UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA

André Antonio de Oliveira

Inferências e aplicações no modelo de regressão beta com dispersão variável

João Pessoa, 24 de Maio de 2017

André Antonio de Oliveira

Inferências e aplicações no modelo de regressão beta com dispersão variável

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

Orientadora: Prof[a]. Dra. Tatiene Correia de Souza

João Pessoa, 24 de Maio de 2017

André Antonio de Oliveira

Inferências e aplicações no modelo de regressão beta com dispersão variável

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

Aprovado em 24 de Maio de 2017.

BANCA EXAMINADORA

Profª. Dra. Tatiene Correia de Souza - Orientadora
UFPB

Profª. Dra. Maria Lídia Coco Terra
UFPB

Profº Dr. Marcelo Rodrigo Portela Ferreira
UFPB

Dedico este trabalho...

A Deus, por me dar forças e coragem para enfrentar os desafios que surgem a cada dia. Aos meus pais, Marques e Cícera. A minha orientadora, Tatiene. Aos meus amigos e familiares por fazerem parte direta ou indiretamente da minha jornada. Dedico este trabalho com muito carinho a todos vocês.

Agradecimentos

Agradeço primeiramente a Deus, por iluminar o meu caminho, me dar forças e coragem para enfrentar todos os desafios que surgem a cada dia e por estar sempre comigo.

Aos meus pais, Marques e Cícera, por todo o apoio que me deram e têm me dado, seja emocional ou financeiro, por todo o carinho, dedicação, e paciência que tiveram na minha criação. Agradeço por todos os conselhos e por estarem comigo em todos os momentos e sempre me apoiarem em todas as minhas decisões. Sou grato por tudo a vocês.

Aos meus irmãos, Sara, Amanda, Andrea e Marquinho, por mesmo estando longe, se fazerem presentes no meu cotidiano.

Aos meus avós (*in memorian*), Dona Zizi, Francisca e Antonio, que mesmo não estando mais entre nós, tiveram um papel fundamental na criação dos meus pais e na minha também.

A toda minha família, meus tios, primos, cunhados, sobrinhos e todos aqueles que direta ou indiretamente estão envolvidos em minha vida.

A minha orientadora, Tatiene, por ter me orientado nestes últimos anos, por ter acreditado no meu potencial desde o início, pela amizade, por todos os conselhos dados, pela paciência e disponibilidade, por todos os ensinamentos e por sempre querer o melhor para mim, independente da situação. Obrigado por sempre me incentivar e por ter acompanhado de perto meu crescimento pessoal e profissional.

A professora Ana Flávia, por ser uma pessoa muito especial, pela amizade, pelos conselhos, pelo cuidado comigo nas viagens com a turma, pelo acolhimento, pelos sábados de *Just Dance*, pelos momentos sérios e de descontração.

A todos os professores do DE-UFPB, em especial, Tarciana, Luiz, Hemílio, Marcelo, João Agnaldo, Maria Lídia, Jozemar, Neir, Renata, Ronei, Ulisses, Rodrigo e Izabel, por todos os ensinamentos ao longo das disciplinas do curso e por sempre estarem disponíveis para tirar minhas dúvidas.

Aos meus amigos e companheiros de curso, Adenice, Diogo, Anny, José e Lukas, pessoas muito especiais que conheci a quatro anos atrás e que tiveram um papel fundamental durante a minha graduação, obrigado pela amizade, pela paciência, por terem estado comigo nos momentos felizes e de descontração, e também nos momentos de estresse e de desespero. Obrigado por terem me ensinado que uma família também pode ser formada por laços de amizade.

Aos amigos que conheci ao longo do curso, em especial, Clarissa, Danilo, Roberto, Matheus, Kelfânio, Lígia, Diego, Saul e Ianne, pessoas que também tiveram um papel fundamental na minha graduação e que passaram a fazer parte desta família estatística ao longo do tempo.

Aos professores do CNEC, por terem me dado uma boa base, em especial ao professor Édino, por ter sido um ótimo professor de matemática.

A Josélia por ter me acompanhado de perto desde os tempos da pré-escola até o vestibular. A todos os funcionários do CNEC por cada um ter exercido seu papel com excelência e terem garantido o bom funcionamento da escola ao longo dos anos que estive lá.

Aos meus amigos de escola, que mesmo não os vendo com frequência ainda me garantem bons momentos de descontração quando nos encontramos.

A natureza, por ser perfeita e por nos ensinar coisas novas a cada dia.

Ao CNPq pelo apoio financeiro nos projetos aqui apresentados.

Aos participantes da banca examinadora pelas sugestões.

"Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes."

(Martin Luther King Jr.)

Resumo

A análise de regressão é uma das técnicas estatísticas mais utilizadas para investigar o comportamento de uma variável resposta quando o mesmo é influenciado por um conjunto de outras variáveis. Modelos de regressão beta são adequados para os casos em que a variável resposta está restrita ao intervalo $(0, 1)$, a exemplo de taxas e proporções. O modelo de regressão beta com dispersão variável, que é o foco deste trabalho, é um tanto geral e contêm duas estruturas de regressão, a saber: para a média e dispersão/precisão. Estas estruturas de regressão contêm covariáveis, parâmetros desconhecidos e funções de ligação que permitem a modelagem dos parâmetros de interesse. Neste trabalho abordamos tanto a aplicação prática quanto teórica do modelo de regressão beta com dispersão variável. Em relação a abordagem prática, um dos nossos primeiros objetivos foi de avaliar e explicar a proporção de crianças obesas beneficiadas pelo Programa Bolsa Família nas regiões do Brasil através de variáveis relacionadas as condições sociais, econômicas, demográficas e nutricionais dos beneficiários do Bolsa Família nos municípios brasileiros. O principal resultado encontrado foi que em duas das cinco regiões do Brasil, a variável gasto per capita com o Bolsa Família foi significativa e apresentou influência positiva na obesidade, ou seja, nos municípios destas duas regiões que mais receberam benefícios do Bolsa Família houve uma tendência a apresentar uma maior proporção de crianças obesas, resultado este bastante preocupante e que exige atenção. Ainda abordando o tema obesidade, um segundo objetivo nosso foi de avaliar a proporção de adultos obesos nos Estados Unidos. Para isto, consideramos algumas variáveis relacionadas as condições sociais e nutricionais da população de 50 estados daquele país. Os resultados revelaram que os estados com maiores porcentagens de inatividade física tenderam a apresentar uma maior proporção de indivíduos obesos, enquanto que os estados com maior escore de bem-estar tenderam a ter uma menor proporção. Adicionalmente, estimamos o impacto da inatividade física sobre a proporção média de adultos obesos e os resultados revelaram que o efeito deste impacto apresenta forma positiva para todos os possíveis valores de inatividade física. Por fim, um dos nossos últimos objetivos, agora abordando o enfoque teórico, foi de avaliar os efeitos de erros de especificação nas inferências do modelo de regressão beta com dispersão variável. Para a avaliação destes erros realizamos um estudo de simulação considerando diferentes cenários. Os resultados destas simulações revelaram que os erros de especificação envolvendo a estrutura de regressão do parâmetro de precisão apresentaram uma influência considerável nas inferências do modelo, indicando uma necessidade de maior atenção na modelagem desta estrutura.

Palavras-chave: Modelo de regressão beta; obesidade; erros de especificação; dispersão variável.

Abstract

Regression analysis is one of the statistical techniques widely used to investigate the behavior of a response variable when it is influenced by a set of other variables. Beta regression models are suitable for cases where the response variable is restricted to the interval $(0, 1)$, such as rates and proportions. The beta regression model with varying dispersion, which is the focus of this work, is somewhat general and contains two regression structures, namely: for the mean and dispersion/precision. These regression structures contain covariates, unknown parameters and link functions that allow the modeling of the parameters of interest. In this work we approach both the practical and theoretical application of the beta regression model with varying dispersion. Regarding the practical approach, one of our first objectives was to evaluate and explain the proportion of obese children benefited by the Bolsa Família Program in the regions of Brazil through variables related to the social, economic, demographic and nutritional conditions of Bolsa Família recipients in the municipalities of Brazil. The main result was that in two of the five regions of Brazil, the variable per capita spending with Bolsa Família was significant and had a positive influence on obesity, that is, in the municipalities of the two regions that received the most Bolsa Família benefits, there was a tendency to present a higher proportion of obese children, a result that is very worrying and requires attention. Still addressing the obesity topic, a second objective was to evaluate the proportion of obese adults in the United States. For this, we consider some variables related to the social and nutritional conditions of the population of 50 states in that country. The results showed that the states with the highest percentages of physical inactivity tended to present a higher proportion of obese individuals, whereas states with higher values of well-being score tended to have a lower proportion. Additionally, we estimated the impact of physical inactivity on the average proportion of obese adults and the results showed that the effect of this impact is positive for all possible values of physical inactivity. Finally, one of our last objectives, now considering the theoretical approach, was to evaluate the effects of specification errors on the inferences of the beta regression model with varying dispersion. For the evaluation of these errors we performed a simulation study considering different scenarios. The results of these simulations showed that the specification errors involving the regression structure of the precision parameter had a considerable influence on the model inferences, indicating a need for more attention in the modeling of this structure.

Keywords: Beta regression model; obesity; misspecification, varying dispersion.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A análise de regressão é uma técnica estatística adequada para avaliar a relação de dependência de uma variável, denominada variável dependente ou resposta, em relação a uma ou mais variáveis, denominadas variáveis independentes ou explicativas. Esta técnica permite estimar e/ou prever o valor médio da variável resposta em termos de valores conhecidos ou fixados das variáveis explicativas, sendo possível a realização de inferências a respeito do fenômeno em estudo. Caso o modelo escolhido não seja adequado, inferências imprecisas, e consequentemente conclusões equivocadas, podem vir a ocorrer.

Um dos modelos mais utilizados em diversas análises é o modelo de regressão normal linear, porém o mesmo torna-se inapropriado quando a variável resposta está restrita ao intervalo unitário $(0, 1)$, a exemplo de taxas e proporções contínuas. Dados desta natureza não se distribuem sobre todos os reais, que é o domínio da distribuição normal, e geralmente apresentam assimetria, fazendo com que inferências baseadas na suposição de normalidade possam estar incorretas. Além disso, o uso desta técnica de regressão para modelar taxas e proporções pode fazer com que o modelo ajustado gere valores preditos para a variável resposta que excedam os limites de seu intervalo. Baseado nisso, uma das práticas mais utilizadas é a transformação da variável dependente para que a mesma possa assumir valores na reta real, e posterior modelagem de sua média através do modelo normal linear. Porém, essa solução apresenta alguns intervenientes, como o fato de que os parâmetros do modelo podem não ser facilmente interpretados em termos da variável original, e que algumas suposições ainda podem estar sendo violadas.

Kieschnick e McCullough (2003) analisaram a aplicação de sete diferentes tipos de modelos de regressão que usualmente são utilizados para analisar taxas e proporções. Os modelos considerados no estudo foram o modelo normal linear, o modelo normal não-linear, o modelo normal censurado (modelo Tobit), o modelo logito e os modelos que usam a distribuição beta, simplex e de quasi-verossimilhança. O objetivo principal do estudo foi de analisar o comportamento de duas bases de dados, cuja variável resposta estava restrita ao intervalo $(0, 1)$, ao efeito da violação das suposições de cada modelo, além comparar as estimativas e inferências obtidas para então se determinar qual dos modelos melhor descreveria os dados considerados. Os autores chegaram a conclusão de que modelos baseados na distribuição beta e de quasi-verossimilhança se mostraram superiores nas análises realizadas, com a ressalva de que os modelos baseados na distribuição beta seriam mais recomendados caso o tamanho amostral não justificasse a aproximação por quasi-verossimilhança.

Smithson e Verkuilen (2006) fizeram comparações entre modelos baseados na distribuição beta e técnicas alternativas (como o uso de modelos baseados na distribuição normal) para se modelar variáveis que se distribuem em um intervalo limitado na reta. Segundo os autores, modelar variáveis limitadas é particularmente difícil, principalmente quando os valores da variável estão concentrados próximos a um dos limites do intervalo. Além disto, muitos pesquisadores acreditam que as suposições do modelo normal linear são robustas a violações, visão esta que pode levar a conclusões equivocadas à respeito do fenômeno em estudo. Os autores ainda afirmam que a abordagem considerando a distribuição beta é uma alternativa eficaz a distribuição normal, particularmente para variáveis que são limitadas acima e abaixo. De acordo com Cribari-Neto e Zeleis (2010), a principal motivação para o uso de modelos de regressão baseados na distribuição beta para modelar variáveis limitadas está na flexibilidade da distribuição. A densidade beta pode assumir diferentes formas, dependendo apenas da combinação dos valores dos parâmetros que a indexam.

Baseado nestes fatos, o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) é uma abordagem eficaz para se modelar variáveis que se distribuem no intervalo unitário $(0, 1)$. Este modelo de regressão assume que a variável resposta possui distribuição beta e que sua média é relacionada a um preditor linear por meio de uma função de ligação, preditor este que contém covariáveis e parâmetros de regressão desconhecidos. Este modelo também é indexado por um parâmetro de dispersão, que pode variar ao longo das observações, podendo também ser modelado em termos de covariáveis, parâmetros desconhecidos e uma função de ligação (SIMAS et al., 2010). Uma das principais vantagens do uso deste modelo de regressão é justamente a possibilidade de se modelar a variabilidade da variável resposta de maneira natural e explícita, sendo um artifício útil na identificação de fontes de variação do fenômeno em estudo.

Neste contexto, este trabalho apresenta como objetivo principal a apresentação de resultados práticos e teóricos no modelo de regressão beta com dispersão variável. Em relação ao enfoque prático direcionamos nosso objetivo a duas bases de dados. Na primeira, o nosso interesse consistiu em avaliar os fatores que influenciaram na situação nutricional de crianças beneficiárias do Programa Bolsa Família nas regiões do Brasil, já que alguns estudos revelaram que a maior parte do benefício recebido mensalmente pelas famílias beneficiárias era gasto com alimentação e que uma das principais mudanças ocorridas nos hábitos alimentares destas famílias foi o aumento no consumo de alimentos com maior densidade calórica e menor valor nutritivo, influenciando diretamente na situação nutricional das crianças pertencentes as mesmas. Em relação ao segundo enfoque prático, ainda abordando o tema obesidade, avaliamos os fatores que influenciaram na obesidade adulta nos Estados Unidos, uma vez que esse país está entre as nações desenvolvidas que mais sofrem com os problemas relacionados a este tema. No que diz respeito ao enfoque teórico, avaliamos através de um estudo de simulação os efeitos de diferentes erros de especificação nas inferências do modelo de regressão beta com dispersão variável, pois ao se realizar qualquer análise de regressão esperamos que o modelo ajustado represente adequadamente a realidade do fenômeno em estudo, porém caso a especificação do modelo escolhido esteja incorreta, inferências imprecisas podem vir a ocorrer.

O presente trabalho encontra-se organizado em cinco capítulos. Os Capítulos 2, 3, e 4 são auto-suficientes, ou seja, cada capítulo pode ser lido separadamente, em qualquer ordem. Desta forma, algumas notações e resultados são apresentados mais de uma vez. No Capítulo 2, avaliamos e explicamos a proporção de crianças obesas, entre 0 e 5 anos de idade, beneficiadas pelo Programa Bolsa Família no ano de 2014, e identificamos para cada uma das 5 regiões do Brasil os fatores que influenciaram na obesidade desses indivíduos através de variáveis relacionadas as condições sociais, econômicas, demográficas e nutricionais dos beneficiários do Programa Bolsa Família nos municípios brasileiros.

No Capítulo 3 modelamos a proporção de adultos obesos nos Estados Unidos, para o ano de 2014, considerando indivíduos que apresentaram IMC (Índice de Massa Corporal) maior ou igual a 30.0 kg/m^2. Para isto, consideramos algumas covariáveis, tais como porcentagem de adultos fisicamente inativos, porcentagem de adultos que consumiam vegetais menos de uma vez ao dia, porcentagem de fumantes de cigarro, porcentagem de residentes desempregados ou empregados em tempo parcial, taxa de insegurança alimentar, escore de bem-estar e porcentagem de residentes que não tinham cobertura do seguro de saúde. Vale salientar que toda a análise estatística foi baseada em 50 observações referentes aos estados dos Estados Unidos. Adicionalmente, estimamos o impacto da inatividade física sobre a proporção média de adultos obesos.

No Capítulo 4, avaliamos os efeitos de alguns erros de especificação nas inferências do modelo de regressão beta com dispersão variável. Para isto, um estudo de simulação foi realizado. Neste estudo, a variável resposta foi gerada com distribuição beta assumindo covariáveis e funções de ligação conhecidas, em sequência o modelo foi ajustado sob a especificação correta e incorreta considerando seis tipos de erros de especificação. Avaliamos os efeitos destes erros através de taxas de rejeição e taxas de cobertura em relação a um dos parâmetros do submodelo da média e, além disso, avaliamos também o viés relativo e o erro quadrático médio em relação às estimativas das respostas médias. Realizamos ainda uma aplicação a dados reais com o objetivo de verificar na prática os efeitos de diferentes formas de especificação nas inferências do modelo de regressão beta com dispersão variável.

Por fim, no último capítulo apresentamos as conclusões, considerações finais e possíveis direcionamentos para trabalhos futuros. Vale salientar que todas as análises aqui realizadas foram feitas utilizando o *software* estatístico `R` (R DEVELOPMENT CORE TEAM, 2014), sendo que os *scripts* utilizados nestas análises encontram-se no Apêndice deste trabalho.

Capítulo 2

Avaliação da proporção de crianças obesas beneficiadas pelo Programa Bolsa Família nas regiões do Brasil

2.1 Introdução

A obesidade é um dos problemas de saúde publica cada vez mais comuns nos dias atuais e que afeta toda a população independente de sexo, faixa etária ou classe social. O Ministério da Saúde define a obesidade como sendo uma doença crônica caracterizada pelo excesso de gordura corporal, que causa prejuízos à saúde do indivíduo. Este problema pode estar relacionado principalmente com a alimentação, falta de exercícios físicos ou fatores genéticos e que pode desencadear outras doenças mais graves como diabetes e doenças cardiovasculares (BRASIL, 2015).

Essa questão têm se tornado ainda mais preocupante por atingir crianças e adolescentes em grande escala. Segundo a Organização Mundial de Saúde (OMS), a obesidade infantil é um problema global que está atingindo muitos países de baixa e média renda, sendo que em 2013, o número de crianças com excesso de peso com idade inferior a cinco anos estava estimado em mais de 42 milhões, sendo que 31 milhões vivendo em países em desenvolvimento (WORLD HEALTH ORGANIZATION, 2015).

No Brasil, segundo dados divulgados pela Secretaria de Direitos Humanos da Presidência da República em um estudo sobre a alimentação adequada de crianças e adolescentes, apenas 1.9% das crianças com menos de 5 anos de idade apresentaram baixo peso. Em contrapartida, constatou-se 7.3% das crianças nessa faixa etária com excesso de peso, resultados esses referentes ao ano de 2006. Além do mais, segundo esse mesmo estudo, o estado nutricional na primeira infância repercute na vida adulta, e a incidência de obesidade em adultos têm crescido nos últimos anos em todas as regiões brasileiras (BRASIL, 2015).

Nesse contexto, diversos estudos buscaram um maior aprofundamento no tema da obesidade infantil. Abrantes et al. (2002) realizaram um estudo sobre a prevalência de sobrepeso e obesidade em crianças e adolescentes das Regiões Nordeste e Sudeste, e concluíram que a prevalência de obesidade foi maior em crianças do que em adolescentes. Além do mais, a Região Sudeste apresentou uma maior prevalência de crianças obesas comparada à Região Nordeste. Moreira et al. (2012) objetivaram identificar a prevalência de excesso de peso e sua associação com fatores econômicos, biológicos e maternos em me-

nores de 5 anos da região semiárida do estado de Alagoas. Contudo, não foi evidenciada associação significativa entre o excesso de peso e as variáveis socioeconômicas estudadas.

Oliveira et al. (2011) afirmaram que o governo brasileiro vem implantando programas de transferência de renda, como o Programa Bolsa Família, partindo do fato de que um incremento na renda familiar pode promover uma melhora no estado nutricional das crianças que nelas vivem. Segundo Segall-Corrêa et al. (2008), as políticas de transferência de renda podem exercer um papel fundamental na melhoria das condições sociais da população, principalmente entre aquelas pessoas que se encontram em situação de extrema pobreza.

Neste cenário, um dos principais programas de transferência de renda é o Programa Bolsa Família. Criado em 2003 no governo do então presidente Lula, o Bolsa Família beneficia famílias em situação de pobreza e extrema pobreza em todo o país e integra o Plano Brasil sem Miséria, que tem como foco de atuação os milhões de brasileiros com renda familiar per capita inferior a $R\$$ 77.00 mensais e está baseado na garantia de renda, inclusão produtiva e no acesso aos serviços públicos (BRASIL, 2015). O Programa Bolsa Família é reconhecido internacionalmente como o maior programa de transferência de renda do mundo.

Segundo uma pesquisa realizada em 2008 pelo Instituto Brasileiro de Análises Sociais e Econômicas, IBASE, com titulares do Cartão Bolsa Família, a maior parte do beneficio recebido mensalmente era gasto principalmente com alimentação, e que uma das principais mudanças ocorridas nos hábitos alimentares após o recebimento deste auxílio foi o aumento no consumo de açúcares, 78% dos titulares disseram que passaram a comprar mais desse grupo alimentar. A pesquisa também concluiu que no geral, o que prevalece na decisão de consumo da dieta das famílias são os alimentos de maior densidade calórica e menor valor nutritivo, contribuindo para o aumento da prevalência de excesso de peso e obesidade (IBASE, 2008).

