e 94)
Intercepto	3,3361	$-2,2170$	1,8307	$-0,2182$
Renda	$-6,5483$	4,2609	$-4,2762$	0,8651
Gini	$-0,8301$	$-2,0941$	$-1,1190$	$-2,1361$
Densidade	4,8811	$-3,7763$	2,8219	$-0,9318$
Gasto	2,3771	4,3377	1,6842	5,5613
Parâmetro ϕ				
Intercepto	$-2,6816$	$-1,4074$	0,3614	$-0,2092$
PP	$-16,1249$	$-7,7844$	$-0,6844$	$-2,4371$
PR	$-6,7132$	$-0,7287$	$-9,1937$	$-5,0233$

Como visto anteriormente, após a validação do modelo é possível verificar as relações entre a variável resposta e as variáveis explicativas selecionadas. Portanto, consideramos o modelo estimado

$$\hat{\mu}_i = \frac{\exp(-1,81130 - 0,00260Renda + 2,11577Gini + 0,00751Densidade + 0,00271Gasto)}{1 + \exp(-1,81130 - 0,00260Renda + 2,11577Gini + 0,00751Densidade + 0,00271Gasto)}$$

e

$$\hat{\phi}_i = \exp\left(5,25800 - 0,02340PP + 0,00007PR\right)$$

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 importante 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, utilizando o modelo de regressão GAMLSS, introduzido por RIGBY &

STASINOPOULOS (2001, 2005). Utilizou-se a classe de modelos GAMLSS para o ajuste do modelo proposto, a fim de verificar, detalhadamente, a contribuição da variável (*Gasto*) suavizada. Vale salientar que o GAMLSS utiliza outra parametrização diferente da *betareg*, no entanto, a recorrência na utilização do modelo GAMLSS foi uma tentativa de verificar a relação entre a covariável *Gasto* e a variável resposta, e não nas estimativas dos parâmetros. A classe de modelos utilizada torna-se satisfatória, uma vez que é possível estimar a função entre as variáveis observadas. A Figura 5.6 mostra a curva suavizada da variável *Gasto*. Observa-se que a contribuição parcial desta variável é crescente na maior parte do seu domínio. No entanto, verifica-se uma contribuição negativa para a taxa de analfabetismo para valores inferiores a R$ 165,00 reais.

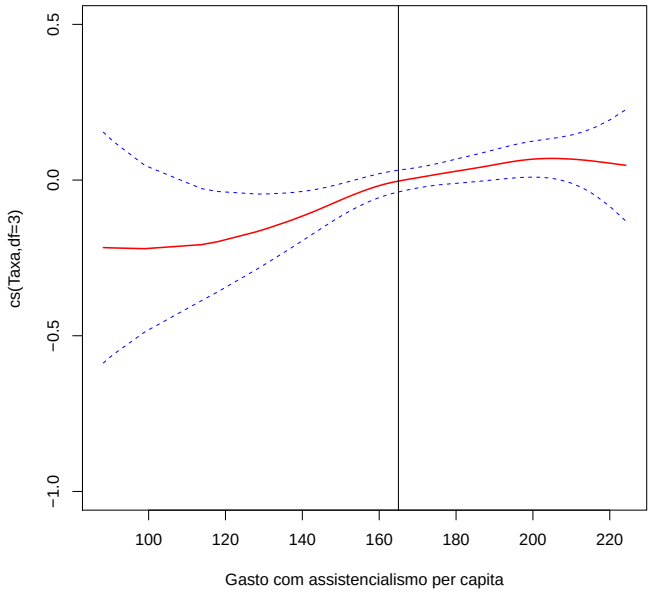

Figura 5.6: Gráfico da curva suavizada para a variável gasto com assistencialismo per capita.