Diversos autores buscaram associar o Programa Bolsa Família à situação nutricional dos beneficiários. Lima et al. (2011) apresentaram como objetivo em seu estudo a verificação do estado nutricional da população adulta beneficiária do Programa Bolsa Família no município de Curitiba, no estado do Paraná, e observaram uma prevalência de sobrepeso e obesidade em 56% dessa população. Cabral et al. (2013) estudaram beneficiários desse mesmo programa em Maceió, no estado de Alagoas, e encontraram alta prevalência de desnutrição em crianças e adolescentes, mas excesso de peso em adultos. Acreditou-se no estudo que o excesso de peso tenha sido influenciado pelo aumento no consumo de alimentos com alta densidade energética, devido a renda extra proveniente do benefício.

Saldiva et al. (2010) avaliaram as condições de saúde e nutrição de crianças menores de cinco anos e associaram a qualidade do consumo alimentar aos beneficiários do Bolsa Família de um município do semiárido brasileiro. Um aspecto relevante identificado no estudo foi o consumo excessivo de guloseimas associado positivamente com crianças que pertenciam às famílias beneficiárias. Baseado neste resultado, foi formulada a hipótese de que com o aumento da renda mensal, as famílias passaram a consumir mais alimentos com baixo valor nutricional. Cotta e Machado (2013), Monteiro et al. (2014), Silva (2011) e Wolf e Filho (2014) também apresentaram como objetivo em seus estudos a avaliação

da situação nutricional de beneficiários do Programa Bolsa Família.

Baseado nestes fatos, o nosso objetivo neste capítulo é avaliar e explicar a proporção de crianças obesas, entre 0 e 5 anos de idade, beneficiadas pelo Programa Bolsa Família no ano de 2014, e identificar para cada uma das cinco regiões do Brasil os fatores que influenciam na obesidade desses indivíduos através de variáveis relacionadas às condições sociais, econômicas, demográficas e nutricionais dos beneficiários do Bolsa Família nos municípios brasileiros. Como a variável de interesse é uma proporção, é necessário o uso de modelos apropriados para essas situações. Para isso, utilizamos o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004). 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.

O presente capítulo encontra-se dividido em cinco seções. A Seção 2 apresenta o modelo de regressão beta. Uma breve descrição dos dados encontra-se na Seção 3. Na Seção 4 são apresentados os modelos de regressão beta ajustados considerando as cinco regiões brasileiras. Por último, na Seção 5 encontram-se as conclusões e considerações finais do capítulo.

2.2 Modelo de regressão beta

A classe de modelos de regressão beta é comumente utilizada em modelagens de variáveis que assumem valores no intervalo unitário $(0, 1)$. Ferrari e Cribari-Neto (2004), Kieschnick e McCullough (2003), Ospina et al. (2006) e Paolino (2001) utilizaram modelos de regressão para situações em que a variável resposta segue distribuição beta. Em tais modelos, assume-se que a resposta média é relacionada com um preditor linear por meio de uma função de ligação. O preditor linear envolve covariáveis e parâmetros de regressão desconhecidos. Estes modelos também são indexados por um parâmetro de dispersão, que em certas situações pode variar ao longo das observações (ALMEIDA JUNIOR; SOUZA, 2015; CRIBARI-NETO; SOUZA, 2012, 2013; ESPINHEIRA et al., 2008a, 2008b; SILVA; SOUZA, 2014; SIMAS et al., 2010; SMITHSON; VERKUILEN, 2006).

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

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\ 0 < y < 1, \tag{2.1}$$

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

Sejam $y_1, ..., y_n$ variáveis aleatórias independentes, em que cada y_t, $t = 1, ..., n$, segue a densidade apresentada em (2.1) com média μ_t e parâmetro de precisão ϕ_t sendo desconhecidos, o modelo de regressão beta (FERRARI; CRIBARI-NETO, 2004) assume que a média satisfaz a seguinte relação funcional

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t, \tag{2.2}$$

em que $\beta = (\beta_1, ..., \beta_k)^\top$ é um vetor de parâmetros de regressão desconhecidos ($\beta \in \mathbb{R}^k$), $x_{t1}, ..., x_{tk}$ são observações de k covariáveis, η_t 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 função de ligação. Portanto $\mu_t = g^{-1}(\eta_t)$ e $\text{var}(y_t) = \mu_t(1 - \mu_t)/(1 + \phi)$. Além disso, podemos considerar ainda que o parâmetro de precisão ϕ_t varia ao longo das observações (SIMAS et al., 2010). Deste modo, podemos admitir que a estrutura de regressão para o parâmetro de precisão é dada por

$$h(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \vartheta_j, \tag{2.3}$$

em que $\gamma = (\gamma_1, ..., \gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1}, ..., z_{tq}$ são observações de q covariáveis ($k + q < n$) assumidas fixas e conhecidas e $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável que mapeia os pontos positivos da reta. Há várias possíveis escolhas para as funções de ligação $g(\cdot)$ e $h(\cdot)$. Entre elas podemos utilizar a função de ligação logit

$$g(\mu) = \log\Big(\frac{\mu}{1 - \mu}\Big), \tag{2.4}$$

ou a função probit

$$g(\mu) = \Phi^{-1}(\mu), \tag{2.5}$$

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

Segue de (2.1) que o logaritmo da função de verossimilhança é

$$\ell(\beta, \gamma) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t), \tag{2.6}$$

em que

$$\begin{aligned} \ell_t(\mu_t, \phi_t) &= \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1 - \mu_t)\phi_t) + (\mu_t\phi_t - 1)\log y_t \\ &+ \{(1 - \mu_t)\phi_t - 1\}\log(1 - y_t). \end{aligned} \tag{2.7}$$

Como os estimadores de máxima verossimilhança de β e γ não possuem forma fechada, eles precisam ser obtidos numericamente maximizando a função de log-verossimilhança através de um algoritmo de maximização não-linear. Usualmente, utiliza-se o método quasi-Newton BFGS (PRESS et al., 1992). Para maiores detalhes inferenciais e expressões matriciais do vetor escore e da matriz de informação de Fisher, ver Simas et al. (2010).

Sob certas 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:

$$\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\beta\\ \gamma\end{pmatrix}, K^{-1}\right), \tag{2.8}$$

aproximadamente, sendo $\hat{\beta}$ e $\hat{\gamma}$ os estimadores de máxima verossimilhança de β e γ, respectivamente, e K^{-1} é a inversa da matriz de informação de Fisher.

2.3 Descrição dos dados

A descrição das variáveis utilizadas nesse estudo está apresentada na Tabela 2.1. As fontes de dados consultadas foram o Atlas do Desenvolvimento Humano no Brasil, o Ministério do Desenvolvimento Social e Combate à Fome (MDS), o Sistema de Vigilância Alimentar e Nutricional (SISVAN), o Departamento de Informática do Sistema Único de Saúde (DATASUS) e o Instituto Brasileiro de Geografia e Estatística (IBGE). Vale salientar que através de uma extensa pesquisa bibliográfica, selecionamos para esse estudo algumas das variáveis disponíveis nestas fontes de dados que poderiam ter relação com a variável resposta e com o contexto proposto.

Tabela 2.1: Descrição das variáveis utilizadas.

Variáveis	**Definição**
$OB2014$	Proporção de crianças beneficiadas pelo Programa Bolsa Família entre 0 e 5 anos de idade com obesidade em 2014
$OB2010$	Proporção de crianças beneficiadas pelo Programa Bolsa Família entre 0 e 5 anos de idade com obesidade em 2010
$SB2010$	Proporção de crianças beneficiadas pelo Programa Bolsa Família entre 0 e 5 anos de idade com sobrepeso em 2010
$GPER$	Gasto com assistencialismo per capita em 2014 (Programa Bolsa Família)
$RENDA$	Renda per capita em 2010
$POBRES$	Percentual de pobres em 2010
IDH	Índice de Desenvolvimento Humano em 2010
$TDES$	Taxa de desemprego em 2010
$GINI$	Índice de Gini em 2010
MI	Taxa de Mortalidade Infantil em 2010
$ANBT$	Taxa de Analfabetismo entre pessoas acima de 15 anos de idade em 2010
PIB	Produto Interno Bruto per capita em 2011
POP	População estimada do município em 2014
DU	Variável *dummy*: 1 se $POP > 50000$ habitantes, 0 caso contrário

Na Tabela 2.2, encontram-se algumas estatísticas descritivas como mínimo, primeiro quartil ($Q_{1/4}$), mediana, média, terceiro quartil ($Q_{3/4}$), máximo e coeficiente de variação

(CV) das variáveis. Essas estatísticas são baseadas em 4957 observações referentes a uma amostra dos 5571 municípios brasileiros. Com base nas suas análises algumas conclusões podem ser feitas a respeito das variáveis. Em relação a variável obesidade no ano de 2014, o valor máximo encontrado foi de 0.9074, ou seja, no município correspondente a esse valor cerca de 91% das crianças que receberam o benefício do Programa Bolsa Família apresentaram obesidade, enquanto que 75% dos municípios obtiveram valores para essa mesma variável inferiores ou iguais a 0.1111. Já considerando o ano de 2010, 25% dos municípios obtiveram um percentual de obesos menor ou igual a 5.36%. Em relação ao sobrepeso, o valor máximo encontrado foi de 0.4462. Vale destacar que estas variáveis são referentes as informações das crianças acompanhadas pelo Sistema de Gestão do Bolsa Família.

Tabela 2.2: Estatísticas descritivas das variáveis utilizadas.

Variáveis	**Mínimo**	$\mathbf{Q}_{1/4}$	**Mediana**	**Média**	$\mathbf{Q}_{3/4}$	**Máximo**	**CV(%)**
$OB2014$	0.0045	0.0588	0.0818	0.0919	0.1111	0.9074	61.64
$OB2010$	0.0039	0.0536	0.0779	0.0897	0.1087	0.8994	66.99
$SB2010$	0.0041	0.0629	0.0806	0.0853	0.1013	0.4462	42.17
$GPER$	2.30	65.75	132.30	155.80	234.70	2839.00	70.69
$RENDA$	96.25	274.20	441.94	476.24	630.12	2043.74	49.18
$POBRES$	0.42	7.82	20.57	24.43	39.62	78.59	73.23
IDH	0.4200	0.6000	0.6600	0.6550	0.7100	0.8600	10.93
$TDES$	0.10	4.10	6.05	6.59	8.34	39.15	55.06
$GINI$	0.32	0.46	0.50	0.49	0.54	0.80	12.76
MI	8.49	14.10	17.40	19.65	24.30	46.80	36.48
$ANBT$	1.04	8.49	13.98	16.68	24.80	44.40	58.80
PIB	2462.00	5594.00	10208.00	13765.00	16622.00	387137.00	116.43
POP	1000	6207	12702	36773	26640	6453682	401.46

A variável gasto per capita com o Programa Bolsa Família no ano de 2014 foi obtida dividindo-se os gastos, em reais, com o referido programa assistencialista pela população estimada dos municípios no ano considerado. Com isso, temos que 50% dos municípios apresentaram um gasto per capita com o referido programa assistencialista menor ou igual a $R\$132.30$, enquanto que o valor máximo encontrado foi de $R\$2839.00$. Considerando o percentual de pobres, temos que os valores de mínimo e máximo foram 0.42% e 78.59%, respectivamente.

O Índice de Desenvolvimento Humano (IDH) médio encontrado foi de 0.6550, enquanto que o coeficiente de variação foi igual a 10.93%, indicando uma baixa dispersão da variável. No caso do Índice de Gini ($GINI$), que de acordo com a fonte consultada (`http://atlasbrasil.org.br/2013/pt/`) mede o grau de desigualdade na distribuição de indivíduos segundo a renda domiciliar per capita, o valor médio encontrado foi de 0.49, sendo que o valor 0 corresponde a quando não há desigualdade (a renda domiciliar per capita de todos os indivíduos tem o mesmo valor) e o índice tende a 1 a medida que essa desigualdade aumenta. Já em relação a taxa de desemprego, o valor médio encontrado foi de 6.59 com coeficiente de variação igual a 55.06%, indicando uma alta dispersão da variável. Para a variável mortalidade infantil, que considera o número de crianças que não deverão sobreviver ao primeiro ano de vida em cada 1000 crianças nascidas vivas, temos que 75% dos municípios apresentaram um valor menor ou igual a 24.30, ou seja,

a cada 1000 crianças nascidas vivas, cerca de 25 não irão sobreviver ao primeiro ano de vida. O valor mínimo para o Produto Interno Bruto per capita corresponde a $R\$2462.00$. Enquanto que 75% das observações apresentaram uma taxa de analfabetismo menor ou igual a 24.80.

Considerando a variável população, temos que o valor médio foi de 36773 habitantes. Com o objetivo de complementar a análise sobre a influência do número de habitantes no estudo, foi criada a variável *dummy*, DU, que classifica os municípios como urbanos ou rurais. Neste estudo, assim como em Souza e Cribari-Neto (2013), o município é classificado como urbano se sua população estimada no ano de 2014 exceder 50000 habitantes ($DU = 1$), e é classificado como rural caso contrário ($DU = 0$). Dos 4957 municípios considerados, 624 foram classificados como urbanos, representado um percentual de aproximadamente 12.59% do total de municípios.

Destacamos que a maior proporção de crianças obesas no ano de 2014 foi encontrada no Estado do Maranhão, no município de Paulo Ramos, enquanto que em Tumiritinga (Minas Gerais) foi encontrada a menor proporção. Já em relação ao sobrepeso infantil, a cidade de Calçado (Pernambuco) apresentou o maior valor. Para a variável gasto per capita, Campo dos Goytacazes e Carapebus, ambos no Estado do Rio de Janeiro, tiveram o maior e o menor gasto per capita com o Programa Bolsa Família, respectivamente.

O município de Fernando Prestes (São Paulo) registrou o menor percentual de pobres para o ano de 2010, enquanto que Uiramutã, em Roraima, registrou o maior percentual para essa mesma variável. Santa Terezinha, no Estado de Santa Catarina, e Campo Alegre do Fidalgo, no Piauí, registraram os valores de mínimo e máximo para a variável taxa de desemprego, respectivamente. O menor Produto Interno Bruto per capita encontrado, no ano de 2011, corresponde a cidade de Curralinho (Pará), enquanto que a maior taxa de analfabetismo para 2010, corresponde a Alagoinha do Piauí (Piauí).

2.4 Especificação dos modelos

Nesta seção apresentamos as modelagens referentes à proporção de crianças obesas beneficiadas pelo Programa Bolsa Família para o ano de 2014 nas cinco regiões do Brasil. Aqui o interesse consiste em explicar a proporção de crianças obesas por região e, para isso, utilizamos o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) que é adequado para os casos em que a variável resposta é uma proporção, ou seja, restrita ao intervalo $(0, 1)$. O procedimento computacional foi desenvolvido utilizando o pacote `betareg` (CRIBARI-NETO; ZELEIS, 2010) do *software* estatístico `R` (KLEIBER; ZELEIS, 2008; R DEVELOPMENT CORE TEAM, 2014).

Na seleção de covariáveis utilizadas para explicar a obesidade em crianças nas cinco regiões brasileiras, utilizamos os critérios de seleção de modelos AIC (AKAIKE, 1974) e BIC (SCHWARZ, 1978). Vale salientar que para as Regiões Centro-Oeste, Sudeste e Sul, utilizamos o critério de seleção de modelos AIC, enquanto que para as Regiões Norte e Nordeste, utilizamos o critério BIC. Isso se deve ao fato de que para essas duas regiões o melhor ajuste foi obtido a partir deste último critério.

Inicialmente, ao usar o modelo de regressão beta, estamos interessados em determinar se a precisão é fixa, ou seja, se há ou não estrutura de regressão para o parâmetro de precisão. Para tanto, empregamos o teste score (ESPINHEIRA, 2007) em que a hipótese nula é $H_0 : \phi_1 = ... = \phi_n = \phi$, ou seja, testamos a hipótese de que a precisão é constante, e concluímos que a mesma é variável para todos os cenários estudados, isto é, para cada uma das cinco regiões do Brasil há o ajuste considerando os parâmetros da média e da precisão. Para essas modelagens consideramos diferentes funções de ligação para as duas estruturas de regressão, e selecionamos os ajustes mais adequados para cada região.

Nas Tabelas 2.3 a 2.7, encontram-se os ajustes referentes às Regiões Sudeste, Sul, Centro-Oeste, Norte e Nordeste, respectivamente, com seus coeficientes estimados, erros-padrão e p-valores. Vale salientar que as funções de ligação utilizadas na estrutura de regressão da média foram cloglog para as Regiões Norte e Centro-Oeste e loglog para Nordeste, Sul e Sudeste, enquanto que para a estrutura de regressão da precisão verificamos que a função de ligação log foi mais adequada para os cinco ajustes.

Para os cinco modelos de regressão beta aplicamos o teste de especificação *RESET* (LIMA, 2007; RAMSEY, 1969). Esse teste considera como hipótese nula que o modelo selecionado está bem especificado versus a hipótese alternativa de que ele está mal especificado. O teste *RESET* realizado considerou o preditor linear estimado elevado a terceira potência como variável de teste. Para os cinco modelos considerados concluiu-se que a especificação correta dos mesmos não foi rejeitada aos níveis usuais de significância.

Em relação ao modelo de regressão beta selecionado para explicar a proporção de crianças obesas nos municípios da Região Sudeste (ver Tabela 2.3), temos que as covariáveis $SB2010$, $TDES$ e $GPER$ tiveram influência positiva na variável resposta, isto é, municípios que apresentaram uma maior proporção de crianças com sobrepeso, uma taxa de desemprego elevada e um maior gasto com assistencialismo per capita, tenderam a ter uma maior incidência de indivíduos com obesidade para o ano de 2014. Além dessas

Tabela 2.3: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão beta com dispersão variável para os dados referentes a Região Sudeste.

Modelo para μ			
Variáveis	**Estimativa**	**Erro-padrão**	***p*-valor**
$INTERCEPTO$	-1.002	2.567×10^{-2}	< 0.001
$SB2010$	8.703×10^{-1}	1.498×10^{-1}	< 0.001
$ANBT$	-3.544×10^{-3}	9.671×10^{-4}	< 0.001
$TDES$	6.046×10^{-3}	2.818×10^{-3}	0.031
$GPER$	4.321×10^{-4}	2.153×10^{-4}	0.044
$INT1$	-4.686×10^{-5}	2.130×10^{-5}	0.027
Modelo para ϕ			
Variáveis	**Estimativa**	**Erro-padrão**	***p*-valor**
$INTERCEPTO$	2.396	0.397	< 0.001
$GPER$	0.001	4×10^{-4}	0.011
DU	0.292	0.106	0.006
$GINI$	4.268	0.690	< 0.001
MI	-0.051	0.015	< 0.001

covariáveis, $ANBT$ e $INT1$, iteração entre $TDES$ e $GPER$, também foram selecionadas, sendo que a taxa de analfabetismo teve influência negativa na variável resposta, ou seja, municípios com maiores valores para esta variável tenderam a apresentar uma menor incidência de crianças obesas. O fato do gasto com assistencialismo per capita ter influenciado positivamente na obesidade pode indicar que com uma maior renda extra, as famílias passaram a consumir mais alimentos e isso pode ter influenciado diretamente no estado nutricional dos indivíduos que as compõe.

Considerando o modelo para o parâmetro de precisão, apenas a covariável MI apresentou influência negativa na precisão das respostas, enquanto que $GPER$, DU e $GINI$ apresentaram influência positiva, ou seja, o gasto com assistencialismo per capita, o fato de um município ter sido classificado como urbano e apresentado um Índice de Gini mais elevado fez com que a precisão das respostas fosse maior, isto é, as respostas tenderam a ser menos dispersas.

Com o objetivo de verificar possíveis afastamentos das suposições feitas para o modelo, a Figura 2.1 apresenta os gráficos de resíduos ponderados versus índices de observações e o gráfico de probabilidade normal com envelopes simulados. O modelo de regressão beta selecionado para explicar a proporção de crianças obesas nos municípios da Região Sudeste parece estar bem ajustado, visto que no gráfico de resíduos ponderados versus índices de observações, os resíduos permanecem dentro do intervalo $(-2, 2)$, e em geral, permanecem dentro das bandas de confiança no gráfico de probabilidade normal com envelopes simulados, ou seja, não há indícios de afastamento da suposição de que o modelo de regressão beta selecionado fornece uma boa representação para os dados.

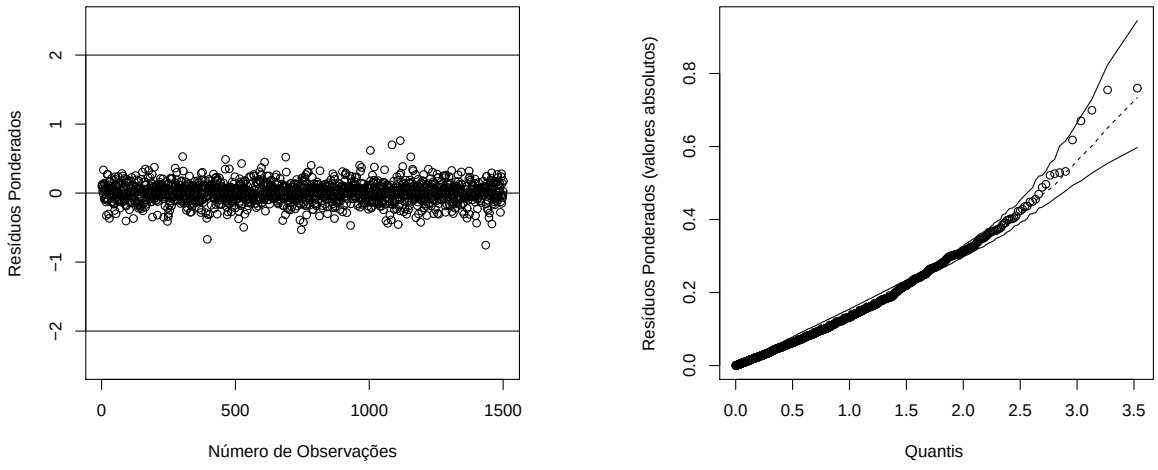


Figura 2.1: Gráfico dos resíduos quantis aleatorizados versus os índices de observações e gráfico de probabilidade normal com envelopes simulados - Região Sudeste.