A análise realizada ainda não retrata de forma minuciosa a variável *Gasto*, visto que os dados observados na amostra, com relação a variável *Gasto*, não permitem a averiguação do comportamento dessa variável para valores superiores a R$ 224,17 reais. Desse modo, seria interessante investigar o comportamento dessa variável considerando um novo cenário. Na realidade, a finalidade é verificar o impacto estimado na taxa de

analfabetismo a medida que o gasto com assistencialismo aumenta. O impacto é obtido da seguinte forma:

$$\frac{\partial E(y_i)}{\partial Gasto_i} = \frac{\partial \mu_i}{\partial Gasto_i}, \tag{5.1}$$

Considerando a função de ligação logit, a expressão (5.1), com as covariáveis selecionadas para o modelo é dada por

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

Dessa forma, foram gerados valores para a variável gasto com assistencialismo (*Gasto*), que variam de R$ 85 a R$ 600. Consideramos a variável renda per capita (*Renda*) fixada no 1º, 2º e 3º quartil. As demais variáveis do modelo foram fixadas na mediana. Desse modo, foi possível obter uma maior compreensão dessa variável em relação aos dados observados. Conforme a Figura 5.7, verifica-se que a contribuição dessa variável é crescente para valores inferiores a R$ 411,00 reais. Por outro lado, para valores superiores, 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.

Com base na Figura 5.7, foi possível captar algumas características da variável gasto com assistencialismo. Podemos observar que a variável gasto não se restringe a contribuir apenas de forma positiva na variável resposta e sim evidenciar uma possível proposta ou estratégia de abordagem para que se busque a redução da taxa de analfabetismo no estado da Paraíba.

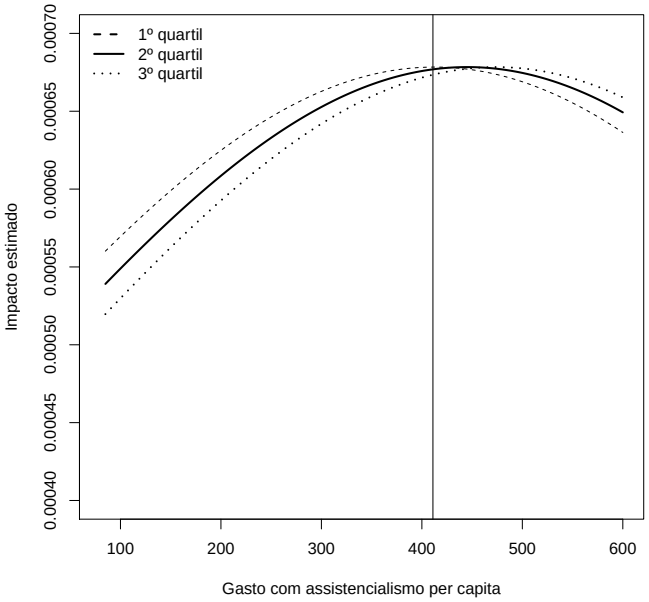

Figura 5.7: Gráfico do gasto com assistencialismo per capita versus impacto estimado na taxa de analfabetismo.

Capítulo 6

Conclusão

Para modelar a taxa de analfabetismo do Estado da Paraíba referente aos dados do censo (2010), utilizamos a classe de modelos de regressão beta proposta por FERRARI & CRIBARI-NETO (2004). Após uma exaustiva seleção de modelos, os mesmos foram comparados com base nos critérios de informação de Akaike (AIC), Bayesiano (BIC). Após a seleção verificamos que o modelo selecionado foi o modelo com função de ligação logit. Para este modelo constatamos que as variáveis renda per capita, índice de Gini, percentual da população em domicílios com densidade maior que 2 e gasto com assistencialismo foram significativas ao nível de 5% para predizer a taxa de analfabetismo do Estado da Paraíba. Adicionalmente, realizou-se uma modelagem para o parâmetro de precisão do modelo, uma vez que a suposição de dispersão constante foi rejeitada nos testes da razão de verossimilhança e Wald.