Analisando os coeficientes estimados referentes ao ajuste da Região Sul apresentados na Tabela 2.4, é possível observar que as covariáveis $OB2010$, $RENDA$ e IDH foram selecionadas para compor a estrutura de regressão para a média, sendo que $OB2010$ e IDH influenciaram positivamente na proporção de obesos em 2014, isto é, municípios com uma maior quantidade de crianças obesas em 2010 e um maior IDH tenderam a apresentar

uma maior incidência de crianças obesas em 2014, enquanto municípios com uma maior renda per capita tenderam a apresentar uma menor proporção. Este último resultado pode estar relacionado ao fato de que nesta região, nos municípios que apresentaram uma maior renda per capita, seus habitantes tenderam a se alimentar de maneira mais adequada e consumir alimentos mais saudáveis, devido ao rendimento maior, reduzindo assim a incidência de obesos. Além das covariáveis citadas acima, algumas iterações também foram selecionadas, sendo que $INT1$, iteração entre as variáveis $OB2010$ e $RENDA$, e $INT2$, iteração entre $OB2010$ e IDH.

Tabela 2.4: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão beta com dispersão variável para os dados referentes a Região Sul.

Modelo para μ			
Variáveis	**Estimativa**	**Erro-padrão**	***p*-valor**
$INTERCEPTO$	-2.065	2.910×10^{-1}	< 0.001
$OB2010$	1.010×10	2.607	< 0.001
$RENDA$	-5.256×10^{-4}	1.071×10^{-4}	< 0.001
IDH	2.054	4.875×10^{-1}	< 0.001
$INT1$	5.030×10^{-3}	1.053×10^{-3}	< 0.001
$INT2$	-1.747×10	$4,425$	< 0.001
Modelo para ϕ			
Variáveis	**Estimativa**	**Erro-padrão**	***p*-valor**
$INTERCEPTO$	8.537	1.587	< 0.001
DU	1.148	0.161	< 0.001
$TDES$	0.095	0.023	< 0.001
IDH	-8.716	2.067	< 0.001
$GINI$	3.874	1.046	< 0.001
MI	-0.068	0.025	0.007
$POBRES$	-0.035	0.013	0.005

Em relação à estrutura de regressão para o parâmetro de precisão, as covariáveis DU, $TDES$ e $GINI$ influenciaram positivamente na precisão das respostas, ou seja, o fato de um município ter sido classificado como urbano, apresentado uma maior taxa de desemprego e um maior Índice de Gini fez com que ele apresentasse respostas mais precisas, ou seja, menos dispersas. Por outro lado, o IDH do município, a taxa de mortalidade infantil e o percentual de pobres, todos para o ano de 2010, exerceram efeito negativo na precisão, isto é, conforme um determinado município apresentou maiores valores para essas variáveis, a precisão das respostas diminuiu.

De acordo com a Figura 2.2, o modelo de regressão beta selecionado para explicar a obesidade em crianças na Região Sul parece estar bem ajustado, visto que os resíduos permanecem dentro do intervalo $(-2, 2)$ e, em geral, permanecem dentro das bandas de confiança dos envelopes simulados, isto é, não há indícios de afastamento da suposição de que o modelo de regressão beta selecionado fornece boa representação para os dados.

Considerando o ajuste selecionado para explicar a obesidade em crianças referente à Região Centro-Oeste (ver Tabela 2.5), é possível observar que as covariáveis $OB2010$ e $GINI$ foram selecionadas e tiveram influência positiva na proporção de obesos em 2014,

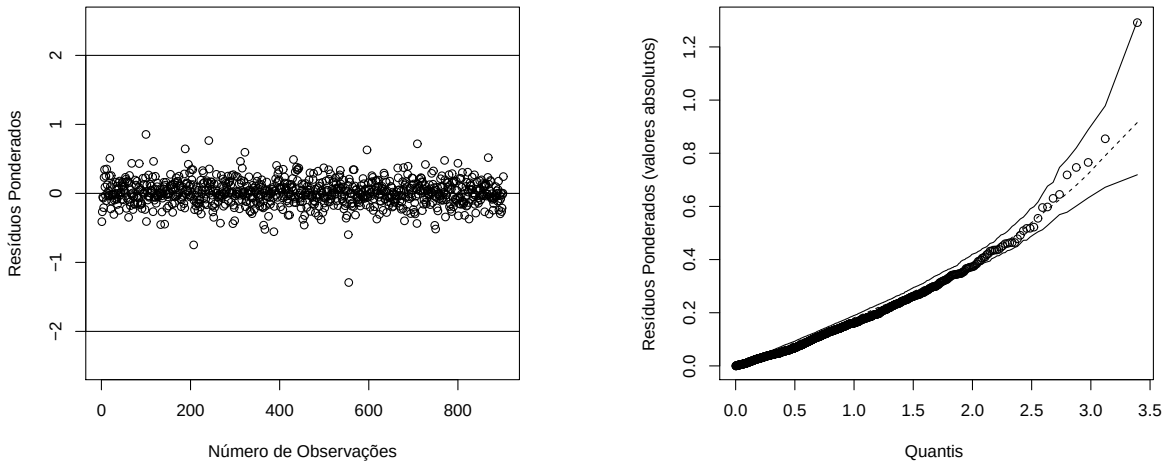


Figura 2.2: Gráfico dos resíduos quantis aleatorizados versus os índices de observações e gráfico de probabilidade normal com envelopes simulados - Região Sul.

enquanto que $RENDA$ influenciou negativamente, ou seja, nos municípios desta região que apresentaram uma maior proporção de indivíduos obesos em 2010 e um maior Índice de Gini, houve uma tendência de aumento na obesidade em crianças no ano de 2014, enquanto que, nos municípios com maior renda per capita, a tendência apresentada foi de diminuição na incidência de obesidade. Também foram selecionadas duas iterações, a primeira entre $OB2010$ e $GINI$ e a segunda entre $OB2010$ e $RENDA$.

Tabela 2.5: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão beta com dispersão variável para os dados referentes a Região Centro-Oeste.

Modelo para μ			
Variáveis	**Estimativa**	**Erro-Padrão**	***p*-valor**
$INTERCEPTO$	-2.967	3.972×10^{-1}	< 0.001
$OB2010$	1.123×10	3.070	< 0.001
$GINI$	1.571	7.690×10^{-1}	0.041
$RENDA$	-6.128×10^{-4}	2.481×10^{-4}	0.013
$INT1$	-2.310×10	5.954	< 0.001
$INT2$	5.305×10^{-3}	2.149×10^{-3}	0.013
Modelo para ϕ			
Variáveis	**Estimativa**	**Erro-Padrão**	***p*-valor**
$INTERCEPTO$	2.485	0.596	< 0.001
$SB2010$	-4.488	1.779	0.011
DU	0.719	0.229	0.001
$GINI$	2.486	1.163	0.032

Já em relação ao ajuste para o parâmetro de precisão, as variáveis $SB2010$, DU e $GINI$ foram selecionadas, sendo que $SB2010$ influenciou negativamente na precisão das respostas, ou seja, o fato de um município ter apresentado uma maior proporção de indivíduos com sobrepeso no ano de 2010, fez com que as respostas fossem menos precisas.

Em contrapartida, o fato de um município ter apresentado uma população urbana e um elevado Índice de Gini, fez com que houvesse uma maior precisão nas respostas.

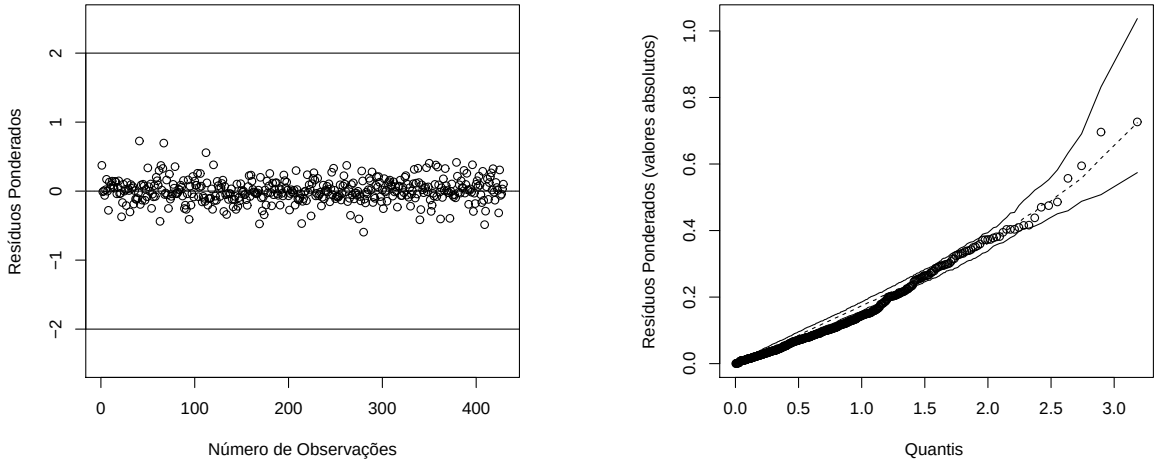


Figura 2.3: Gráfico dos resíduos quantis aleatorizados versus os índices de observações e gráfico de probabilidade normal com envelopes simulados - Região Centro-Oeste.

O modelo de regressão beta selecionado para explicar a proporção de crianças obesas na Região Centro-Oeste parece estar bem ajustado, visto que os resíduos permanecem dentro do intervalo $(-2, 2)$ e se encontram, em sua maioria, dentro das bandas de confiança dos envelopes simulados (ver Figura 2.3).

Analisando o modelo de regressão beta selecionado referente à Região Norte (Tabela 2.6), pode-se concluir que as variáveis $OB2010$, DU, $GPER$ e POP exerceram efeito positivo na variável resposta, ou seja, quanto maiores os valores destas variáveis nos mu-

Tabela 2.6: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão beta com dispersão variável para os dados referentes a Região Norte.

Modelo para μ			
Variáveis	**Estimativa**	**Erro-Padrão**	***p*-valor**
$INTERCEPTO$	-2.931	1.219×10^{-1}	< 0.001
$OB2010$	5.596	1.159	< 0.001
DU	1.837×10^{-1}	7.528×10^{-2}	0.014
$POBRES$	-8.490×10^{-3}	2.110×10^{-3}	< 0.001
$GPER$	2.456×10^{-3}	5.987×10^{-4}	< 0.001
POP	7.089×10^{-6}	1.935×10^{-6}	< 0.001
$INT1$	-1.679×10^{-2}	5.126×10^{-3}	0.001
$INT2$	-7.378×10^{-6}	1.956×10^{-6}	< 0.001
Modelo para ϕ			
Variáveis	**Estimativa**	**Erro-Padrão**	***p*-valor**
$INTERCEPTO$	3.657	0.338	< 0.001
$TDES$	-0.035	0.017	0.045
MI	0.030	0.014	0.029

nicípios desta região, maior a tendência de aumento na obesidade em crianças no ano de 2014. Por outro lado, a variável $POBRES$ apresentou influência negativa na variável resposta. Foram selecionadas duas iterações, sendo a $INT1$ iteração entre $OB2010$ e $GPER$, e $INT2$ iteração entre DU e POP.

Em relação à estrutura de regressão para a precisão, as variáveis $TDES$ e MI foram selecionadas, sendo que $TDES$ exerceu efeito negativo, enquanto que MI influenciou positivamente, ou seja, municípios que apresentaram uma taxa de desemprego maior tenderam a apresentar respostas menos precisas, enquanto que os que apresentaram uma taxa de mortalidade mais elevada tenderam a ter mais precisão nas respostas, isto é, nesses municípios as respostas tenderam a ser menos dispersas.

A Figura 2.4 apresenta os gráficos de resíduos ponderados versus índices de observações e o gráfico de probabilidade normal com envelopes simulados. O modelo de regressão beta selecionado para a Região Norte parece estar bem ajustado, visto que os resíduos permanecem dentro do intervalo $(-2, 2)$ e, em geral, dentro das bandas de confiança dos envelopes simulados.

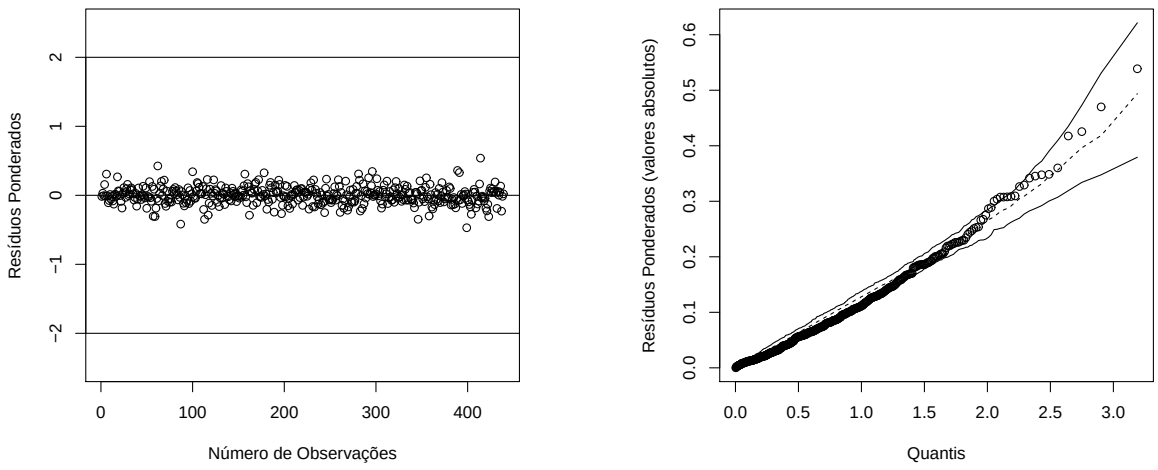


Figura 2.4: Gráfico dos resíduos quantis aleatorizados versus os índices de observações e gráfico de probabilidade normal com envelopes simulados - Região Norte.

As covariáveis selecionadas para explicar a proporção de crianças obesas em 2014 referentes ao ajuste da Região Nordeste se encontram apresentadas na Tabela 2.7. Através de sua análise, verifica-se que $POBRES$ e $TDES$ influenciaram positivamente na variável resposta, isto é, municípios com um maior percentual de pobres e uma maior taxa de desemprego, tenderam a apresentar uma maior incidência de indivíduos obesos. Este resultado pode estar relacionado ao fato de que famílias com menos renda ou que possuem indivíduos desempregados tendem a se alimentar de maneira inadequada, consumindo alimentos com alto teor calórico e influenciando diretamente na alimentação das crianças. Já as covariáveis $OB2010$, $GINI$ e MI influenciaram negativamente na proporção de crianças obesas em 2014. Além disso, três iterações foram selecionadas, sendo $INT1$ iteração entre as covariáveis $OB2010$ e MI, $INT2$ iteração entre $POBRES$ e $TDES$, e $INT3$ iteração entre $OB2010$ e $GINI$.

Tabela 2.7: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão beta com dispersão variável para os dados referentes a Região Nordeste.

Modelo para μ			
Variáveis	**Estimativa**	**Erro-Padrão**	***p*-valor**
$INTERCEPTO$	-6.422×10^{-1}	9.928×10^{-2}	< 0.001
$OB2010$	-3.146	8.057×10^{-1}	< 0.001
$POBRES$	2.898×10^{-3}	9.288×10^{-4}	0.001
$GINI$	-5.281×10^{-1}	1.660×10^{-1}	0.001
$TDES$	1.935×10^{-2}	3.781×10^{-3}	< 0.001
MI	-7.135×10^{-3}	1.490×10^{-3}	< 0.001
$INT1$	7.051×10^{-2}	1.179×10^{-2}	< 0.001
$INT2$	-3.936×10^{-4}	8.687×10^{-5}	< 0.001
$INT3$	4.692	1.383	< 0.001
Modelo para ϕ			
Variáveis	**Estimativa**	**Erro-Padrão**	***p*-valor**
$INTERCEPTO$	2.753	3.751×10^{-1}	< 0.001
$GINI$	2.104	7.081×10^{-1}	0.002
PIB	1.315×10^{-5}	5.185×10^{-6}	0.011
DU	-3.692×10^{-1}	1.125×10^{-1}	0.001

Considerando a modelagem para o parâmetro de precisão, as variáveis $GINI$ e PIB foram selecionadas e exerceram efeito positivo na precisão, enquanto que a variável DU, também selecionada, influenciou negativamente, ou seja, municípios que apresentaram um Índice de Gini e PIB elevados tenderam a ter mais precisão nas respostas, enquanto que o fato do município ter sido classificado como urbano, no ano de 2014, fez com que a precisão diminuísse, ou seja, as respostas tendessem a ser mais dispersas.

Através da análise dos gráficos apresentados na Figura 2.5, conclui-se que o modelo de regressão beta selecionado para a Região Nordeste parece estar bem ajustado, visto que os resíduos permanecem dentro do intervalo $(-2, 2)$. Porém, há resíduos que se encontram fora das bandas de confiança dos envelopes simulados, mas não há fortes indícios de afastamento da suposição de que o modelo de regressão beta selecionado é adequado para os dados.

Os modelos de regressão beta ajustados para as cinco regiões do Brasil conduzem a algumas conclusões relevantes. Nos municípios das Regiões Sul e Centro-Oeste a renda per capita influenciou negativamente na variável resposta, enquanto que a proporção de obesos para o ano de 2010 apresentou influência positiva. Na Região Sudeste o sobrepeso infantil influenciou positivamente a proporção de obesos em 2014. Nos municípios desta mesma região com uma taxa de desemprego mais elevada, houve uma tendência a apresentar uma maior incidência de crianças obesas, resultado este semelhante ao encontrado na Região Nordeste. Além disso, em duas das cinco regiões do Brasil, o gasto com assistencialismo per capita apresentou influência positiva na obesidade de crianças. No que se referem as diferenças encontradas entre os modelos de regressão, ainda considerando

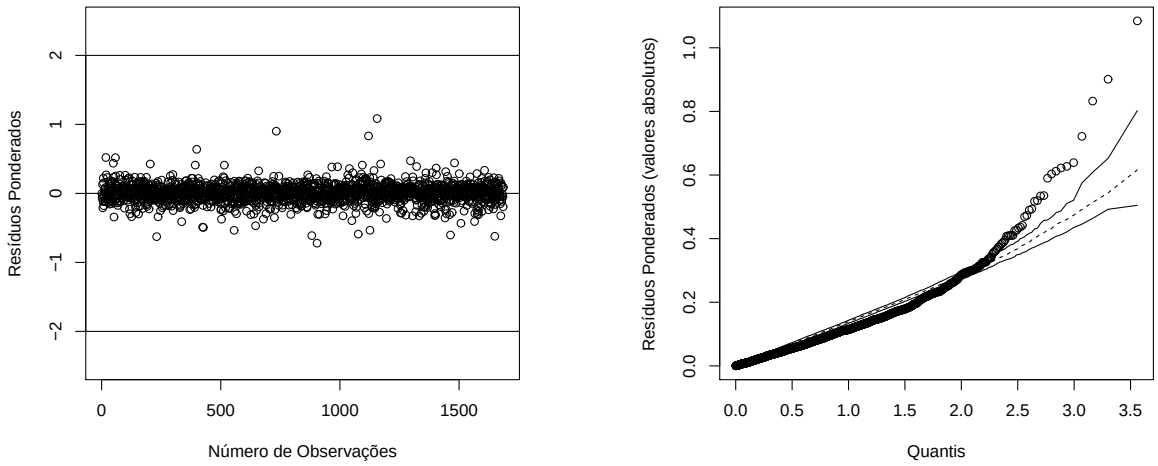


Figura 2.5: Gráfico dos resíduos quantis aleatorizados versus os índices de observações e gráfico de probabilidade normal com envelopes simulados - Região Nordeste.

a estrutura de regressão para a média, podemos destacar o fato de que o percentual de pobres dos municípios apresentou diferentes influências entre as Regiões Norte e Nordeste. Nos municípios da Região Norte esta variável apresentou influência negativa na variável resposta, enquanto que na Região Nordeste, houve uma influência positiva nesta mesma variável.

Em relação a estrutura de regressão para o parâmetro de precisão, nos municípios das Regiões Sudeste, Sul e Centro-Oeste classificados como urbanos, houve uma tendência a apresentar uma maior precisão nas respostas, ao contrário do resultado encontrado na Região Nordeste. Vale destacar também, que para as Regiões Sudeste e Sul a taxa de mortalidade infantil influenciou negativamente na precisão das respostas, ou seja, conforme um determinado município apresentou um maior valor para essa variável, a precisão das respostas tendeu a diminuir.

2.5 Conclusões

Neste capítulo avaliamos e explicamos a proporção de crianças obesas, entre 0 e 5 anos de idade, beneficiadas pelo Programa Bolsa Família no ano de 2014 e identificamos os fatores que influenciaram na obesidade desses indivíduos em cada uma das cinco regiões brasileiras. Para isso, utilizamos o modelo de regressão beta que é apropriado para situações em que a variável resposta é uma proporção, ou seja, restrita ao intervalo $(0, 1)$. Para cada uma das cinco regiões brasileiras, há modelagens considerando a estrutura de regressão da média e precisão do modelo de regressão beta com dispersão variável.

Verificamos que para os municípios da Região Sudeste, variáveis como o sobrepeso e taxa de desemprego tiveram influência positiva na obesidade. Vale destacar que nesta região a variável gasto per capita com assistencialismo foi significativa, e influenciou positivamente na variável resposta, ou seja, municípios com maiores gastos per capita com o

Programa Bolsa Família tenderam a apresentar uma maior incidência de crianças obesas. Nos municípios da Região Sul, a obesidade no ano de 2010 e o Índice de Desenvolvimento Humano apresentaram influência positiva, enquanto que nestes municípios, a renda per capita apresentou influência negativa na proporção de indivíduos obesos. Para a Região Centro-Oeste, as variáveis obesidade em 2010 e renda per capita dos municípios também foram selecionadas, sendo que a obesidade no ano de 2010 influenciou positivamente e a renda per capita negativamente, resultados semelhantes aos encontrados na Região Sul. Adicionalmente, na Região Centro-Oeste o Índice de Gini apresentou influência positiva na proporção de obesos.