Para validar o modelo estimado é indispensável verificar as suposições inerentes ao modelo de regressão beta. Para isso, realizou-se uma análise de diagnóstico com base no resíduo padronizado ponderado 2 proposto por ESPINHEIRA (2007). Com base nos gráficos dos resíduos verificou-se que o modelo selecionado é adequado para modelar a taxa de analfabetismo. O gráfico da distância de Cook versus ordem das observações destacou a observação 40, referente ao município de Cabedelo, como ponto de influência. Com relação ao gráfico da alavancagem generalizada versus valores ajustados, pode-se observar que os municípios de João Pessoa e Campina Grande foram classificados como pontos de alavanca. Com o objetivo de avaliar a influência do município de Cabedelo nas estimativas dos parâmetros, ajustou-se novamente o modelo sem essa observação. Vimos que a retirada da observação 40 não afetou substancialmente as estimativas dos

parâmetros, evidenciando a robustez do modelo ajustado, bem como, a exclusão das observações que se caraterizaram como pontos de alavanca (50 e 94).

De acordo com as estimativas dos parâmetros é possível verificar o sentido da relação entre a variável resposta e as variáveis explicativas. 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. A partir do gráfico da curva suavizada da variável *Gasto*, foi possível analisar de forma mais detalhada a contribuição desta variável ao longo de seu domínio na taxa de analfabetismo. Desse modo, apesar de contribuir de forma crescente para a variável resposta, foi possível observar que esta covariável contribui de forma negativa para a taxa de analfabetismo no Estado da Paraíba, em situações em que o gasto com assistencialismo é inferior a R$ 165,00 reais, segundo os dados da amostra.

Além disso, através do modelo de regressão beta, modelamos o impacto causado na taxa de analfabetismo no estado da Paraíba considerando outro cenário. A partir da análise, verificamos que para valores do gasto com assistencialismo superior a R$ 411,00 reais, o impacto causado na variável resposta é negativo. Sendo assim, é necessário investimento de aproximadamente R$ 411,00 per capita para que se obtenha uma redução na taxa de analfabetismo do Estado da Paraíba. Salientando que, essa redução é mais expressiva no grupo da população com renda per capita de até R$ 230,12.

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.

6.1 Sugestões para trabalhos futuros

Com a flexibilidade da classe de modelos de regressão beta, para a modelagem de variáveis pertencentes ao intervalo unitário, é possível obter estimativas mais precisas e que condizem de fato, com a relaidade apresentada pelos dados. Dessa forma, algumas extensões desta monografia podem ser sugeridas para trabalhos futuros, segue abaixo algumas delas;

- Nesta monografia, considerou-se apenas o gasto referente a um programa de assistencialismo. Dessa forma, torna-se interessante investigar a relação de outros programas de assistência social com a educação da população;
- Inclusão de novas variáveis referentes a aspectos sociais da população;
- Incluir a latitude e a longitude para verificação de dependência espacial entre os municípios;
- Realizar uma comparação com os modelos usuais de regressão (Regressão logística, Modelos Lineares Generalizados) com o modelo de regressão beta, com o objetivo de ressaltar a flexibilidade e adequabilidade dessa classe de modelos;
- Explicar a taxa de analfabetismo nas regiões Sul, Sudeste, Norte, Nordeste e Centro-Oeste e avaliar o impacto do gasto com o assistencialismo;
- Ajustar um modelo de regressão beta para explicar a taxa de analfabetismo nos municípios brasileiros e adicionalmente incluir variáveis dummy's para identificar as cinco regiões do Brasil.

Referências Bibliográficas

[1] AKAIKE, H. (1983). Information measures and model selection. International Statistical Institute, Voorburg, v. 44, p. 277-291.

[2] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

[3] ATKINSON, A. C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. New York: Oxford University Press.

[4] BOX, G. E. P. ; COX, D. R. (1964). An analysis of transformations. *Journal of the Royal Statical Society* B 26,211-252.

[5] BOX, G. E. P. ; DRAPER, N. R. (1987). *Empirical Model-Building and Response Surfaces*,Wiley.

[6] BREHM, J. ; GATES, S. (1993). Donut shops and speed traps: Evaluating models of supervision on police behavior. *American Journal of Political Science*,37(2),555-581.

[7] BATES, D. M. ; WATTS, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

[8] COOK, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19,15-18.

[9] COOK, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B*, 48, 133-169.

[10] COOK, R. D. ; PEÑA, D. ; WEISBERG, S. (1988). The likelihood displacement: A unifying principle for influence measures. *Communications in Statistics, Theory and Methods*, 17, 623-640.