O modelo de regressão ajustado para a Região Norte revelou que variáveis como a obesidade em 2010, o fato do município ter sido classificado como urbano e sua população influenciaram positivamente na variável resposta. Neste ajuste a variável gasto per capita com assistencialismo foi selecionada e apresentou influência positiva na obesidade de crianças, ou seja, assim como na Região Sudeste, municípios com maiores gastos com o Programa Bolsa Família tenderam a apresentar uma maior incidência de crianças obesas. Já em relação ao ajuste referente a Região Nordeste, o percentual de pobres e a taxa de desemprego tiveram influência positiva. Em contrapartida, nesta mesma região, os municípios com maiores taxas de mortalidade infantil tenderam a apresentar uma menor incidência de obesidade em crianças.

Observamos que, para os modelos de regressão beta de duas das cinco regiões brasileiras, a variável gasto per capita com o Programa Bolsa Família foi selecionada e apresentou influência positiva na obesidade. Este resultado exige atenção, pois o rendimento extra proveniente do benefício nestas duas regiões pode ter feito com que as famílias participantes do programa tenham aumentado o consumo de produtos industrializados e com alta densidade calórica, influenciando assim no estado nutricional das crianças pertencentes as mesmas. Como forma de tentar reverter este cenário, algumas intervenções por parte do governo poderiam ser realizadas, com enfoque na implementação de programas direcionados a educação alimentar dos beneficiários, a intensificação de políticas de assistência social com o objetivo de melhorar as condições de saúde e educação das famílias, e o reforço a programas de segurança alimentar de modo a facilitar o acesso dos beneficiários a alimentos nutricionalmente adequados, como frutas, legumes e verduras.

Capítulo 3

Modelagem da proporção de obesos nos Estados Unidos utilizando o modelo de regressão beta com dispersão variável

3.1 Introdução

A obesidade é uma doença de abrangência mundial podendo afetar tanto países desenvolvidos quanto subdesenvolvidos. Segundo a Organização Mundial de Saúde (OMS), a obesidade é definida como a excessiva concentração de gordura que pode prejudicar a saúde do indivíduo, sendo a falta de atividade física e o consumo exagerado de alimentos altamente energéticos dois dos principais fatores para o surgimento dessa morbidade. Além disso, ela pode estar associada a outras doenças, a exemplo dos problemas respiratórios, problemas circulatórios, diabetes ou até mesmo o surgimento do câncer (WORLD HEALTH ORGANIZATION, 2015).

Cabrera e Filho (2001) apresentaram três medidas antropométricas distintas para avaliar a concentração e o volume de gordura no indivíduo, a saber: o IMC (Índice de Massa Corporal), o RCQ (Razão Cintura-Quadril) e o CA (Circunferência Abdominal). O IMC é definido como a razão entre o peso do indivíduo dado em quilogramas (kg) e sua altura ao quadrado (m^2). Dessa forma, Stol et al. (2011) apresentaram três classificações para a obesidade com base nessa medida: grau I com $30.0 \leq \text{IMC} \leq 34.9\ kg/m^2$, grau II com $35.0 \leq \text{IMC} \leq 39.9\ kg/m^2$ e grau III com $\text{IMC} \geq 40.0\ kg/m^2$. O RCQ é uma medida utilizada para verificar o risco do indivíduo apresentar doenças cardiovasculares, definida como a razão entre a circunferência da cintura e a circunferência do quadril. Assim, homens que apresentarem $\text{RCQ} \geq 0.9$ e mulheres que apresentarem $\text{RCQ} \geq 0.85$ estão mais sujeitas a tais riscos (GABRIELE, 2011). O CA é também uma medida de risco para doenças cardiovasculares, de modo que homens e mulheres que apresentarem $\text{CA} > 0.94$ cm e $\text{CA} > 0.80$ cm, respectivamente, apresentam maiores riscos de adquirirem doenças cardíacas (RESENDE et al., 2006).

Segundo estimativas da OMS, cerca de 13% da população mundial adulta, 11% dos homens e 15% das mulheres, eram obesos em 2014. Em 2008 verificou-se que a maior prevalência de pessoas com sobrepeso ou obesas encontravam-se na América, com cerca de 62% de pessoas com sobrepeso e 26% obesas, e a menor no Sudeste da Ásia, com cerca de 14% com excesso de peso e 3% obesas. Além disso, é verificado que a obesidade é responsável por cerca de 2.8 milhões de mortes no mundo devido às suas complicações (WORLD HEALTH ORGANIZATION, 2015). O consumo de refrigerantes tem um papel

fundamental para explicar o grande aumento da proporção de obesos ao redor do mundo. Basu et al. (2013), por exemplo, mostraram através de um modelo de regressão que a variável consumo de refrigerantes está relacionada de forma significativa com o aumento do sobrepeso, obesidade e diabetes no mundo. Mostrando assim, que os cuidados com a alimentação são de fundamental importância para prevenir essas e outras doenças.

Os Estados Unidos é um dos países desenvolvidos que mais sofrem com os problemas relacionados a obesidade. De 2011 à 2012 cerca de um terço da população adulta era obesa, sendo encontrada uma maior prevalência entre negros não-hispânicos (47.8%), hispânicos (42.5%), brancos não-hispânicos (32.6%) do que entre os asiáticos não-hispânicos (10.8%), contudo não encontrou-se diferenças entre as prevalências de obesidade de homens e mulheres (OGDEN et al., 2014). A obesidade é uma doença que pode ser prevenida ou tratada. O tratamento dessa doença e dos problemas relacionados a ela causam um grande impacto nos cofres públicos, que poderiam ser evitados se existissem políticas públicas no setor de saúde que conscientizassem as pessoas sobre os cuidados com a saúde, a importância de uma boa alimentação ou as vantagens de se praticar atividades físicas. Segundo dados de Arterburn et al. (2005), os gastos no setor de saúde pública relacionados a obesidade adulta nos Estados Unidos custaram cerca de 11 bilhões de dólares no ano de 2000. Ou seja, o simples cuidado com a saúde poderia significar uma redução nos custos relacionados a medicamentos, internações hospitalares, atendimentos médicos, redução de doenças associadas e uma melhor qualidade de vida.

Danaei et al. (2009) apontaram que as mortes nos Estados Unidos no ano de 2005 relacionados a inatividade física, obesidade/sobrepeso e glicemia elevada mataram de 24% a 27% dos estadunidenses, já o tabagismo foi responsável por cerca de 16.7% a 20% das mortes adultas. Pietiläinen et al. (2008) apresentaram em seu estudo que a inatividade física na adolescência se mostra como um fator de risco para a obesidade e a obesidade abdominal em adultos com 25 anos, ou seja, a falta de atividade física aumentava os riscos dessas condições em 3.9 e 4.8 vezes, respectivamente, constituindo um dos principais fatores para o aumento dessa morbidade ao longo dos anos.

Neste cenário, o nosso objetivo com o presente capítulo é avaliar a proporção de adultos obesos nos Estados Unidos, uma vez que esse país está entre as nações desenvolvidas que mais sofrem com os problemas relacionados à obesidade. Para isso, utilizamos o modelo de regressão beta com dispersão variável (FERRARI; CRIBARI-NETO, 2004; SIMAS et al. 2010). A classe de modelos de regressão beta tem como objetivo permitir a modelagem de respostas restritas ao intervalo $(0, 1)$, por meio de uma estrutura de regressão que contém funções de ligação para modelar a média e a precisão, além de covariáveis e parâmetros de regressão desconhecidos. Os dados utilizados nesse capítulo foram extraídos a partir de algumas fontes de informações públicas. O procedimento computacional foi desenvolvido utilizando o pacote `betareg` (CRIBARI-NETO; ZELEIS, 2010) do *software* estatístico `R` (KLEIBER; ZELEIS, 2008; R DEVELOPMENT CORE TEAM, 2014).

O capítulo encontra-se dividido em cinco seções. A Seção 2 apresenta o modelo de regressão beta com dispersão variável. Uma breve descrição dos dados encontra-se na Seção 3. Na Seção 4 são apresentados os resultados obtidos a partir do ajuste do modelo selecionado. Por último, na Seção 5 são apresentadas as conclusões e considerações finais.

3.2 Modelo de regressão beta

A classe de modelos de regressão beta é comumente utilizada em modelagens de variáveis que assumem valores no intervalo unitário $(0, 1)$, a exemplo de taxas e proporções. Estes modelos são baseados na suposição de que a variável dependente tem distribuição beta e que a sua média é relacionada a um preditor linear por meio de uma função de ligação. O preditor linear envolve covariáveis e parâmetros de regressão desconhecidos. O modelo também inclui um parâmetro de dispersão, que em certas situações podem variar ao longo das observações (ALMEIDA JUNIOR; SOUZA, 2015; CRIBARI-NETO; SOUZA, 2012; ESPINHEIRA et al., 2008a, 2008b; SILVA; SOUZA, 2014; SIMAS et al., 2010; SMITHSON; VERKUILEN, 2006; SOUZA; CRIBARI-NETO, 2015).

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

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{3.1}$$

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

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, em que cada y_t, $t = 1, \ldots, n$, segue a densidade da Equação (3.1) com média μ_t e parâmetro de precisão ϕ_t sendo desconhecidos. O modelo proposto por Ferrari e Cribari-Neto (2004) é obtido assumindo que a média de y_t pode ser escrita como

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t, \tag{3.2}$$

em que $\beta = (\beta_1, \ldots, \beta_k)^\top$ é um vetor de parâmetros de regressão desconhecidos ($\beta \in \mathbb{R}^k$), $x_{t1}, \ldots, x_{tk}$ são observações de k covariáveis e η_t é o preditor linear. Por fim, $g(\cdot)$ a função de ligação, $g : (0, 1) \to \mathbb{R}$, é estritamente monótona e duas vezes diferenciável. Portanto, $\mu_t = g^{-1}(\eta_t)$ e $\mathrm{var}(y_t) = \mu_t(1-\mu_t)/(1+\phi)$.

O modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) considera o parâmetro de precisão constante ao longo das observações. Contudo, admitimos como em Simas et al. (2010) que o parâmetro de precisão é variável, sendo modelado em termos de covariáveis, parâmetros desconhecidos e de uma função de ligação, sendo essa estrutura dada da seguinte forma:

$$h(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \vartheta_t, \tag{3.3}$$

em que $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1}, \ldots, z_{tq}$ são observações de q covariáveis ($k + q < n$), assumidas fixas e conhecidas, ϑ_t é o preditor linear, e $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável que mapeia os pontos positivos da reta, $h : (0, \infty) \to \mathbb{R}$. Portanto, $\phi_t = h^{-1}(\vartheta_t)$. Há várias possíveis

escolhas para as funções de ligação $g(\cdot)$ e $h(\cdot)$. Para $g(\cdot)$ pode-se utilizar a função de ligação logit, $g(\mu) = \log\{\mu/(1-\mu)\}$, ou cloglog, $g(\mu) = \log\{-\log(1-\mu)\}$, entre outras. Já para $h(\cdot)$, pode-se utilizar a função log, $h(\phi) = \log(\phi)$, ou sqrt, $h(\phi) = \sqrt{\phi}$, entre outras. Para maiores detalhes sobre as funções de ligação ver McCullagh e Nelder (1989).

Segue de (3.1) que o logaritmo da função de verossimilhança é

$$\ell(\beta, \gamma) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t), \tag{3.4}$$

em que

$$\begin{aligned} \ell_t(\mu_t, \phi_t) &= \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) + (\mu_t\phi_t - 1)\log y_t \\ &+ \{(1-\mu_t)\phi_t - 1\}\log(1-y_t). \end{aligned} \tag{3.5}$$

Como os estimadores de máxima verossimilhança de β e γ não possuem forma fechada, eles precisam ser obtidos numericamente maximizando a função de log-verossimilhança através de um algoritmo de maximização não-linear. Usualmente, utiliza-se o método quasi-Newton BFGS (PRESS et al., 1992). Para maiores detalhes inferenciais e expressões matriciais do vetor escore e da matriz de informação de Fisher, ver Simas et al. (2010).

Sob certas condições de regularidade, para tamanhos de amostras grandes, a distribuição conjunta de β e γ é aproximadamente normal $(k+q)$-multivariada:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, K^{-1} \right), \tag{3.6}$$

em que $\widehat{\beta}$ e $\widehat{\gamma}$ são os estimadores de máxima verossimilhança de β e γ, respectivamente, e K^{-1} é a inversa da matriz de informação de Fisher.

3.3 Descrição dos dados

A Tabela 3.1 apresenta a descrição das variáveis utilizadas neste estudo. As fontes de dados consultadas foram as páginas da web: `http://stateofobesity.org`, `http://health-stats.azurewebsites.net/`, `http://www.gallup.com` e `http://map.feedingamerica.org`.

A Tabela 3.2 apresenta algumas estatísticas descritivas, como mínimo, primeiro quartil ($Q_{1/4}$), mediana, média, terceiro quartil ($Q_{3/4}$), máximo e coeficiente de variação (CV) das variáveis utilizadas. Essas estatísticas são baseadas em 50 observações referentes aos estados dos Estados Unidos. Algumas conclusões podem ser feitas através de sua análise. Para a variável proporção de adultos obesos, o valor máximo encontrado foi de 0.36, ou seja, para o estado correspondente a este valor, cerca de 36% dos adultos apresentaram obesidade no ano de 2014, enquanto que o valor mínimo encontrado para essa mesma variável foi de 0.21.

Para a variável porcentagem de adultos considerados inativos físicos, temos que 75% dos estados norte-americanos apresentaram uma porcentagem menor do que 25.23%. Considerando a porcentagem de indivíduos que consumiam vegetais menos de uma vez ao dia,

Tabela 3.1: Descrição das variáveis utilizadas.

Variáveis	**Definição**
$OB2014$	Proporção de adultos obesos em 2014
$INAT$	Porcentagem de inatividade física entre adultos em 2014
$VEGET$	Porcentagem de adultos que consumiam vegetais menos de uma vez por dia em 2011
FUM	Porcentagem de fumantes de cigarro em 2012
$DESEMP$	Porcentagem de residentes desempregados ou empregados em tempo parcial em 2014
$INSEG$	Taxa de insegurança alimentar em 2013
BST	Escore de bem-estar em 2014
$DESCOB$	Porcentagem de residentes que não tinham cobertura de seguro de saúde em 2014

50% dos estados apresentaram um valor menor do que 22.90%. Para a variável FUM, que representa a porcentagem de adultos que fumavam cigarro no ano de 2012, o menor valor encontrado foi de 10.60%, enquanto que o coeficiente de variação foi igual a 18.27%, o que indica uma média dispersão da variável.

Em relação ao percentual de indivíduos desempregados ou empregados em tempo parcial, 25% dos estados apresentaram um valor menor que 13.35%. Já para a variável taxa de insegurança alimentar, que segundo a fonte consultada (`http://map.feedingamerica.org`) representa uma medida à falta de acesso a alimentos suficientes para uma vida ativa e saudável e à um acesso limitado ou incerto a alimentos nutricionalmente adequados, temos que 75% dos estados norte-americanos apresentaram um valor menor do que 16.98%.

A variável escore de bem-estar, que de acordo com a fonte consultada (`http://www.gallup.com`) é constituída por cinco elementos de bem-estar que são os principais componentes para uma vida melhor (proposital, social, financeiro, comunitário e físico) apresentou os valores de mínimo e máximo de 59.00 e 64.70, respectivamente. Em relação a variável que refere-se a porcentagem de residentes do estado que não tinham cobertura de seguro de saúde no ano de 2014, 75% dos estados apresentaram um valor menor que 15.30%.

Destacamos que a maior proporção de adultos obesos no ano de 2014 foi registrada no estado do Arkansas, enquanto que a menor proporção foi encontrada no estado do

Tabela 3.2: Estatísticas descritivas das variáveis utilizadas.

Variáveis	**Mínimo**	$\mathbf{Q}_{1/4}$	**Mediana**	**Média**	$\mathbf{Q}_{3/4}$	**Máximo**	**CV(%)**
$OB2014$	0.21	0.27	0.29	0.29	0.31	0.36	11.15
$INAT$	16.40	20.30	22.90	23.02	25.23	31.60	15.84
$VEGET$	15.30	20.70	22.90	23.21	25.85	32.50	15.42
FUM	10.60	17.32	19.55	19.84	22.38	28.30	18.27
$DESEMP$	9.00	13.35	15.20	15.07	16.88	20.70	17.46
$INSEG$	7.80	13.30	14.60	14.97	16.98	22.70	17.55
BST	59.00	61.10	61.90	61.86	62.62	64.70	1.98
$DESCOB$	4.60	10.15	12.75	12.73	15.30	24.40	30.72

Colorado. Para a variável inatividade física, os valores de mínimo e máximo foram encontrados em Colorado e Mississipi, respectivamente. Enquanto que Oregon apresentou a menor proporção de indivíduos que consumiam vegetais menos de uma vez por dia. Já o estado de Kentucky apresentou o maior percentual de adultos fumantes de cigarro no ano de 2012.

Ao analisarmos a porcentagem de residentes desempregados ou empregados em tempo parcial, verificamos que Nevada apresentou o maior percentual para esta variável, enquanto que North Dakota apresentou a maior taxa de insegurança alimentar no ano de 2013. O valor máximo para a variável escore de bem-estar foi encontrado no estado do Alaska, enquanto que o menor percentual de residentes que não tinham cobertura de seguro de saúde no ano de 2014 foi registrado em Massachussets.

A Figura 3.1 apresenta o histograma e o *Box-plot*, respectivamente, da variável proporção de adultos obesos nos Estados Unidos. Dessa forma, é possível visualizar a distribuição dos dados e uma certa assimetria à esquerda, uma vez que a mediana está mais próxima do terceiro quartil. No *Box-plot* é possível destacar uma observação discrepante que excede seus limites, referente ao estado do Colorado. Portanto a utilização do modelo de regressão beta se faz necessário, visto que a variável resposta é uma proporção e foi verificada uma certa assimetria em sua distribuição.

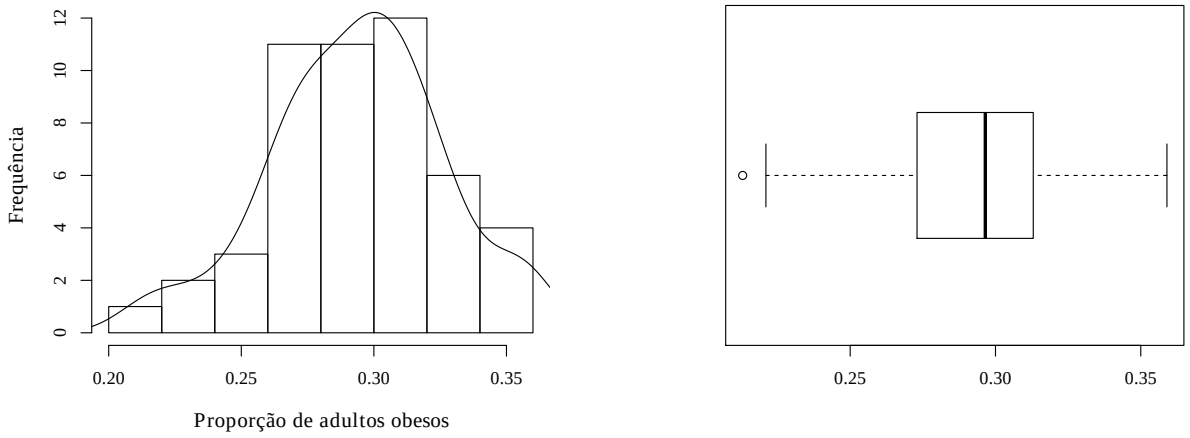


Figura 3.1: Histograma e *Box-plot* da variável proporção de adultos obesos nos Estados Unidos em 2014.

3.4 Especificação do modelo

Inicialmente, ao ajustarmos o modelo de regressão beta, estamos interessados em testar a hipótese nula que a dispersão dos dados é fixa versus a hipótese alternativa de que a dispersão dos dados é variável. Para tanto, utilizamos o teste da razão de verossimilhanças (NEYMAN; PEARSON, 1928; SILVA; SOUZA, 2014) e obtivemos um p-valor < 0.001 (valor obtido a partir dos dados amostrais e reflete a probabilidade de rejeitar a hipótese

nula dado que ela é verdadeira), ou seja, ao nível de significância de 5% rejeitamos a hipótese nula de que a dispersão é fixa, portanto se faz necessário um ajuste para modelar a dispersão dos dados, aqui representada pelo parâmetro de precisão ϕ_t. O modelo de regressão beta com dispersão variável selecionado foi:

$$\begin{aligned} \text{cloglog}(\mu_t) &= \beta_0 + \beta_1 INAT_t + \beta_2 VEGETt + \beta_3 FUM_t + \beta_4 DESEMP_t \\ &+ \beta_5 INSEG_t + \beta_6 BST_t + \beta_7 INT1_t, \\ \log(\phi_t) &= \gamma_0 + \gamma_1 INSEG_t + \gamma_2 DESCOB_t + \gamma_3 VEGET_t, \end{aligned} \quad (3.7)$$

com $t = 1, \ldots, 50$.

A análise de diagnóstico é uma etapa da regressão que permite verificar algumas suposições do modelo, tais como: aleatoriedade dos resíduos, adequação da distribuição de probabilidade suposta para a variável resposta e a identificação de possíveis pontos de influência e de alavanca. Neste estudo utilizamos os resíduos ponderados padronizados (ESPINHEIRA et al., 2008a). Para verificarmos a qualidade do ajuste do modelo foi utilizado o coeficiente de determinação ajustado (pseudo-R^2), o teste *RESET* (RAMSEY, 1969) e o gráfico de probabilidade normal com envelopes simulados.

O pseudo-R^2 é uma medida global da variação explicada e é análogo ao coeficiente de determinação, utilizado em modelos lineares de regressão. Ferrari e Cribari-Neto (2004) propuseram um pseudo-R^2 para os modelos de regressão beta, definido como o quadrado do coeficiente de correlação entre $\hat{\eta}$ e $g(y)$. Dessa forma com um pseudo-R^2=0.75 constatamos que as variáveis independentes foram capazes de explicar cerca de 75% da variabilidade total da proporção de adultos obesos. Para testar a correta especificação do modelo, utilizamos o teste *RESET* para modelos de regressão beta (LIMA, 2007). A hipótese nula deste teste sugere que o modelo proposto está bem especificado contra a hipótese alternativa de que o modelo está mal especificado. O teste realizado consistiu em adicionar como variável de teste o preditor linear estimado elevado a segunda potência ($\widehat{\eta}^2$) ao submodelo da média. Desta forma, obtivemos um p-valor=0.208, ou seja, como a variável de teste não mostrou-se significativa, podemos concluir que o modelo proposto não apresentou nenhum erro de especificação ao nível de significância de 5%.

A Figura 3.2 apresenta os gráficos dos resíduos ponderados padronizados versus os índices das observações e o gráfico de probabilidade normal com envelopes simulados. No primeiro gráfico é possível visualizar que os resíduos estão distribuídos de forma aleatória entre os limites $(-2, 2)$ com apenas as observações 2, 6 e 11 ultrapassando esse intervalo, correspondendo aos estados do Alaska, Colorado e Hawaii. Portanto, temos que a suposição de que os resíduos são uma sequência aleatória não foi violada. Já o gráfico de probabilidade normal com envelopes simulados é uma técnica gráfica que permite identificar possíveis observações discrepantes, bem como a adequação da distribuição de probabilidade que foi suposta para o modelo. Na gráfico notamos que as observações encontram-se distribuídas de forma aleatória dentro do envelope e próximo a linha central, não sendo possível detectar observações discrepantes. Portanto, não temos evidências de que o modelo especificado não está adequado.

A distância de Cook (COOK, 1977) é uma medida de influência utilizada para quantificar o impacto de cada observação na estimativa dos parâmetros desconhecidos. Espinheira

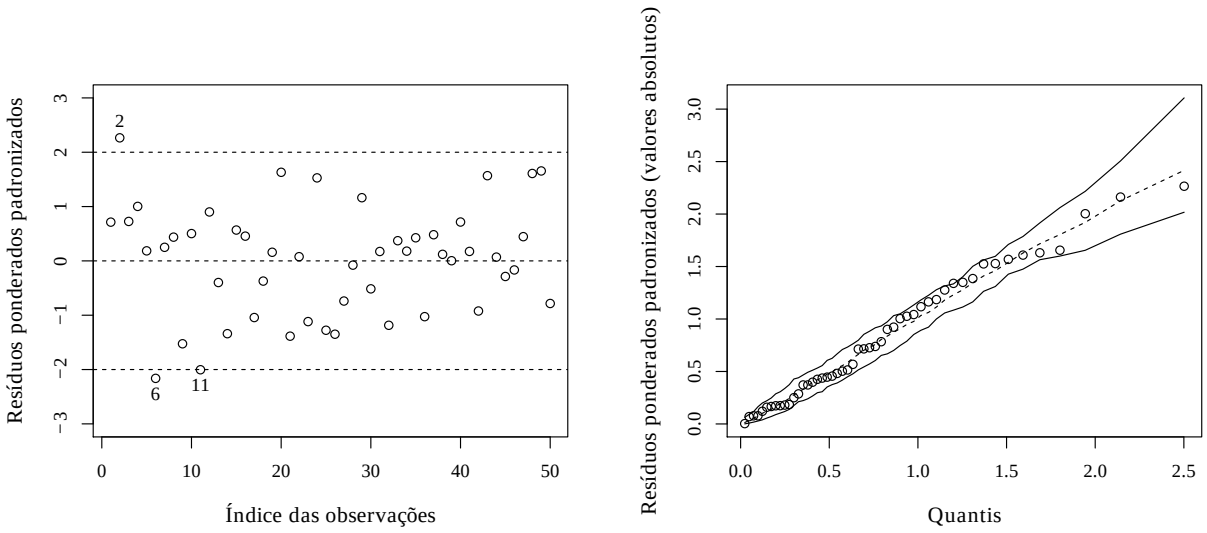


Figura 3.2: Gráfico dos resíduos ponderados padronizados versus os índices de observações e gráfico de probabilidade normal com envelopes simulados.

et al. (2008b) propuseram uma medida similar a distância de Cook e medidas de influência local para modelos de regressão beta. Por outro lado, a alavancagem generalizada proposta por Wei et al. (1998) é definida em modelos de regressão como uma medida da importância individual das observações. Ferrari et al. (2011) propuseram a alavancagem generalizada para modelos de regressão beta com dispersão variável.

A Figura 3.3 apresenta o gráfico das distâncias de Cook versus os valores preditos e da alavancagem generalizada versus os valores preditos. Essas técnicas gráficas permitem identificar observações que interferem nas estimativas dos parâmetros produzindo resultados distorcidos. No gráfico da distância de Cook não foi possível identificar nenhum

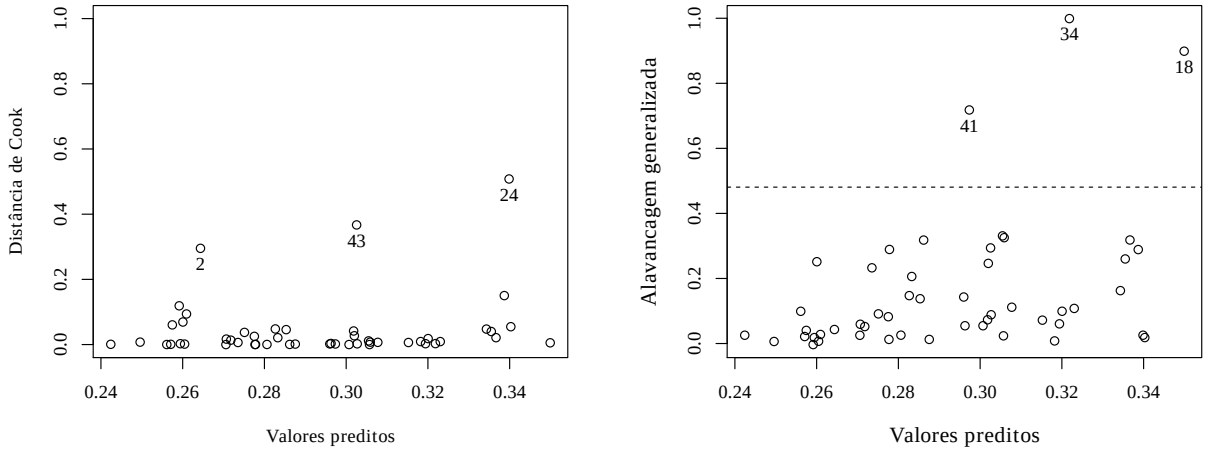


Figura 3.3: Gráfico da distância de Cook e de alavancagem generalizada.

ponto de influência. Entretanto algumas observações apresentam valores de Cook diferenciados, a saber, as observações 2, 24 e 43 referentes, respectivamente, aos estados do Alaska, Mississippi e Texas. No gráfico da alavancagem generalizada é possível identificar três pontos de alavanca referentes aos estados da Louisiana, North Dakota e South Dakota.

A Tabela 3.3 apresenta os resultados obtidos a partir do ajuste do modelo de regressão beta com dispersão variável. Esse modelo utiliza as funções de ligação cloglog e log para modelar a média e a precisão, respectivamente, uma vez que estas forneceram um melhor ajuste. Através do teste de Wald (WALD, 1943; CRIBARI-NETO; ZELEIS, 2010) verificamos que as variáveis relevantes para explicar a proporção de adultos obesos na modelagem da média foram: $INAT$, $VEGET$, FUM, $DESEMP$, $INSEG$, BST e a interação entre as variáveis taxas de fumantes e de insegurança alimentar, denotada por $INT1$, pois apresentaram p-valores menores que o nível de significância de 5%, rejeitando assim a hipótese nula de que $\beta_j = 0$. Em relação a modelagem da precisão verificamos que as variáveis $INSEG$, $DESCOB$ e $VEGET$ foram significativas, pois apresentaram p-valores menores que o nível de significância de 5%.

Através da análise dos coeficientes do modelo proposto, é possível verificar que as variáveis $INAT$, $VEGET$, FUM e $INSEG$ apresentaram efeito positivo na variável resposta, ou seja, os estados que apresentaram maiores valores para estas variáveis tenderam a apresentar uma maior incidência de adultos obesos, por outro lado, os estados com maiores porcentagens de desempregados ou empregados em tempo parcial ($DESEMP$) e maiores escores de bem estar (BST) tenderam a apresentar uma menor incidência de obesidade em adultos. Considerando a estrutura de regressão para o parâmetro de precisão, temos que à medida que a covariável $INSEG$ aumenta a precisão diminui, ou seja, os estados que apresentaram maiores valores de insegurança alimentar tenderam a apre-

Tabela 3.3: Estimativas dos coeficientes, erros-padrão e p-valores do modelo de regressão beta com dispersão variável para os dados da obesidade adulta nos Estados Unidos.

Modelo para μ			
Variáveis	**Estimativa**	**Erro-Padrão**	**p-valor**
$INTERCEPTO$	-1.473	0.485	0.002
$INAT$	0.012	0.004	0.002
$VEGET$	0.011	0.002	< 0.001
FUM	0.055	0.016	< 0.001
$DESEMP$	-0.014	0.002	< 0.001
$INSEG$	0.069	0.025	0.005
BST	-0.018	0.006	0.003
$INT1$	-0.003	0.001	0.004
Modelo para ϕ			
Variáveis	**Estimativa**	**Erro-Padrão**	**p-valor**
$INTERCEPTO$	3.679	1.556	0.018
$INSEG$	-0.509	0.094	< 0.001
$DESCOB$	0.196	0.061	0.001
$VEGET$	0.370	0.058	< 0.001
Pseudo-R^2	0.75		

sentar respostas menos precisas. Por outro lado, à medida que as covariáveis $DESCOB$ e $VEGET$ aumentaram, a precisão também aumentou, ou seja, os estados com maiores valores de descobertos e consumo de vegetais tenderam a apresentar respostas mais precisas.

Os resultados obtidos na análise inicial da regressão beta concordam com alguns resultados encontrados na literatura. Primeiro, no estudo de Cavalcanti et al. (2010) por meio da regressão logística mostrou-se que a prática de atividade física é um fator de proteção para a obesidade abdominal em adolescentes de 14 a 19 anos de idade, independente da presença do excesso de peso. Segundo, no estudo de Castanho et al. (2013) por meio da regressão logística verificou-se também que o consumo de frutas de maneira adequada reduz as chances de se adquirir obesidade abdominal. Ainda a ingestão de frutas, verduras e legumes apresentou um efeito significativo reduzindo o risco de se adquirir doenças cardiovasculares. Terceiro, no estudo de Flegal (2007) concluiu-se que grandes mudanças na prevalência de tabagismo causavam um pequeno efeito na prevalência da obesidade, ocasionando um aumento muitas vezes menor do que um ponto percentual na prevalência de obesos. Quarto, no estudo de Zhang et al. (2014) verificou-se uma associação entre as taxas de desemprego e o peso dos indivíduos nos estados e cidades. Concluindo a partir da regressão logística que as taxas de desemprego nos estados estão associadas negativamente com o IMC individual ao longo dos anos.

Quinto, Dharod et al. (2013) por meio de um estudo transversal analisaram a associação entre insegurança alimentar, consumo alimentar e índice de massa corporal entre mulheres refugiadas Somalis que viviam nos Estados Unidos no período de outubro de 2006 até dezembro de 2007. Verificou-se que a insegurança alimentar estava associada positivamente ao sobrepeso e a obesidade, ou seja, ela apresentava-se como um fator de risco aumentando as chances do indivíduo vir a ser obeso. Sexto, Jagielski et al. (2014) exploraram a associação entre adiposidade, qualidade de vida e bem-estar mental entre indivíduos obesos que entraram em um serviço de gestão de peso. Verificando-se que a adiposidade está relacionada negativamente com o bem-estar e a qualidade de vida, sendo identificada uma alta prevalência de comorbidades psicológicas e uma redução da qualidade de vida dos indivíduos obesos. Segundo Holben (2010) a insegurança alimentar é uma ameaça a saúde que pode ser evitada. Além disso, ela está relacionada a diabetes, a incidência e o risco de doenças crônicas, ao excesso de peso e a obesidade. Entretanto é necessário maiores estudos para direcionar a relação entre a insegurança alimentar e o excesso de peso ou obesidade nos adultos.

Dando continuidade a análise de regressão do modelo ajustado, com o objetivo de verificarmos o impacto que as observações de alta alavancagem podem causar na estimativa dos parâmetros, decidimos por excluí-las individualmente e conjuntamente, sendo assim as variações percentuais das estimativas devido as observações 18, 34 e 41 podendo ser visualizadas na Tabela 3.4. A partir da análise descritiva das variáveis relevantes ao modelo foi possível verificar que o estado da Louisiana, observação 18, se destacou por apresentar a maior taxa de $VEGET$ e uma das menores taxas de BST. Ao excluirmos a observação 18, temos que a variação percentual de $\hat{\beta}_2$ é 25.10%. Em relação ao submodelo da precisão, temos que o maior impacto ocorreu no intercepto, $\hat{\gamma}_0$, com uma redução de -37.49%, diminuindo a precisão das respostas.

Tabela 3.4: Variações percentuais nas estimativas dos parâmetros ao se retirar observações influentes. Proporção de adultos obesos nos Estados Unidos.

Casos	**18**	**34**	**41**	**18, 34, 41**
$\hat{\beta}_0$	8.09	16.73	−11.64	10.44
$\hat{\beta}_1$	9.65	9.73	1.60	15.61
$\hat{\beta}_2$	25.10	−16.97	0.38	−13.45
$\hat{\beta}_3$	−12.34	−34.98	2.49	−24.07
$\hat{\beta}_4$	−11.21	−13.38	5.70	−8.53
$\hat{\beta}_5$	−13.95	−31.48	−0.98	−0.98
$\hat{\beta}_6$	−13.31	−51.17	12.71	−29.42
$\hat{\beta}_7$	−14.56	−39.64	−1.70	−25.64
$\hat{\gamma}_0$	−37.49	−76.35	−0.86	−61.67
$\hat{\gamma}_1$	−23.60	−52.55	−11.53	−74.77
$\hat{\gamma}_2$	−20.95	−10.12	−10.86	−35.03
$\hat{\gamma}_3$	1.35	−6.94	−6.94	−30.24
$\hat{\lambda}$	−63.28	−91.79	−56.18	−98.91

O estado North Dakota, observação 34, apresentou a menor taxa de $DESEMP$ e $INSEG$, sendo que a exclusão dessa observação diminui de forma considerável as estimativas de $\hat{\beta}_6$ e $\hat{\beta}_7$ em respectivamente -51.17% e -39.64%. Em relação ao submodelo da precisão, a exclusão da observação 34 diminui as estimativas de $\hat{\gamma}_0$ e $\hat{\gamma}_1$ em respectivamente -76.35% e -52.55%, influenciando negativamente a precisão. O estado South Dakota, observação 41, se destacou por apresentar uma das menores taxas de $INAT$, além das maiores taxas de $DESEMP$ e $DESCOB$. A exclusão da observação 41 não causou grandes variações nas estimativas dos parâmetros. O maior impacto ocorreu na estimativa de $\hat{\beta}_6$, aumentando seu valor em 12.71%.

Por fim, a exclusão das três observações causam um maior impacto nas estimativas dos parâmetros referentes ao submodelo da precisão, a saber, $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ e $\hat{\gamma}_3$, reduzindo em respectivamente -61.67%, -74.77%, -35.03% e -30.24% a estimativa dos parâmetros. Portanto, o estado North Dakota, observação 34, foi o que mais contribui com a variação percentual das estimativas dos parâmetros nos submodelos da média e da precisão. Temos ainda a variação do λ, que representa o grau de heterogeneidade da precisão dos dados, sendo definido como a razão $\max(\phi_t)/\min(\phi_t)$. Nas suas estimativas para os diferentes casos, ocorre redução de seus valores, refletindo assim a intensidade de não-constância da precisão.

Um dos objetivos desse estudo é estimar o impacto exercido pela inatividade física na proporção de adultos obesos nos diferentes estados. Como apresentado em Cribari-Neto e Souza (2013) o impacto pode ser obtido da seguinte maneira:

$$\frac{\partial \mathbb{E}(y_t)}{\partial INAT_t} = \frac{\partial \mu_t}{\partial INAT_t}, \tag{3.8}$$

em que

$$\begin{aligned}\mu_t &= g^{-1}(\beta_0 + \beta_1 INAT_t + \beta_2 VEGETt + \beta_3 FUM_t + \beta_4 DESEMP_t \\ &+ \beta_5 INSEG_t + \beta_6 BST_t + \beta_7 INT1_t).\end{aligned} \tag{3.9}$$

com $t = 1, \ldots, 50$. Considerando que a função de ligação para a média é cloglog, o impacto pode ser expresso como:

$$\begin{aligned}\frac{\partial \mathbb{E}(y_t)}{\partial INAT_t} &= \exp[-\exp(\beta_0 + \beta_1 INAT_t + \beta_2 VEGET_t + \beta_3 FUM_t \\ &+ \beta_4 DESEMP_t + \beta_5 INSEG_t + \beta_6 BST_t + \beta_7 INT1_t)] \\ &\times \exp(\beta_0 + \beta_1 INAT_t + \beta_2 VEGET_t + \beta_3 FUM_t + \beta_4 DESEMP_t \\ &+ \beta_5 INSEG_t + \beta_6 BST_t + \beta_7 INT1_t) \times \beta_1.\end{aligned} \tag{3.10}$$

para $t = 1, \ldots, 50$. Dessa forma, considerou-se o cenário em que as variáveis $VEGET$, FUM, $DESEMP$, $INSEG$, BST e $INT1$ estão fixadas no primeiro, segundo e terceiro quartis. A Figura 3.4 apresenta o impacto exercido pela inatividade física sobre a proporção de adultos obesos nos diferentes estados. Neste cenário é possível notar que o impacto da variável $INAT$ é positivo e assume forma acelerada para valores menores que 0.85. Para valores de $INAT > 0.85$ o efeito continua sendo positivo, contudo apresentando um crescimento mais lento. Notamos também que as curvas de impacto para o segundo e terceiro quartis não apresentam grandes diferenças, contudo quando as covariáveis estão fixadas no primeiro quartil a proporção de obesos é relativamente menor que nos demais quartis.

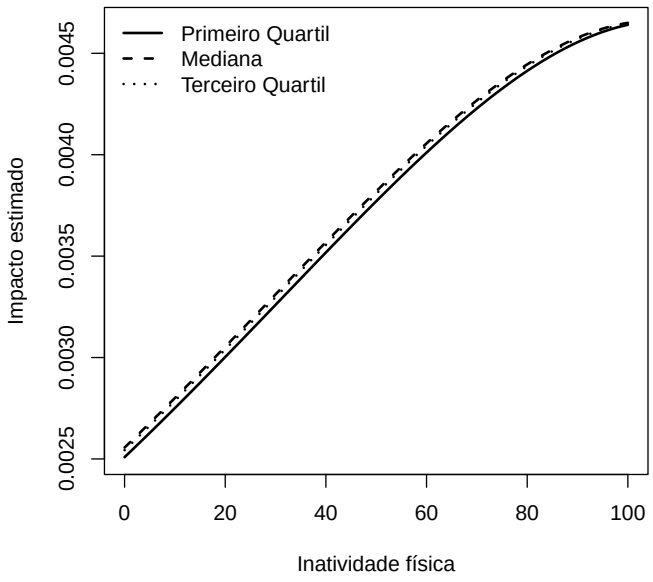


Figura 3.4: Impacto da porcentagem de adultos fisicamente inativos sobre a proporção de adultos obesos fixando-se a demais variáveis no primeiro, segundo e terceiro quartis.

3.5 Conclusões

Neste capítulo, através do modelo de regressão beta com dispersão variável, modelamos a proporção de adultos obesos e verificamos através do teste de Wald que as covariáveis inatividade física, consumo de vegetais menos de uma vez por dia, o hábito de fumar e a taxa de insegurança alimentar estavam relacionadas de forma positiva com a proporção média de adultos obesos. Já as taxas de desempregados e o escore de bem-estar mostraram-se relacionados negativamente com a resposta média. A modelagem do submodelo da precisão permitiu constatar que os estados que apresentavam maiores taxas de descobertos quanto ao seguro de saúde e maiores taxas de consumo de vegetais, apresentavam respostas mais precisas, ou seja, menos dispersas. Estimamos ainda o impacto exercido pela taxa de inatividade física sobre a proporção média de adultos obesos nos Estados Unidos. Os resultados revelaram que o efeito desse impacto é positivo e apresenta uma forma acelerada para valores de inatividade física menores do que 0.85, considerando que as demais variáveis foram fixadas em um determinado quartil.

De modo geral, concluímos que o ajuste obtido por meio do modelo de regressão beta com dispersão variável se mostrou uma ferramenta bastante útil para avaliar a proporção de adultos obesos nos Estados Unidos. Verificamos que os resultados encontrados se mostraram coerentes aos obtidos por outros autores em seus estudos, o que mostra a adequabilidade deste modelo de regressão para analisar dados do tipo proporção. Adicionalmente, o mesmo permitiu modelar a variabilidade dos dados, que é um artifício que permite melhorar os resultados inferenciais. Por fim, foi possível ainda analisar o impacto individual de uma determinada variável, que se mostrou relevante durante o estudo, na variável resposta, permitindo assim ampliar as conclusões a respeito do tema.

Capítulo 4

Erros de especificação no modelo de regressão beta com dispersão variável

4.1 Introdução

A análise de regressão é uma das técnicas estatísticas mais utilizadas sendo útil para investigar o comportamento de uma variável aleatória de interesse (variável dependente) quando o mesmo é influenciado por um conjunto de outras variáveis (variáveis independentes). Um dos modelos mais utilizados em análises empíricas é o modelo de regressão normal linear. Porém, o mesmo torna-se inapropriado quando a variável resposta assume valores pertencentes a um intervalo limitado na reta, tais como taxas e proporções contínuas. Dados desta natureza usualmente se distribuem assimetricamente, não sendo adequado o uso do modelo de regressão normal linear (CRIBARI-NETO; ZELEIS, 2010; KIESCHNICK; MCCULLOUCH, 2003).