[11] COOK, R. D. ; WEISBERG, S. (1982). *Residuals and influence in Regressions*. London: Chapman and Hall.

[12] CORDEIRO, G. M. ; PAULA, G. A. (1992). Estimation, large sample parametric tests and diagnostics for non-exponential family nonlinear models. *Communications in Statistics, Simulation and Computation*, 21, 149-172.

[13] CRIBARI-NETO, F. ; LIMA, L. B. (2007). *A misspecification test for beta regressions*.

[14] CRIBARI-NETO, F. ; ZARCOS, S. G. (1999). R: Yet another aconometric programming environment, *Journal of Aplied Econometrics* 14, 319-329.

[15] CRIBARI-NETO, F. ; ZELEIS, A. (2010). Beta regression in R. *Journal of Statistical Software* 34(2),1-24.

[16] DUNN, P. K. ; SMYTH, G. K. (1996). Randomized quantile residuals. *Journal of Computacional and Graphical*, 5, 1-10.

[17] ESPINHEIRA, P. L. (2007). *Regressão beta*. Tese de Doutorado, Instituto de Matemática e Estatística, Universidade de São Paulo (USP), São Paulo.

[18] ESPINHEIRA, P. L. ; FERRARI, S. L. P. ; CRIBARI-NETO, F. (2008a). On beta regression residuals. *Journal of Applied Statistics* 35, 407-419.

[19] ESPINHEIRA, P. L. ; FERRARI, S. L. P. ; CRIBARI-NETO, F. (2008b). Influence diagnostics in beta regression. *Computational Statistics and Data Analysis* 52, 4417-4431.

[20] FERRARI, S. L. P. ; CRIBARI-NETO, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics* 31, 799-815.

[21] FERRARI, S. L. P. ; PINHEIRO, E. C. (2011). Improved likelihood inference in beta regression. *Journal of Statistical Computation and Simulation* 81, 4, 431-443.

[22] FREIRE. P. (1979). *Educação e mudança*. 24 ed. (trad. Moacir Gadotti e Lilian Lopes Martins) Rio de Janeiro: Paz e Terra.

[23] HANCOX,D.; HOSKIN, C. J. ; WILSON, R.S. (2010). Evening up the score: Sexual selection favours both alternatives in the colour-polymorphic ornate rainbowfish. *Animal Behaviour*, 80(5),845-851.

[24] HASTIE, T. J. ; TIBSHIRANI,R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

[25] HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* 46, 1251–1271.

[26] HOUAISS. (2001). *Dicionário Eletrônico Houaiss* 1.0 (BR). Disponível em www.dicionariohouaiss.com.br/.

[27] HOAGLIN, D. C. ; WELSCH, R. E. (1978). The hat matriz in regression and ANOVA. *The American Statistician*, 32, 17-22.

[28] KIESCHNICK, R. ; McCULLOUGH, B. D. (2003). Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, 3, 19-213.

[29] LAMPORT, L. (1994). A Document Preparation System, 2nd. edn, Addison-Wesley, Massachusetts.

[30] McCULLAGH, P. ; NELDER, J.A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

[31] NELDER, J. A. ; WEDDERBURN, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical* A 135,370-384.

[32] PAULA, G. A. (2004). *Modelos de Regressão com Apoio Computacional*. São Paulo : IME/USP.

[33] PAOLINO, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9, 325-346.

[34] PREGIBON, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705-724.

[35] RAMSEY, J. B. (1969). Tests for specification erros in classical linear least squares regression analysis. *Journal of the Royal Statistical Society* B 31, 350-371.

[36] RAMSEY, J. B.; GILBERT, R. (1972). A monte carlo study of some small sample properties of tests for specification error. *Journal of the American Statistical Association* 67, 180-186.

[37] RAMSEY, J. B.; SCHMIDT, P. (1976). Some further results on the use of ols and blus residuals in specification error tests. *Journal of the American Statistical Association* 71,389-390.

[38] RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley, p.33.