Ferrari e Cribari–Neto (2004) propuseram um modelo de regressão que é de ampla utilidade para modelar variáveis pertencentes ao intervalo contínuo $(0, 1)$. O modelo de regressão beta proposto por estes autores assume que a variável resposta possui distribuição beta e que sua média é relacionada a um preditor linear por meio de uma função de ligação, preditor este que envolve covariáveis e parâmetros de regressão desconhecidos. Este modelo também é indexado por um parâmetro de dispersão, que neste caso, é constante ao longo das observações. Contudo, Simas et al. (2010) apresentaram uma extensão do modelo proposto por Ferrari e Cribari–Neto (2004), denominado modelo de regressão beta com dispersão variável. Nesta abordagem, o parâmetro de dispersão varia ao longo das observações, sendo modelado também por uma estrutura de regressão que contém covariáveis, parâmetros desconhecidos e uma função de ligação. Segundo Espinheira et al. (2008b), a classe de modelos de regressão beta é similar em muitos aspectos à classe dos modelos lineares generalizados (MCCULLACH; NELDER, 1989).

Diversas aplicações do modelo de regressão beta podem ser encontradas na literatura. Oliveira e Souza (2016), por exemplo, utilizaram este modelo para investigar a proporção de crianças obesas beneficiadas pelo Programa Bolsa Família nas regiões do Brasil. Almeida Junior e Souza (2015) avaliaram o impacto exercido pelo Programa Bolsa Família nas eleições presidenciais do ano de 2010, enquanto que Silva e Souza (2014) tiveram por finalidade modelar a taxa de analfabetismo nos municípios do estado da Paraíba neste mesmo ano. Sant'Anna e Caten (2010) modelaram a fração ou proporção de itens não conformes às especificações de um processo industrial com enfoque no modelo de regressão

beta e no modelo linear generalizado, por outro lado, Pinto et al. (2011) fizeram uso destes mesmos modelos de regressão para um estudo relacionado à infartos em pacientes. Outros exemplos de aplicações podem ser encontradas em Souza e Cribari–Neto (2015) e Cribari–Neto e Pereira (2013).

Ao se realizar qualquer análise de regressão não é possível saber de fato se o modelo estimado retrata adequadamente a realidade do fenômeno em estudo. Segundo Pereira e Cribari–Neto (2014), caso a especificação do modelo escolhido esteja incorreta, inferências imprecisas podem vir a ocorrer. No modelo de regressão beta com dispersão variável a escolha das funções de ligação e das variáveis independentes são processos tipicamente necessários no ajuste de um determinado modelo, e que podem ser auxiliados por conclusões de estudos anteriores. Porém, em termos práticos, é comum se cometer erros de especificação neste processo.

Neste contexto, alguns autores estudaram diferentes formas de especificações na classe de modelos de regressão beta. Bayer e Cribari–Neto (2017) exploraram o tema de seleção das covariáveis importantes nas estruturas de regressão dos submodelos da média e da dispersão. Andrade (2007) realizou um extenso estudo de simulação com o objetivo de avaliar o impacto da especificação incorreta da função de ligação da média e comparou, através de uma aplicação prática, os resultados obtidos através do uso de diferentes funções de ligação. Lima (2007) propôs um teste de erro de especificação para modelos de regressão beta baseado no teste *RESET* (RAMSEY, 1969), e concluiu através de simulações que o mesmo é útil para detecção do uso de função de ligação incorreta bem como de não-linearidades no preditor linear. Canterle et al. (2015) abordaram o problema da má especificação na função de ligação do submodelo da dispersão e verificaram que a incorreta especificação desta função de ligação tem uma influência considerável nas inferências do modelo, incluindo implicações diretas na eficiência dos estimadores dos parâmetros da média.

Loose et al. (2014) abordaram o desempenho dos estimadores pontuais e intervalares no modelo de regressão beta com dispersão variável e, através de simulações de Monte Carlo, confirmaram a consistência destes estimadores. Contudo, observaram que os estimadores que modelam a precisão (inverso da dispersão) são consideravelmente mais viesados do que os que modelam a média, indicando uma necessidade de maior atenção na modelagem da estrutura de regressão deste parâmetro. Considerando a dificuldade em se modelar esta estrutura e que na prática é comum se cometer erros de especificação, Cribari–Neto e Souza (2012) propuseram uma nova abordagem em modelos de regressão beta como forma de solucionar estes problemas. Esta abordagem é baseada em estimadores do tipo sanduíche para casos em que a estrutura de regressão para o parâmetro de dispersão é negligenciada. Os autores concluíram que as inferências considerando esta metodologia são precisas mesmo sob dispersão variável, o que indica uma possível solução para os erros de especificação que usualmente são cometidos nesta estrutura.

Neste contexto de erros de especificação, o nosso objetivo é avaliar o efeito dos mesmos nas inferências do modelo de regressão beta com dispersão variável. Um estudo de simulação considerando diferentes cenários foi realizado com este propósito. Nestas simulações, a variável resposta foi gerada com distribuição beta assumindo covariáveis e funções de ligação conhecidas, o modelo foi então ajustado considerando a especificação

correta e incorreta. Em particular, seis tipos de erros de especificação foram avaliados, englobando tanto erros nos preditores quanto nas funções de ligação das duas estruturas de regressão. Para avaliar o efeito destes erros, consideramos as taxas de rejeição e as taxas de cobertura em relação a um dos parâmetros do submodelo da média (μ). Computamos ainda algumas medidas considerando as estimativas para as respostas médias, a saber: o viés relativo médio e o erro quadrático médio. Por fim, realizamos uma aplicação a dados reais.

O presente capítulo encontra-se dividido em cinco seções. A Seção 2 apresenta o modelo de regressão beta com dispersão variável. Os resultados numéricos e as discussões são apresentados na Seção 3. Na Seção 4 uma aplicação a dados reais é realizada. Por último, na Seção 5 são apresentadas as conclusões e considerações finais.

4.2 Modelo de regressão beta

O modelo de regressão beta, introduzido por Ferrari e Cribari–Neto (2004), é comumente utilizado para modelar variáveis que assumem valores no intervalo $(0, 1)$, a exemplo de taxas e proporções. Almeida Junior e Souza (2015), Oliveira e Souza (2016), Smithson e Verkuilen (2006) e Souza e Cribari–Neto (2015) utilizaram modelos de regressão para a situação em que a variável resposta segue distribuição beta. Em tais modelos assume-se que a resposta média é relacionada a um preditor linear através de uma função de ligação, preditor este que envolve covariáveis e parâmetros de regressão desconhecidos. Estes modelos também são indexados por um parâmetro de dispersão que em certas ocasiões podem variar ao longo das observações (CRIBARI–NETO; SOUZA, 2012, 2013; ESPINHEIRA et al., 2008a, 2008b; SILVA; SOUZA, 2014; SIMAS et al., 2010).

Para a definição do modelo de regressão beta, Ferrari e Cribari–Neto (2004) sugerem uma parametrização da distribuição beta em termos de sua média e um parâmetro de precisão. Com essa parametrização a função densidade da distribuição beta pode ser reescrita como

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\ 0<y<1, \tag{4.1}$$

em que $0 < \mu < 1$ e $\phi > 0$. Aqui, $\mathrm{E}(y) = \mu$ e $\mathrm{var}(y) = \frac{V(\mu)}{1+\phi}$, sendo $V(\mu) = \mu(1-\mu)$, a "função variância", μ é a média da variável resposta e ϕ pode ser interpretado como o parâmetro de precisão no sentido que, para um valor fixo de μ, quanto maior o valor de ϕ, menor a variância de y. O parâmetro de dispersão é obtido considerando ϕ^{-1}.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, em que cada y_t, $t = 1, \ldots, n$, segue a densidade apresentada em (4.1) com média μ_t e parâmetro de precisão ϕ_t sendo desconhecidos. O modelo de regressão beta assume que uma função da média μ_t pode ser igualada ao preditor linear η_t, sendo esta estrutura definida por

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t, \tag{4.2}$$

em que $\beta = (\beta_1, \ldots, \beta_k)^\top$ é um vetor de parâmetros de regressão desconhecidos ($\beta \in \mathbb{R}^k$), $x_{t1}, \ldots, x_{tk}$ são observações de k covariáveis e $g(\cdot)$ é denominada função de ligação. Portanto, $\mu_t = g^{-1}(\eta_t)$ e $\text{var}(y_t) = \mu_t(1-\mu_t)/(1+\phi)$, para $t = 1, \ldots, n$.

O modelo de regressão beta proposto por Ferrari e Cribari–Neto (2004) considera o parâmetro de precisão constante ao longo das observações. Porém, admitimos como em Simas et al. (2010) que o parâmetro de precisão é variável, sendo modelado através de uma estrutura de regressão que contém covariáveis, parâmetros de regressão desconhecidos e uma função de ligação, sendo esta estrutura definida por

$$h(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \vartheta_t, \tag{4.3}$$

em que $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1}, \ldots, z_{tq}$ são observações de q covariáveis ($k + q < n$) assumidas fixas e conhecidas, ϑ_t é o preditor linear e $h(\cdot)$ é uma função de ligação. Aqui, $\phi_t = h^{-1}(\vartheta_t)$, em que $t = 1, \ldots, n$.

As estimativas dos parâmetros são obtidas maximizando numericamente a função de log-verossimilhança através de um algoritmo de maximização não-linear. Usualmente, utiliza-se o método quasi-Newton BFGS (PRESS et al., 1992). A distribuição dos estimadores de máxima verossimilhança de β e γ, ditos $\widehat{\beta}$ e $\widehat{\gamma}$, é aproximadamente normal em grandes amostras. Esta aproximação pode ser usada na construção de intervalos de confiança e testes de hipóteses. Para maiores detalhes inferenciais e matriciais do vetor escore e da matriz de informação de Fisher, ver Simas et al. (2010).

4.2.1 Teste de Hipóteses e Intervalos de Confiança

De acordo com Ferrari e Cribari–Neto (2004) inferências em grandes amostras no modelo de regressão beta podem ser realizadas utilizando o teste de Wald (WALD, 1943). A estatística de teste para testar a hipótese nula $\mathcal{H}_0 : \beta_1 = \beta_1^{(0)}$ é dada por:

$$\omega = (\widehat{\beta}_1 - \beta_1^{(0)})^\top (\widehat{K}_{11}^{\beta\beta})^{-1} (\widehat{\beta}_1 - \beta_1^{(0)}), \tag{4.4}$$

em que $\widehat{K}_{11}^{\beta\beta}$ é igual $K_{11}^{\beta\beta}$ (matriz $m \times m$ obtida da inversa da matriz de informação de Fisher K^{-1}) avaliado no estimador de máxima verossimilhança irrestrito, e $\widehat{\beta}_1$ é o estimador de máxima verosimilhança de β_1. Sob fracas condições de regularidade e sob $\mathcal{H}_0$, $\omega \xrightarrow{D} \chi^2_m$. Em particular, para testar a significância do i-ésimo parâmetro de regressão (β_i), $i = 1, \ldots, k$, pode-se utilizar a raiz quadrada da estatística de Wald (teste z), isto é, $\widehat{\beta}_i / ep(\widehat{\beta}_i)$, onde $ep(\widehat{\beta}_i)$ é o erro padrão assintótico do estimador de máxima verossimilhança de $\widehat{\beta}_i$ obtido da inversa da matriz de informação de Fisher avaliada nas estimativas de máxima verossimilhança. A distribuição nula restrita da estatística de teste é normal padrão.

Um intervalo de confiança $(1-\alpha) \times 100\%$ para os parâmetros dos modelos, sendo $\theta = (\beta^\top, \gamma^\top)^\top$ o vetor de parâmetros, é dado por:

$$\widehat{\theta}_i \pm \Phi^{-1}(1-\alpha/2) ep(\widehat{\theta}_i), \tag{4.5}$$

em que Φ^{-1} é a função de distribuição acumulada de uma variável aleatória normal padrão, $ep(\widehat{\theta}_i)$ é o erro padrão para $\widehat{\theta}_i$ e α é o nível nominal do intervalo de confiança.

4.2.2 Funções de Ligação

As funções $g(\cdot)$ e $h(\cdot)$ são conhecidas como funções de ligação e existem muitas possibilidades para suas escolhas. Considerando o parâmetro da média temos que, $g(\mu_t) = \eta_t$, $t = 1, \ldots, n$, em que $g(\cdot)$ é estritamente monótona e duas vezes diferenciável, com domínio em $(0, 1)$ e imagem em $\mathbb{R}$. Portanto, $\mu_t = g^{-1}(\eta_t)$. Alguns exemplos destas funções de ligação são:

- logit: $g(\mu_t) = \log(\frac{\mu_t}{1-\mu_t})$; $\mu_t = \frac{\exp(\eta_t)}{1+\exp(\eta_t)}$;
- probit: $g(\mu_t) = \Phi^{-1}(\mu_t)$; em que $\Phi^{-1}(\cdot)$ é a função de distribuição acumulada de uma variável normal padrão; $\mu_t = \Phi(\eta_t)$;
- cloglog: $g(\mu_t) = \log\{-\log(1-\mu_t)\}$; $\mu_t = 1 - \exp\{-\exp(\eta_t)\}$;
- loglog: $g(\mu_t) = -\log\{-\log(\mu_t)\}$; $\mu_t = \exp\{-\exp(-\eta_t)\}$.

em que $t = 1, \ldots, n$. A Figura 4.1 apresenta o comportamento de η como função de μ para as funções de ligação citadas acima. É possível perceber, por exemplo, que as funções logit e probit tem comportamento parecido, que a função cloglog tem comportamento similar à logit para valores de μ próximos de 0 e a função de ligação loglog tem comportamento similar à logit para valores de μ próximos de 1.

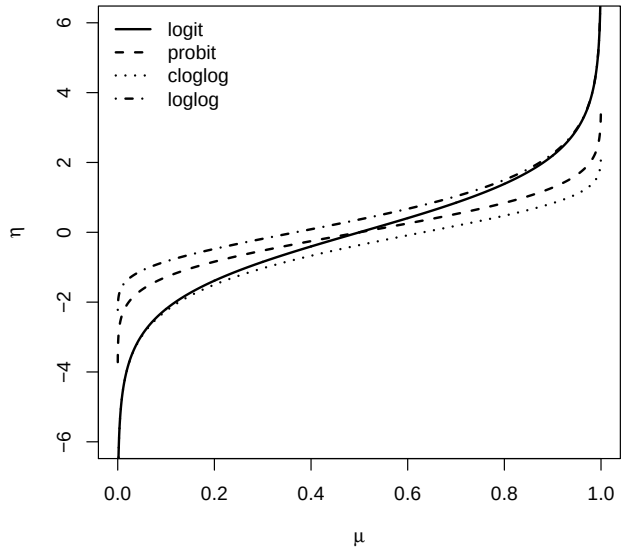


Figura 4.1: Gráfico de funções de ligação para a média.

Para o parâmetro de precisão temos que, $h(\phi_t) = \vartheta_t$, em que $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável que mapeia os pontos positivos da reta. Portanto, $\phi_t = h^{-1}(\vartheta_t)$. Alguns exemplos destas funções são:

- log: $h(\phi_t) = \log(\phi_t)$; $\phi_t = \exp\{\vartheta_t\}$;
- sqrt: $h(\phi_t) = \sqrt{\phi_t}$; $\phi_t = \vartheta_t^2$;
- identity: $h(\phi_t) = \phi_t$; $\phi_t = \vartheta_t$.

em que $t = 1, \ldots, n$. Para maiores detalhes sobre as funções de ligação ver McCullagh e Nelder (1989).

4.3 Avaliação numérica

Através de simulações de Monte Carlo nós avaliamos, em amostras de tamanho finito, o efeito de diferentes erros de especificação no modelo de regressão beta com dispersão variável. Os erros de especificação cometidos estão apresentados na Tabela 4.1 e exemplificam alguns erros que usualmente são cometidos em modelagens práticas, como a escolha incorreta de funções de ligação ou de covariáveis para compor as estruturas de regressão da média e/ou precisão. Estes erros de especificação foram avaliados em diferentes cenários, considerando tamanhos amostrais e intervalos de médias distintos. A implementação computacional foi desenvolvida no *software* estatístico `R` (KLEIBER; ZELEIS, 2008; R DEVELOPMENT CORE TEAM, 2014;) utilizando o pacote `betareg` (CRIBARI–NETO; ZELEIS, 2010).

O número de réplicas de Monte Carlo foi fixado em 10000. Para cada réplica de Monte Carlo foram geradas amostras aleatórias da variável aleatória y_t, $t = 1, \ldots, n$, com função de densidade dada em (4.1), parâmetro de média definido por $\mu_t = g^{-1}(\eta_t)$ e parâmetro de precisão definido por $\phi_t = h^{-1}(\vartheta_t)$, em que:

$$\begin{aligned} \eta_t &= \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}, \\ \vartheta_t &= \gamma_0 + \gamma_1 z_{1t}, \end{aligned} \tag{4.6}$$

para $t = 1, \ldots, n$. As variáveis independentes foram geradas a partir da distribuição uniforme $(0, 1)$. Na geração dos dados, consideramos o modelo corretamente especificado (MCE) definido pela estrutura apresentada acima, e consideramos ainda $g(\cdot)$ a função de ligação logit e $h(\cdot)$ a função de ligação sqrt. Medimos a não-constância da precisão dos dados através da quantidade $\lambda = \frac{\max(\phi_t)}{\min(\phi_t)}$, para $t = 1, \ldots, n$. Note que $\lambda = 1$ indica que a precisão é constante para todas as observações. No nosso estudo $\lambda \approx 13$.

Com o objetivo de avaliar os efeitos dos erros de especificação em diferentes cenários, consideramos três intervalos para as médias na geração dos dados. No primeiro cenário, utilizamos $\beta_0 = -1.9$, $\beta_1 = 1.5$ e $\beta_2 = 0.0$ como verdadeiros valores dos parâmetros, que conduziram a valores de médias próximos à 0, mais precisamente $\mu_t \in [0.1374; 0.3764]$. No segundo cenário, os valores para os parâmetros foram $\beta_0 = -0.5$, $\beta_1 = 1.2$ e $\beta_2 = 0.0$, que conduziram a $\mu_t \in [0.3894; 0.6493]$, isto é, médias próximas à 0.5. No terceiro cenário,

Tabela 4.1: Descrição dos cenários considerados no estudo de simulação de Monte Carlo.

Especificação dos modelos	**Geração**	
	$\boldsymbol{\mu}$	$\boldsymbol{\phi}$
Modelo Corretamente Especificado - MCE	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\text{sqrt}(\phi_t) = \gamma_0 + \gamma_1 z_{1t}$
Especificação dos modelos	**Estimação**	
	$\boldsymbol{\mu}$	$\boldsymbol{\phi}$
Erro no preditor linear da média e nas funções de ligação da média e precisão - E1	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}^2$	$\log(\phi_t) = \gamma_0 + \gamma_1 z_{1t}$
Erro no preditor linear da precisão e na função de ligação da precisão - E2	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\log(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Erros nos preditores lineares da média e da precisão e nas funções de ligação da média e precisão - E3	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}^2$	$\log(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Erro no preditor linear da precisão e na função de ligação da média - E4	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\text{sqrt}(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Erros nos preditores lineares da média e da precisão - E5	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}^2$	$\text{sqrt}(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Precisão fixa - E6	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\text{sqrt}(\phi_t) = \gamma_0$

utilizamos $\beta_0 = 2.3$, $\beta_1 = -1.8$ e $\beta_2 = 0.0$, que neste caso conduziu a $\mu_t \in [0.6515; 0.9024]$, ou seja, valores de médias próximas à 1. No caso da precisão, os verdadeiros valores para os parâmetros foram $\gamma_0 = 1.0$ e $\gamma_1 = 7.9$, que produziram valores de ϕ_t no intervalo $[5.8420; 76.6020]$, o que corresponde a um cenário de baixa precisão, ou ainda, de alta dispersão.

Consideramos ainda três diferentes tamanhos amostrais, $n = 25, 50, 100$, sendo que são geradas 25 observações das variáveis x_1, x_2, z_1 e z_2, que são replicadas duas e quatro vezes, respectivamente, para os tamanhos amostrais $n = 50, 100$. Este procedimento de replicação de valores assegura que o grau de heterogeneidade na precisão dos dados mantenha-se constante à medida em que se aumenta o tamanho amostral.

Na avaliação dos efeitos dos erros de especificação nas inferências do modelo de regressão beta foram computadas as taxas de rejeição sob a hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ e os intervalos de confiança com nível nominal de confiança de 95% para o parâmetro β_2. A partir de 10000 intervalos de confiança foram obtidas as taxas de cobertura para β_2. A taxa de cobertura representa a proporção de vezes em que o intervalo de confiança conteve o parâmetro, sendo esperada uma taxa em torno de 95%. Em relação as estimativas das médias μ_t, foram avaliados os vieses relativos médios (VRm) e o erro quadrático médio (EQM), sendo definidos por:

$$VRm = \frac{1}{n} \sum_{t=1}^{n} \frac{|\hat{\mu}_t - \mu_t|}{\mu_t}, \tag{4.7}$$

$$EQM = \frac{1}{n}\sum_{t=1}^{n} E[(\hat{\mu}_t - \mu_t)^2], \tag{4.8}$$

em que $t = 1, \ldots, n$, $\hat{\mu}_t$ é a estimativa para as médias considerando as n observações e E denota a operação de valor esperado sob a quantidade $(\hat{\mu}_t - \mu_t)^2$. Para essas medidas espera-se valores próximos de zero com o aumento do tamanho amostral.

Primeiramente, nosso interesse consiste em avaliar as taxas de rejeição (tamanho dos testes z) sob a hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ versus $\mathcal{H}_1 : \beta_2 \neq 0$ considerando seis tipos de erros de especificação (ver Tabela 4.1). Na Tabela 4.2 são apresentados os resultados das taxas de rejeição aos níveis nominais de 10%, 5% e 1% considerando diferentes intervalos para as médias μ_t. As principais conclusões serão resumidas a seguir. Em primeiro lugar, considerando o modelo corretamente especificado, é possível observar que as taxas de rejeição estão acima dos níveis nominais considerados para os três intervalos das médias e para os diferentes tamanhos amostrais. Vale salientar que as mesmas foram ainda mais distantes dos níveis nominais para valores de médias próximas de 0. Como ilustração, considerando o cenário em que as médias estão variando no intervalo $[0.1374; 0.3764]$, $n = 50$ e $\alpha = 10\%$, a taxa de rejeição é de 13.07%. No cenário 2, de médias entre $[0.3894; 0.6493]$, essa mesma taxa é de 12.26%. Já para o cenário 3, cujo intervalo de médias é entre $[0.6515; 0.9024]$ a taxa é de 12.65%.

Segundo, as taxas de rejeição tendem a ser maiores quando há erro no preditor da precisão (E2, E3, E4 e E5) comparado ao ajuste em que não há esse erro (E1) ou quando o modelo é estimado com precisão fixa (E6). Exemplificando, considere o cenário 2, com médias variando próximas à 0.5, $n = 50$ e $\alpha = 5\%$, a taxa de rejeição do modelo em que há erros nas funções de ligação da média e da precisão e no preditor da média (E1) é de 6.71%, enquanto que no modelo em que há esses mesmos erros acrescido ao erro no preditor da precisão (E3), essa taxa é de 10.24%. Considerando o ajuste estimado com precisão fixa (E6), a taxa de rejeição associada é de 7.74%.

Terceiro, no caso em que o modelo foi estimado cometendo erros no preditor da precisão e em uma das funções de ligação, nossos resultados mostraram que no geral, é mais grave se cometer erros no preditor da precisão e na função de ligação da média do que no preditor e na função de ligação do parâmetro de precisão, considerando valores de médias variando próximos de 0 e de 0.5 (cenários 1 e 2). Já para valores de médias próximos de 1 (cenário 3), os resultados se inverteram, as taxas maiores se associaram ao erro no preditor e na função de ligação do parâmetro de precisão. Como exemplo considere o cenário 1, $n = 100$ e $\alpha = 5\%$, a taxa de rejeição para o modelo em que se cometem erros no preditor e na função de ligação da precisão (E2) é de 9.07%, enquanto que para o modelo estimado com erros no preditor da precisão e na função de ligação da média (E4), essa mesma taxa é de 9.73%. No cenário 3, considerando o mesmo tamanho amostral e nível de significância, essas taxas foram de 9.13% e 8.12%, respectivamente.

Quarto, considerando os erros de especificação cometidos, o modelo estimado que no geral apresentou taxas mais distantes dos níveis nominais, para os cenários 1 e 2, foi o modelo com erros nas funções de ligação e nos preditores dos parâmetros da média e da precisão (E3). Por exemplo, para valores de médias próximos de 0.5, $n = 50$ e $\alpha = 5\%$, o tamanho do teste baseado no modelo estimado com os erros de especificação descritos

Tabela 4.2: Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$.

Cenário 1	$\mu_t \in [\mathbf{0.1374; 0.3764}]$								
	$n = 25$			$n = 50$			$n = 100$		
	10%	**5%**	**1%**	**10%**	**5%**	**1%**	**10%**	**5%**	**1%**
Modelo corretamente especificado - MCE	16.10	9.70	5.35	13.07	7.54	3.40	11.52	6.17	2.75
Erro no pred. linear da média e nas funções de lig. - E1	17.37	10.58	6.13	13.12	7.31	3.49	11.18	6.19	2.67
Erro no pred. linear e na função de lig. da precisão - E2	21.15	13.50	7.76	16.89	10.10	5.22	15.05	9.07	4.66
Erros nos pred. lineares e nas funções de lig. - E3	22.07	14.83	8.89	17.69	11.01	6.05	16.98	10.40	5.38
Erro no pred. linear da precisão e na função de lig. da média - E4	21.18	13.68	7.77	17.15	10.42	5.44	16.06	9.73	4.91
Erros nos pred. lineares da média e da precisão - E5	21.90	14.18	8.54	17.24	10.69	5.82	16.22	9.99	5.09
Precisão Fixa - E6	17.14	10.00	5.21	14.91	8.53	4.12	13.70	7.93	3.76
Cenário 2	$\mu_t \in [\mathbf{0.3894; 0.6493}]$								
	$n = 25$			$n = 50$			$n = 100$		
	10%	**5%**	**1%**	**10%**	**5%**	**1%**	**10%**	**5%**	**1%**
Modelo corretamente especificado - MCE	15.88	9.56	5.11	12.26	6.73	3.12	11.36	6.32	2.76
Erro no pred. linear da média e nas funções de lig. - E1	16.13	10.09	5.58	12.00	6.71	3.02	11.26	6.38	2.59
Erro no pred. linear e na função de lig. da precisão - E2	20.91	13.90	8.12	15.42	9.19	4.64	14.21	8.16	3.97
Erros nos pred. lineares e nas funções de lig. - E3	21.19	14.31	8.53	16.04	10.24	5.34	15.31	8.98	4.65
Erro no pred. linear da precisão e na função de lig. da média - E4	20.73	13.72	8.00	15.80	9.39	4.78	14.65	8.41	4.24
Erros nos pred. lineares da média e da precisão - E5	20.32	13.51	7.98	15.80	9.67	5.03	14.52	8.62	4.38
Precisão Fixa - E6	16.06	10.01	5.11	13.67	7.74	3.69	13.40	7.74	3.58
Cenário 3	$\mu_t \in [\mathbf{0.6515; 0.9024}]$								
	$n = 25$			$n = 50$			$n = 100$		
	10%	**5%**	**1%**	**10%**	**5%**	**1%**	**10%**	**5%**	**1%**
Modelo corretamente especificado - MCE	15.87	9.36	5.22	12.65	7.13	3.61	10.70	5.87	2.51
Erro no pred. linear da média e nas funções de lig. - E1	17.10	10.62	5.97	13.05	7.36	3.70	11.76	6.03	2.70
Erro no pred. linear e na função de lig. da precisão - E2	20.40	13.37	7.96	17.38	10.62	5.26	15.30	9.13	4.50
Erros nos pred. lineares e nas funções de lig. - E3	20.88	13.26	7.69	16.81	9.96	5.10	14.67	8.45	4.14
Erro no pred. linear da precisão e na função de lig. da média - E4	19.84	12.70	7.47	16.44	9.92	4.80	13.99	8.12	3.89
Erros nos pred. lineares da média e da precisão - E5	21.46	14.29	8.60	18.40	11.66	6.12	16.43	9.69	5.13
Precisão Fixa - E6	16.47	10.06	5.52	15.21	8.93	4.28	13.66	7.91	3.84

acima rejeita a hipótese nula 10.24% das vezes, o que vale ressaltar que é mais do que o dobro do nível nominal considerado. Em contrapartida, para o cenário 3, o modelo com erros nos preditores da média e da precisão (E5) apresentou as maiores taxas de rejeição comparadas aos demais modelos estimados.

Quinto, as taxas de rejeição do modelo estimado com precisão fixa (E6), considerando os tamanhos amostrais $n = 50$ e $n = 100$, tenderam a serem maiores do que as taxas do modelo estimado com erros nas funções de ligação da média e da precisão e no preditor linear da média (E1). Isso significa que, para tamanhos amostrais maiores omitir a estrutura de regressão do parâmetro de precisão pode ser mais grave do que errar na especificação da média e na função de ligação do parâmetro de precisão, resultado este que não se verifica para amostras pequenas ($n = 25$). Para ilustrar essa conclusão considere médias variando próximas de 1 e $\alpha = 10\%$, as taxas de rejeição do modelo estimado com precisão fixa considerando os três tamanhos amostrais foram de 16.47%, 15.21% e 13.66%, enquanto que no modelo com erros nas funções de ligação e no preditor da média, essas taxas foram de 17.10%, 13.05% e 11.76%, respectivamente.

Na Tabela 4.3 encontram-se os resultados numéricos da avaliação das taxas de cobertura para o parâmetro β_2. Nota-se que os resultados estão dentro do esperado, pois as taxas se aproximam de 95% com o aumento do tamanho amostral, considerando os erros

Tabela 4.3: Resultados das taxas de cobertura para β_2

Cenário 1	$\boldsymbol{\mu_t \in [0.1374; 0.3764]}$		
	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$
Modelo corretamente especificado - MCE	90.30	92.46	93.83
Erro no pred. linear da média e nas funções de lig. - E1	89.42	92.69	93.81
Erro no pred. linear e na função de lig. da precisão - E2	86.50	89.90	90.93
Erros nos pred. lineares e nas funções de lig. - E3	85.17	88.99	89.60
Erro no pred. linear da precisão e na função de lig. da média - E4	86.32	89.58	90.27
Erros nos pred. lineares da média e da precisão - E5	85.82	89.31	90.01
Precisão Fixa - E6	90.00	91.47	92.07
Cenário 2	$\boldsymbol{\mu_t \in [0.3894; 0.6493]}$		
	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$
Modelo corretamente especificado - MCE	90.44	93.27	93.68
Erro no pred. linear da média e nas funções de lig. - E1	89.91	92.29	93.62
Erro no pred. linear e na função de lig. da precisão - E2	86.10	90.81	91.84
Erros nos pred. lineares e nas funções de lig. - E3	85.69	89.76	91.02
Erro no pred. linear da precisão e na função de lig. da média - E4	86.28	90.61	91.59
Erros nos pred. lineares da média e da precisão - E5	86.49	90.33	91.38
Precisão Fixa - E6	89.99	92.26	92.26
Cenário 3	$\boldsymbol{\mu_t \in [0.6515; 0.9024]}$		
	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$
Modelo corretamente especificado - MCE	90.64	92.87	94.13
Erro no pred. linear da média e nas funções de lig. - E1	89.38	92.34	93.97
Erro no pred. linear e na função de lig. da precisão - E2	86.63	89.38	90.87
Erros nos pred. lineares e nas funções de lig. - E3	86.74	90.04	91.55
Erro no pred. linear da precisão e na função de lig. da média - E4	87.30	90.08	91.88
Erros nos pred. lineares da média e da precisão - E5	85.71	88.34	90.31
Precisão Fixa - E6	89.94	91.07	92.09

de especificação cometidos e os diferentes intervalos para as médias. Com base nesses resultados, temos que o modelo estimado com mais erros de especificação (E3) apresentou as taxas de cobertura mais distantes de 95% para os cenários 1 e 2, enquanto que para o cenário 3, o modelo estimado com erros nos preditores da média e da precisão (E5) apresentou as menores taxas de cobertura comparado aos demais modelos.

Assim como ocorreram com as taxas de rejeição, nos modelos em que há o erro no preditor da precisão (E2, E3, E4 e E5) há uma tendência a apresentar resultados mais distantes do esperado comparado ao caso em que não há esse erro (E1) ou quando o modelo é estimado desconsiderando a estrutura de regressão para o parâmetro de precisão (E6). Exemplificando, considere valores de médias próximos de 0 e $n = 50$, a taxa de cobertura para o parâmetro β_2 no modelo com erros nas funções de ligação e no preditor da média (E1) é de 92.69%, enquanto que no modelo estimado com esses mesmos erros acrecido ao erro no preditor da precisão (E3), a taxa é de 88.99%. Já para o modelo estimado com precisão fixa (E6), a mesma foi de 91.47%.

As taxas de cobertura para o modelo estimado com erros no preditor da precisão e na função de ligação da média (E4) tenderam a serem mais distantes do valor esperado do que no modelo estimado com erros no preditor e na função de ligação do parâmetro precisão (E2) considerando os cenários 1 e 2. Resultado que reforça o fato de que, errar no preditor da precisão e na função de ligação da média é mais grave do que errar apenas na especificação do parâmetro de precisão (função de ligação e preditor linear) considerando valores de médias variando até próximos de 0.65.

Confrontando os resultados obtidos em relação apenas aos diferentes intervalos de médias considerados, temos que em geral, as taxas de cobertura tenderam a serem maiores para os cenários com médias variando próximos de 0.5 e de 1. Para ilustrar este resultado, considere o tamanho amostral $n = 100$ e caso em que se cometem erros nos preditores lineares e nas funções de ligação dos parâmetros da média e da precisão (E3), as taxas de cobertura considerando os cenários 1, 2 e 3, foram 89.60%, 91.02% e 91.55%, respectivamente.

A Tabela 4.4 apresenta os vieses relativos médios e o erro quadrático médio das estimativas das médias. Através de sua análise, pode-se observar que em todos os casos, como já era esperado devido a propriedade de consistência dos estimadores de máxima verossimilhança, com o aumento do tamanho amostral os valores tenderam a zero. Além do mais, o modelo estimado corretamente no geral apresentou os menores valores para essas medidas.

Dos intervalos para as médias utilizados na geração dos dados, o cenário 3, de médias próximas de 1, apresentou menores valores comparado aos outros dois cenários. Mais uma vez, modelos estimados com erro no preditor linear do parâmetro da precisão apresentaram resultados mais distantes do esperado. Vale destacar, que no caso do viés relativo médio e do erro quadrático médio, o modelo estimado desconsiderando a estrutura de regressão do parâmetro de precisão apresentou em alguns casos, os resultados mais distantes de zero comparado aos demais modelos aqui estudados. Exemplificando este último resultado, considere $n = 50$ e médias variando próximas de 0, o viés relativo médio e o erro quadrático médio do modelo estimado com precisão fixa (E6) foram de 0.0750 e 0.0113 respectivamente, enquanto que no modelo estimado com erros no preditor linear da média

Tabela 4.4: Vieses Relativos Médios (VRm) e Erro Quadrático Médio (EQM) para o estimador das médias.

Cenário 1	$\mu_t \in [\mathbf{0.1374; 0.3764}]$					
	$\boldsymbol{VRm}$			$\boldsymbol{EQM}$		
	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$
Modelo corretamente especificado - MCE	0.0839	0.0589	0.0413	0.0128	0.0089	0.0070
Erro no pred. linear da média e nas funções de lig. - E1	0.0857	0.0597	0.0418	0.0132	0.0091	0.0071
Erro no pred. linear e na função de lig. da precisão - E2	0.1064	0.0739	0.0522	0.0173	0.0110	0.0081
Erros nos pred. lineares e nas funções de lig. - E3	0.1068	0.0743	0.0525	0.0175	0.0111	0.0082
Erro no pred. linear da precisão e na função de lig. da média - E4	0.1061	0.0740	0.0524	0.0173	0.0111	0.0081
Erros nos pred. lineares da média e da precisão - E5	0.1067	0.0743	0.0524	0.0174	0.0111	0.0081
Precisão Fixa - E6	0.1052	0.0750	0.0532	0.0172	0.0113	0.0083
Cenário 2	$\mu_t \in [\mathbf{0.3894; 0.6493}]$					
	$\boldsymbol{VRm}$			$\boldsymbol{EQM}$		
	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$
Modelo corretamente especificado - MCE	0.0450	0.0313	0.0220	0.0093	0.0078	0.0071
Erro no pred. linear da média e nas funções de lig. - E1	0.0459	0.0318	0.0224	0.0094	0.0079	0.0071
Erro no pred. linear e na função de lig. da precisão - E2	0.0570	0.0390	0.0276	0.0109	0.0085	0.0075
Erros nos pred. lineares e nas funções de lig. - E3	0.0573	0.0394	0.0281	0.0110	0.0086	0.0075
Erro no pred. linear da precisão e na função de lig. da média - E4	0.0570	0.0393	0.0280	0.0109	0.0086	0.0075
Erros nos pred. lineares da média e da precisão - E5	0.0570	0.0392	0.0278	0.0109	0.0086	0.0075
Precisão Fixa - E6	0.0565	0.0397	0.0284	0.0108	0.0086	0.0075
Cenário 3	$\mu_t \in [\mathbf{0.6515; 0.9024}]$					
	$\boldsymbol{VRm}$			$\boldsymbol{EQM}$		
	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$	$\boldsymbol{n = 25}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$
Modelo corretamente especificado - MCE	0.0224	0.0156	0.0109	0.0067	0.0062	0.0059
Erro no pred. linear da média e nas funções de lig. - E1	0.0228	0.0162	0.0117	0.0065	0.0059	0.0057
Erro no pred. linear e na função de lig. da precisão - E2	0.0290	0.0203	0.0142	0.0072	0.0064	0.0060
Erros nos pred. lineares e nas funções de lig. - E3	0.0286	0.0204	0.0147	0.0069	0.0061	0.0058
Erro no pred. linear da precisão e na função de lig. da média - E4	0.0284	0.0203	0.0146	0.0069	0.0061	0.0057
Erros nos pred. lineares da média e da precisão - E5	0.0291	0.0205	0.0143	0.0073	0.0064	0.0060
Precisão Fixa - E6	0.0284	0.0203	0.0143	0.0074	0.0066	0.0061

e nas funções de ligação da média e da precisão (E1), por exemplo, estes valores foram de 0.0597 e 0.0091, respectivamente.

Deste modo, através dos resultados das simulações verificamos que os erros de especificação que envolviam o preditor linear da estrutura de regressão do parâmetro de precisão tiveram uma influência considerável nas inferências do modelo de regressão beta com dispersão variável. De acordo com Bayer e Cribari–Neto (2017), a modelagem do parâmetro de dispersão (recíproco da precisão) pode ser de interesse na identificação de fontes de variabilidade no fenômeno em estudo. Além disso, Canterle et al. (2015) também concluem que a correta modelagem deste parâmetro por meio da seleção de covariáveis e funções de ligação têm influência direta na eficiência dos estimadores dos parâmetros de regressão da média. Estes fatos demonstram a real importância de se modelar corretamente esta estrutura de regressão. Porém, Souza et al. (2016) afirmam que modelar a variabilidade é um processo tipicamente mais complicado do que modelar a média da variável resposta, e como consequência disso, torna-se mais fácil se cometer erros de especificação nesta estrutura. Baseado nisso, Cribari–Neto e Souza (2012) propuseram estimadores do tipo sanduíche para o modelo de regressão beta, sendo esta uma alternativa para se realizar inferências precisas no submodelo da média sem necessariamente ter que se modelar a dispersão. As inferências obtidas utilizando esta abordagem são precisas mesmo sob dispersão variável. Uma outra alternativa para se evitar erros de especificação caso seja importante para o estudo modelar a variabilidade, é a aplicação do teste *RESET* adaptado para a classe de modelos de regressão beta (LIMA, 2007; OLIVEIRA, 2013). Segundo os autores, este teste é útil na identificação de diversos tipos de erros de especificação que usualmente são cometidos.

4.4 Aplicação

Nesta seção apresentamos uma aplicação do modelo de regressão beta com dispersão variável aos dados da obesidade adulta nos Estados Unidos no ano de 2014. Estes dados são referentes ao estudo realizado por Souza et al. (2016). Aqui, nosso interesse consiste em comparar os resultados inferenciais obtidos através do uso de diferentes funções de ligação. Para esses dados temos que a variável resposta, *OB2014*, é a proporção de adultos obesos nos estados e totalizam 50 observações. As variáveis independentes utilizadas para explicar a obesidade nos estados foram: a porcentagem de residentes desempregados ou empregados em tempo parcial em 2014 (*DESEMP*), a porcentagem de adultos que consumiam vegetais menos de uma vez por dia em 2011 (*VEGET*), a porcentagem de residentes que não tinham cobertura de seguro de saúde em 2014 (*DESCOB*), o escore de bem-estar em 2014 (*BST*), a porcentagem de fumantes de cigarro em 2012 (*FUM*) e a taxa de insegurança alimentar em 2013 (*INSEG*). Neste estudo, a especificação do modelo de regressão beta com dispersão variável pode ser definida da seguinte maneira:

$$\begin{aligned} g(\mu_t) &= \beta_0 + \beta_1 DESEMP_t + \beta_2 VEGET_t + \beta_3 DESCOB_t + \beta_4 BST_t + \beta_5 FUM_t, \\ h(\phi_t) &= \gamma_0 + \gamma_1 INSEG_t + \gamma_2 DESCOB_t + \gamma_3 VEGET_t, \end{aligned} \tag{4.9}$$

com $t = 1, \ldots, 50$. Além disso, temos que $g(\cdot)$ e $h(\cdot)$ são as funções de ligação utilizadas para modelar a média e precisão, respectivamente. Vale salientar que foram feitas tentati-

vas com as funções de ligação loglog, cloglog e logit para a estrutura da média e log e sqrt para a precisão, porém apresentamos apenas os ajustes com as funções de ligação cujos resultados inferenciais mostraram-se mais diferenciados, que aqui denominamos de ajuste 1 e ajuste 2 (ver Tabela 4.5). Aplicamos o teste da razão de verossimilhanças (NEYMAN; PEARSON, 1928; SOUZA et al., 2016) para testar a hipótese nula de que a dispersão dos dados é fixa versus a hipótese alternativa de que a mesma é variável, concluindo ao nível de significância de 5% a necessidade de uma estrutura para modelar a dispersão dos dados.

Tabela 4.5: Estimativas dos parâmetros (Est.), erros-padrão (E.P.) e p-valores dos modelos considerando as funções de ligação loglog para a estrutura da média e log e sqrt para a estrutura da precisão.