[39] RIGBY, R. A. ; STASINOPOULOS, D. M. (2001). The GAMLSS project: a flexible approach to statistical modelling. In New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling. Eds:Klein, B. and Korsholm, L., 337-345. Odense: Denmark.

[40] RIGBY, R. A. ; STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape (with discussion), *Applied Statistics* 54, 507-554.

[41] SHUKUR, G.; EDGERTON, D. L. (2002). The small sample properties of the reset test as applied to systems of equations. *Journal of Statistical Computation and Simulation* 72, 909-924.

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

[43] SKOVGAARD, I. M.(2001). Likelihood asymptotics. *Scandinavian Journal of Statistics* 28, (2001), 3-32.

[44] SMITHSON, M. ; VERKUILEN, J. (2006). A better lemon-squeezer? Maximum likelihood regression with beta-distribuited dependent variables. *Psychological Methods*, 11,54-71.

[45] SMYTH, G. K. ; VERBYLA, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics*, 10, 695-709.

[46] St. LAURENT, R.T. ; COOK, R. D. (1992). Leverage and superleverage in nonlinear regression. *Journal of the American Statistical Association*. 87, 985-990.

[47] STASINOPOULOS, W. ; RIGBY, B. ; AKANTZILIOTOU, C. (2008). Instructions on how to use the gamlss package in R. 2ed.

[48] THURSBY, J. G.; SCHMIDT, P. (1977). Some properties of tests for specification error in a linear regression model. *Journal of the American Statistical Association* 72, 635-641.

[49] WEI, B.-C.; HU, Y.-Q.; FUNG, W.-K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, 25, 25-37.

[50] YOSHIZOE, Y. (1991). Leverage points in nonlinear regression models. *Journal of Japan Statistics Society*, 21, 1-11. 25, 25-37.

Apêndice

```
######### Regressão Beta #############
###### Leitura do conjunto de dados #####
library(tcltk) # TCL/TK para abrir o conjunto de dados.
local <- tclvalue(tkgetOpenFile(title="Abrir Banco de Dados"))
dados <- read.table(file = local, header=TRUE, dec=".")
attach(dados)
names(dados)
########## Pacotes requeridos ########
library(fBasics)
library(gamlss)
library(betareg)
library(car)
library(nortest)
library(lmtest)
######### Medidas Descritivas #######
basicStats(Taxa)
cor(dados)
# Gráficos
hist(Taxa,xlab="Taxa de analfabetismo")
boxplot(Taxa,xlab="Taxa de analfabetismo")
######### Análise de Regressão #######
modelo1=betareg(Taxa ~ MI + Renda + Gini + AE + Lixo + Densidade + PP +
PR + PU + Gasto,link="logit")
summary(modelo1)
modelo2=betareg(Taxa ~ MI + Renda + Gini + AE + Lixo + Densidade + PP +
PR + PU + Gasto,link="probit")
summary(modelo2)
modelo3=betareg(Taxa ~ MI + Renda + Gini + AE + Lixo + Densidade + PP +
PR + PU + Gasto,link="loglog")
summary(modelo3)
modelo4=betareg(Taxa ~ MI + Renda + Gini + AE + Lixo + Densidade + PP +
```

```
PR + PU + Gasto,link="cloglog")
summary(modelo4)
modelo5=betareg(Taxa ~ MI + Renda + Gini + AE + Lixo + Densidade + PP +
PR + PU + Gasto,link="cauchit")
summary(modelo5)
## Critérios de seleção
AIC(modelo1)
BIC(modelo1)
# Testes de homocedasticidade
#Modelo selecionado
modelofinal=betareg(Taxa ~ Renda + Gini + Densidade + Gasto)
# Modelo com dispersão variável
modelocompleto=betareg(Taxa ~ Renda + Gini + Densidade + Gasto|MI +
Renda + Gini + AE + Lixo + Densidade + PP + PR + PU + Gasto)
summary(modelocompleto)
#Testes
lrtest(modelofinal,modelocompleto)
waldtest(modelofinal,modelocompleto)

## Modelo com dispersão variável
modelodisp1=betareg(Taxa ~ Renda + Gini + Densidade + Gasto|MI +
Renda + Gini + AE + Lixo + Densidade + PP + PR + PU + Gasto)
summary(modelodisp1)
modelodisp1final=betareg(Taxa ~ Renda + Gini + Densidade + Gasto| PP +
PR)
summary(modelodisp1final)
# Teste Reset
lrtest(modelodisp1final,.~.+I(predict(modelodisp1final,type="link")^2))

## Análise de diagnóstico
par(mfrow=c(1,2))
res=residuals(modelodisp1final)
plot(res, main="(a)",xlab = "Observações", ylab = "Resíduos")
abline(h = c(-3,3), lty = 2)
```

```
identify(res,,n=1)
plot(predict(modelodisp1final, type = "link"),
residuals(modelodisp1final),
main="(b)", xlab = "Preditor linear", ylab = "Resíduos")
abline(h = 0, lty = 2)
cook = cooks.distance(modelodisp1final)
plot(cook, main= "(a)",xlab = "Observações", ylab = "Distância de Cook")
identify(cook,,n = 1)
abline(h=df(8,223-8,0.05))
gl = gleverage(modelodisp1final)
plot(fitted(modelodisp3),gl, main = "(b)", xlab = "Valores ajustados",
ylab = "Alavanca generalizada")
identify(fitted(modelodisp3),gl, n = 2)
abline(h=3*(8/223))
qqPlot(res, xlab = "Quantis N(0,1)", ylab = "Resíduos")

######### Gráfico da variável gasto com assistencialismo ##########
set.seed(2)
bbolsa = seq(from=85, to=600, length=600)
#bbolsa = seq(from=85, to=550, length=250)

eta = modelofinal$coef$mean[1]+modelofinal$coef$mean[2]*quantile
(Renda, 0.25)+modelofinal$coef$mean[3]*median(Gini) +
modelofinal$coef$mean[4]*median(Densidade)+modelofinal$coef$mean[5]
*bbolsa

tQ14 = (exp(eta)/(1+exp(eta))^2)*(modelofinal$coef$mean[5])
eta = modelofinal$coef$mean[1]+modelofinal$coef$mean[2]*quantile
(Renda, 0.5)+ modelofinal$coef$mean[3]*median(Gini)
+modelofinal$coef$mean[4]*median(Densidade)+modelofinal$coef$mean[5]
*bbolsa
tQ24 = (exp(eta)/(1+exp(eta))^2)*(modelofinal$coef$mean[5])
eta = modelofinal$coef$mean[1]+modelofinal$coef$mean[2]*quantile
(Renda, 0.75)+modelofinal$coef$mean[3]*median(Gini)
```

```
+modelofinal$coef$mean[4]*median(Densidade)+modelofinal$coef$mean[5]
*bbolsa
tQ34 = (exp(eta)/(1+exp(eta))^2)*(modelofinal$coef$mean[5])
summary(tQ14)
summary(tQ24)
summary(tQ34)
plot(bbolsa,tQ14, type="l",lty=2, ylab="Impacto estimado",
xlab="Gasto com assistencialismo per capita", main="",
ylim=c(0.0004, 0.0007),xlim=c(80, 600))
lines(bbolsa,tQ24,type="l",lty=1, lwd=2)
lines(bbolsa,tQ34,type="l",lty=3, lwd=2)
legend("topleft", legend=c("1º quartil","2º quartil","3º quartil"),
lty=c(2,1,3),lwd=2, bty="n")
####### Ajuste do modelo e gráfico da variável gasto com
assistencialismo suavizada - Modelo GAMLSS ########
modelogamlss=gamlss(Taxa ~ Renda + Gini + Densidade + cs(Gasto,df=3),
 family = "BE")
summary(modelogamlss)
term.plot(modelogamlss, what = "mu", col.se = "blue", ylim = c(-1.0,0.5),
main="",xlab=c("Renda","Gini","Densidade",
"Gasto com assistencialismo per capita"),ylab=c("Renda","Gini",
"Densidade","cs(Taxa,df=3)"))
abline(h=0,lty=1)
```