Ajuste	**1**			**2**		
Especificação	**loglog(μ_t)** e **log(ϕ_t)**			**loglog(μ_t)** e **sqrt(ϕ_t)**		
			Modelo para μ			
Variáveis	**Est.**	**E.P.**	**p-valor**	**Est.**	**E.P.**	**p-valor**
$INTERCEPTO$	0.546	0.267	0.041	0.853	0.446	0.056
$DESEMP$	-0.005	0.001	< 0.001	-0.006	0.003	0.012
$VEGET$	0.010	0.002	< 0.001	0.009	0.002	< 0.001
$DESCOB$	0.004	0.001	0.004	0.005	0.002	< 0.001
BST	-0.018	0.004	< 0.001	-0.023	0.006	< 0.001
FUM	0.008	0.002	< 0.001	0.008	0.002	< 0.001
			Modelo para ϕ			
Variáveis	**Est.**	**E.P.**	**p-valor**	**Est.**	**E.P.**	**p-valor**
$INTERCEPTO$	3.110	1.557	0.046	-17.409	18.094	0.336
$INSEG$	-0.370	0.094	< 0.001	–	–	–
$DESCOB$	0.138	0.062	0.025	–	–	–
$VEGET$	0.332	0.059	< 0.001	2.146	0.835	0.010
λ	1349.416			11.509		
Pseudo-R^2	0.7731			0.7834		

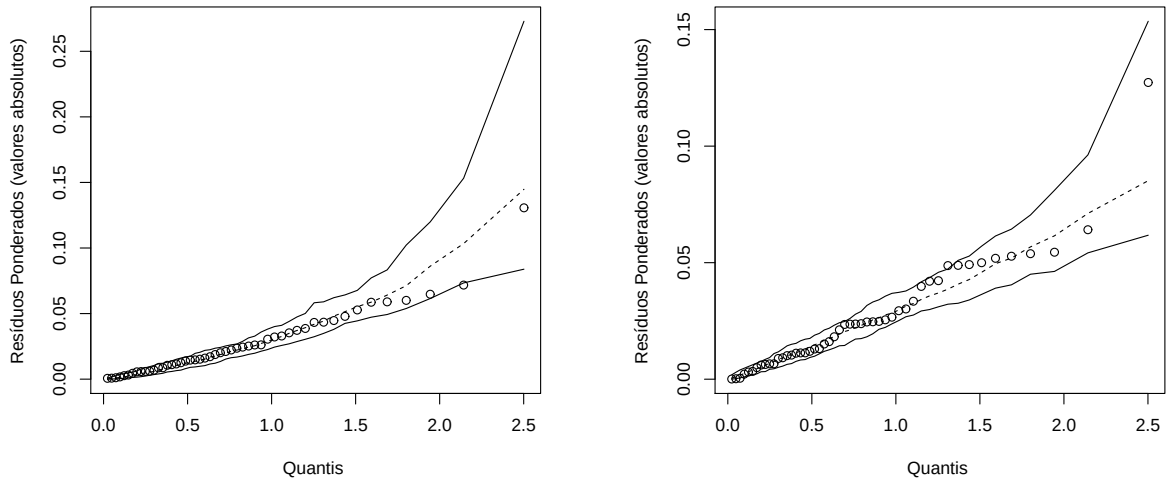
A Tabela 4.5 apresenta as estimativas, erros-padrão e p-valores obtidos após a modelagem dos dados considerando a função de ligação loglog para o submodelo da média, log para o submodelo da precisão no ajuste 1 e sqrt para o submodelo da precisão no ajuste 2. Nesta tabela são apresentados apenas os resultados inferenciais para aquelas variáveis cujo o parâmetro apresentou um p-valor do teste significativo. Desta forma, pôde-se verificar que para o ajuste 2, as variáveis *INSEG* e *DESCOB* não foram significativas no submodelo da precisão, sendo o mesmo reajustado sem essas variáveis. Para cada modelo foi avaliado a razão entre os valores de máximo e mínimo da precisão (λ), que pode ser interpretado como uma medida de não-constância da precisão dos dados. Verificamos que a mudança na estrutura de regressão do parâmetro de precisão ocasionou uma grande variabilidade nesta medida, sendo que para o ajuste 1 o valor encontrado foi de 1349.416, enquanto que para o ajuste 2 este mesmo valor foi de 11.509, resultando em uma mudança considerável no grau de precisão. Além disso, avaliamos o teste *RESET* (LIMA, 2007; OLIVEIRA, 2013), o gráfico de probabilidade normal com envelopes

simulados, o pseudo-R^2 e as medidas de influência para que fosse possível comparar as mudanças inferenciais ocorridas quando se utilizam diferentes formas de especificação na modelagem dos dados. Por fim, aplicamos o teste J (CRIBARI-NETO; LUCENA, 2015) para chegarmos ao modelo melhor especificado.

Ferrari e Cribari–Neto (2004) propuseram uma medida similar ao coeficiente de determinação para modelos de regressão beta, denominado pseudo-R^2. Esta medida permite avaliar a qualidade do ajuste dos modelos e apresenta valores contidos no intervalo $(0, 1)$, sendo que, quanto mais próximo de 1 melhor a capacidade explicativa do modelo proposto. Desta forma, para os ajustes 1 e 2 obtivemos respectivamente os valores de 0.7721 e 0.7834. Portanto, a mudança na estrutura de regressão do parâmetro de precisão ocasionou uma variação da capacidade explicativa dos modelos, sendo o ajuste 2 o que apresentou a maior capacidade explicativa.

Nós utilizamos o teste *RESET* adaptado para modelos de regressão beta (LIMA, 2007; OLIVEIRA, 2013) com o objetivo de testar a correta especificação dos modelos propostos. A hipótese nula sugere que o modelo testado está bem especificado versus a hipótese alternativa de que o mesmo está mal especificado. Para a realização do teste, consideramos o preditor linear estimado elevado a segunda potência ($\widehat{\eta}^2$) como variável de teste incluída no submodelo da média. Desta forma, obtivemos os p-valores de 0.1141 e 0.1522 referentes aos ajustes 1 e 2, respectivamente. Portanto, podemos concluir que tais modelos não apresentam erros na omissão de variáveis ou forma funcional incorreta aos níveis usuais de significância. Vale destacar aqui, que quando consideradas as funções de ligação cloglog e log para modelar as estruturas da média e da precisão, respectivamente, o teste *RESET* sugere a incorreta especificação do modelo ao nível de significância de 5% (p-valor = 0.0444), podendo o mesmo apresentar algum tipo de erro de especificação.

A Figura 4.2 apresenta os gráficos de probabilidade normal com envelopes simulados



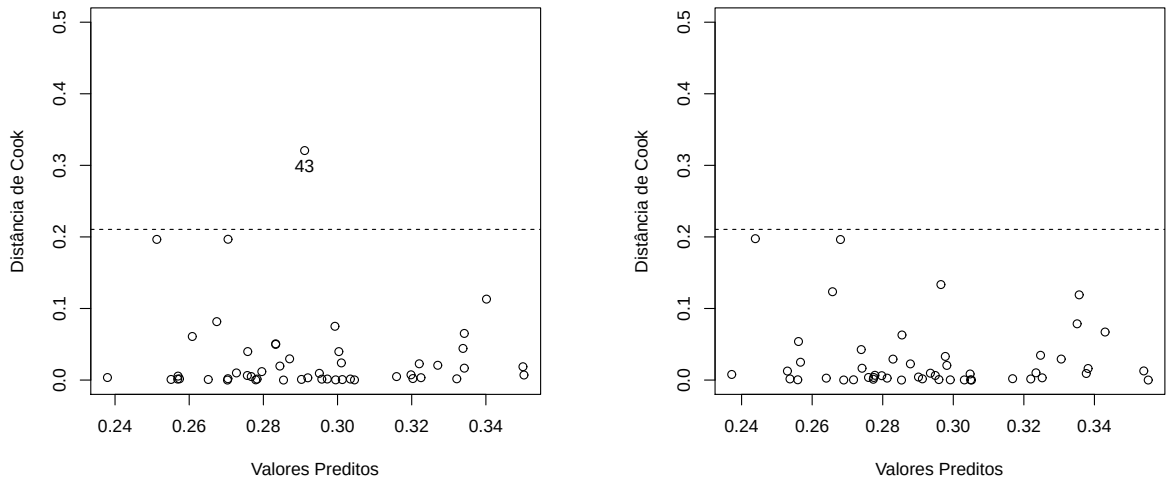
(a) loglog e log

(b) loglog e sqrt

Figura 4.2: Gráficos de probabilidade normal com envelopes simulados.

utilizando os resíduos ponderados (ESPINHEIRA et al., 2008a). A partir destes gráficos é possível concluir que os ajustes avaliados apresentam os resíduos, em sua maioria, dentro das bandas de confiança dos envelopes simulados, mostrando a adequabilidade dos modelos independente da especificação utilizada. Valendo aqui ressaltar que os resíduos do ajuste 1 apresentaram menores desvios comparado aos do ajuste 2.

A partir da análise de influência para modelos de regressão beta (ESPINHEIRA et al., 2008b) foi possível avaliar algumas medidas como distância de Cook e alavancagem generalizada. A distância de Cook foi introduzida por Cook (1977) como uma medida capaz de quantificar o impacto de cada observação nas estimativas dos parâmetros. A Figura 4.3 apresenta os gráficos da distância de Cook versus os valores preditos, sendo possível visualizar que houveram diferenças na classificação das observações de acordo com a especificação utilizada. Para o ajuste 1 apenas a observação 43, referente ao estado do Texas, encontrou-se destacada frente as demais, sendo este estado o que apresentou uma das maiores porcentagens de residentes que não tinham cobertura de seguro de saúde (*DESCOB*). Por outro lado, para o ajuste 2 não foram identificamos pontos de influência.



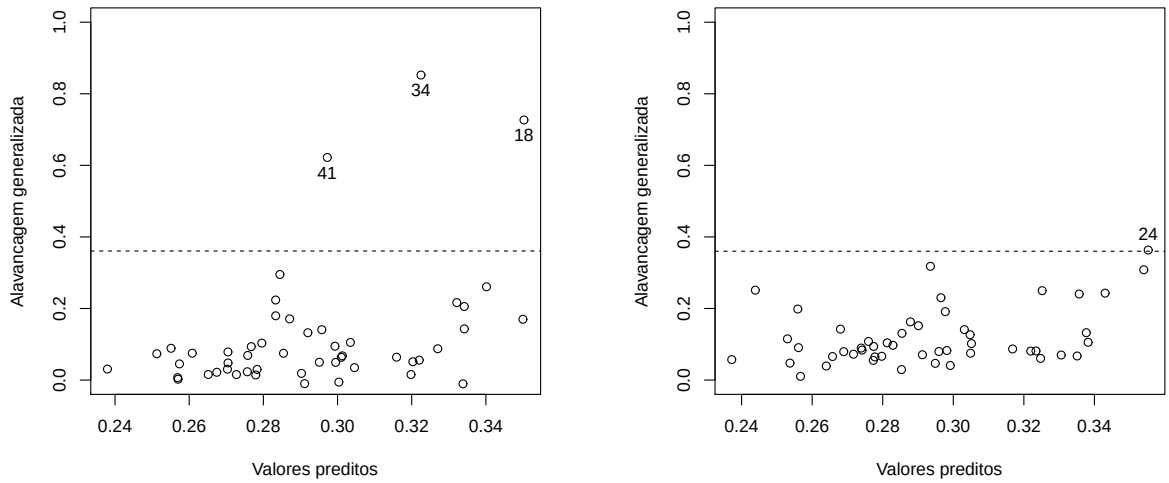
(a) loglog e log

(b) loglog e sqrt

Figura 4.3: Gráficos das distâncias de Cook.

A alavancagem generalizada foi proposta por Wei et al. (1998) como uma medida da importância individual de cada observação. A Figura 4.4 apresenta os gráficos da alavancagem generalizada versus os valores preditos, e a partir dela, é possível visualizar que a alteração na estrutura de regressão da precisão ocasionou mudanças nestes gráficos. Para o ajuste 1 temos que as observações 18, 34 e 41, referentes aos estados de Louisiana, North Dakota e South Dakota, respectivamente, foram consideradas pontos de alavanca, enquanto que para o ajuste 2 apenas a observação 24, referente ao estado do Mississipi, encontrou-se destacada frente as demais. Vale salientar que o estado de Louisiana apresentou a maior porcentagem de adultos que consumiam vegetais menos de uma vez ao dia (*VEGET*), enquanto que North Dakota apresentou a menor porcentagem de residentes desempregados ou empregados em tempo parcial (*DESEMP*) e a menor taxa de insegurança alimentar (*INSEG*). O estado de South Dakota apresentou um dos maiores escores

de bem-estar (*BST*), enquanto que o Mississipi apresentou a maior taxa de insegurança alimentar (*INSEG*) e uma das maiores porcentagens de adultos que consumiam vegetais menos de uma vez ao dia (*VEGET*).



(a) loglog e log (b) loglog e sqrt

Figura 4.4: Gráficos de alavancagem generalizada.

Por fim, a Figura 4.5 apresenta o gráfico dos valores observados versus os valores estimados, permitindo assim a comparação entre as respostas médias estimadas obtidas a partir de cada modelo proposto. Como resultado, verificamos que as estimativas médias de *OB2014* se diferenciaram dependendo da especificação utilizada no modelo, contudo não se foi possível indicar um melhor modelo considerando esta medida, visto que as estimativas foram bem próximas dos valores reais em ambos os ajustes. Valendo ressaltar que o melhor ajuste seria aquele que apresentasse os menores desvios, denotados por $\mu_t - \widehat{\mu}_t$, com $t = 1, \ldots, 50$, ou seja, apresentando observações mais próximas possível da reta.

Uma das maneiras de se decidir entre dois modelos de regressão beta não-encaixados qual que apresenta-se melhor especificado é utilizar o teste J adaptado para este tipo de modelo apresentado por Cribari–Neto e Lucena (2015). A aplicação deste teste é realizada de forma sequencial para cada modelo. Primeiramente, para testar a especificação do ajuste 1, incluímos no mesmo a estimativa do preditor linear do ajuste 2 como variável de teste, em seguida, testamos a significância do modelo por meio do teste da razão de verossimilhanças, obtendo assim um p-valor 0.0440. Em sequência, testamos a especificação do ajuste 2 incluindo a estimativa do preditor linear do ajuste 1 como variável de teste, obtendo um p-valor de 0.2350. Desta forma, concluímos aos níveis usuais de significância que o ajuste 2 apresentou uma melhor especificação, ou seja, a partir do teste J temos que o modelo melhor especificado é aquele com funções de ligação loglog e sqrt para modelar as estruturas de regressão da média e da precisão, respectivamente.

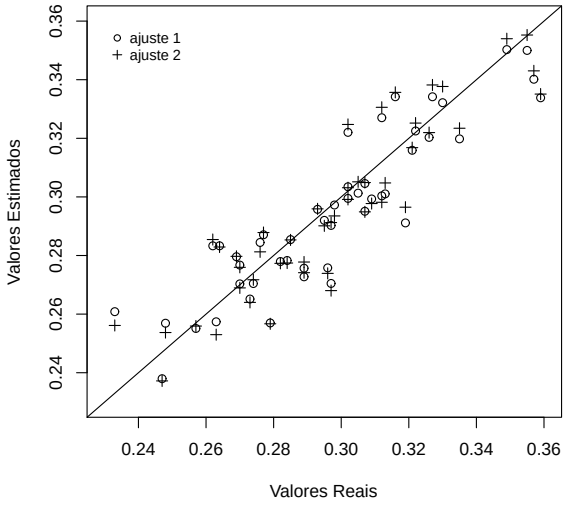


Figura 4.5: Gráfico dos valores observados versus os valores estimados da variável obesidade adulta nos Estados Unidos em 2014, considerando os ajustes com diferentes funções de ligação.

4.5 Conclusões

Neste capítulo avaliamos o efeito de erros de especificação nas inferências do modelo de regressão beta com dispersão variável. Para isto, um estudo de simulação foi realizado considerando diferentes cenários. Neste estudo, o modelo de regressão beta foi ajustado sob a especificação correta e incorreta. Em particular, seis tipos de erros de especificação foram avaliados, englobando tanto erros nos preditores quanto nas funções de ligação dos submodelos da média e da precisão, incluindo o caso em que a estrutura de regressão para o parâmetro de precisão é negligenciada erroneamente.

Verificamos através das taxas de rejeição que os modelos estimados com erro no preditor da precisão tenderam a apresentar valores mais distantes dos níveis nominais, comparados aos demais modelos. Além do mais, o modelo estimado com mais erros de especificação (erros nas funções de ligação e nos preditores) apresentou as maiores taxas de rejeição, considerando médias localizadas próximas a 0 e a 0.5. Em contrapartida, para valores de médias próximas a 1, errar nos preditores das duas estruturas de regressão (média e precisão) se mostrou um erro mais grave. Adicionalmente, negligenciar a estrutura de regressão do parâmetro de precisão, por exemplo, se mostrou mais grave do que errar no preditor linear da média e nas funções de ligação dos dois submodelos. Em relação a análise dos resultados referentes as taxas de cobertura do parâmetro β_2, houve a confirmação de algumas das conclusões citadas acima. Por exemplo, nos modelos estimados com erro no preditor da precisão os resultados tenderam a ser mais distantes do valor esperado (95%) e o modelo estimado com erros nas funções de ligação e nos preditores dos dois submodelos de regressão, no geral, apresentou os piores resultados comparados aos demais. Em relação as medidas utilizadas para avaliar as estimativas para as respostas médias, o modelo estimado com precisão fixa apresentou, em alguns casos, os resultados mais distantes do valor esperado. Por fim, uma aplicação a dados reais foi realizada com o

objetivo de verificar na prática os efeitos de diferentes formas de especificação no modelo de regressão beta com dispersão variável.

Observamos que, os erros de especificação que envolviam o preditor linear da estrutura de regressão do parâmetro de precisão tiveram uma influência considerável nas inferências do modelo de regressão beta. Este resultado confirma o que é encontrado na literatura. Uma das possíveis soluções para contornar este problema, caso seja de interesse no estudo a identificação de fontes de variabilidade, é o uso do teste *RESET* adaptado para a classe de modelos de regressão beta. Este teste é adequado para identificar possíveis erros de especificação que usualmente são cometidos. Uma outra solução, caso não se tenha tanto interesse em modelar a variabilidade, é o uso dos estimadores do tipo sanduíche propostos por Cribari–Neto e Souza (2012). Estes estimadores são adequados para os casos em que a estrutura de regressão que modela a variabilidade é negligenciada. Segundo os autores, as inferências obtidas através do uso destes estimadores apresentam bons resultados mesmo sob dispersão variável.

Capítulo 5

Considerações Finais

5.1 Conclusões

Ao longo dos capítulos ilustramos a aplicabilidade prática e teórica do modelo de regressão beta com dispersão variável. As principais conclusões deste trabalho encontram-se resumidas a seguir:

1. No Capítulo 2, avaliamos e explicamos a proporção de crianças obesas, entre 0 e 5 anos de idade, beneficiadas pelo Programa Bolsa Família no ano de 2014 e identificamos para cada região do Brasil os fatores que influenciaram a obesidade destes indivíduos. Utilizamos o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) com a finalidade de explicar a obesidade infantil por região. Os resultados mostraram que para as Regiões Norte e Sudeste o gasto per capita com o Programa Bolsa Família apresentou influência positiva na obesidade, ou seja, quanto mais se gastou com o referido programa assistencial, maior foi a incidência de crianças obesas. Este resultado exige atenção, pois nestas duas regiões o rendimento extra proveniente do benefício pode ter influenciado nos hábitos alimentares das famílias beneficiárias e consequentemente na situação nutricional das crianças pertencentes às mesmas. Nos municípios das Regiões Sul e Centro-Oeste, a renda per capita influenciou negativamente na obesidade infantil, ou seja, nos municípios que apresentaram uma maior renda per capita, houve uma tendência a apresentar uma menor incidência de crianças obesas. Na Região Nordeste, nos municípios com uma maior taxa de desemprego e um maior percentual de pobres, houve uma tendência a apresentar uma maior incidência de obesidade em crianças, sendo que este resultado pode indicar que nesta região, famílias menos favorecidas economicamente tenderam a se alimentar de maneira inadequada, influenciando diretamente na situação nutricional das crianças pertencentes às mesmas.

2. No Capítulo 3, modelamos a proporção de adultos obesos nos Estados Unidos para o ano de 2014. Para isto, utilizamos o modelo de regressão beta com dispersão variável, uma vez que os dados apresentaram assimetria e estavam restritos ao intervalo $(0, 1)$. Os resultados mostraram que a falta de atividade física, o pouco consumo de vegetais por dia, o hábito de fumar e as taxas de insegurança alimentar nos estados, apresentam um efeito positivo no aumento da proporção média de adultos obesos. Por outro lado, as taxas de desemprego e o escore de bem-estar

exibiram uma relação negativa com o desfecho. Estimamos o impacto da inatividade física sobre a proporção média de adultos obesos e os resultados revelaram que o efeito desse impacto tem forma positiva para todos os possíveis valores de inatividade física.

3. No Capítulo 4, avaliamos os efeitos de alguns erros de especificação nas inferências do modelo de regressão beta com dispersão variável. Em particular, seis tipos de erros de especificação foram considerados. Para a avaliação de seus efeitos, um estudo de simulação foi realizado considerando diferentes cenários. Verificamos, através dos resultados obtidos por estas simulações, que os erros de especificação que envolviam o preditor linear da estrutura de regressão do parâmetro de precisão apresentaram uma influência considerável nas inferências do modelo, demonstrando a real importância de se modelar corretamente esta estrutura. Para contornar este problema de erros de especificação, algumas abordagens são recomendadas, como o uso do teste *RESET* ou de estimadores do tipo sanduíche (CRIBARI-NETO; SOUZA, 2012). Além destas análises, realizamos ainda uma aplicação a dados reais com o objetivo de verificar os efeitos de diferentes formas de especificação no modelo de regressão beta com dispersão variável.

5.2 Trabalhos futuros

Algumas linhas de pesquisa ainda podem ser desenvolvidas. Por exemplo, podem ser foco de futuras pesquisas:

1. Inclusão de outras variáveis sociais, econômicas e demográficas relacionadas ao tema Obesidade e Bolsa Família;

2. Comparação do desempenho dos estimadores do tipo sanduíche e dos estimadores usuais da matriz de covariâncias sob erros de especificação no modelo de regressão beta com dispersão variável, considerando a parâmetrização da densidade beta em termos dos parâmetros de média e de precisão.

5.3 Publicações

O presente trabalho é uma compilação dos seguintes artigos publicados/submetidos:

1. Capítulo 2:

   OLIVEIRA, A.A.; SOUZA, T.C. Avaliação da proporção de crianças obesas beneficiadas pelo Programa Bolsa Família nas regiões do Brasil. *Revista Ciências Exatas e Naturais*, 18(1), p.55–80, 2016.

2. Capítulo 3:

   SOUZA, S.A.; OLIVEIRA, A.A.; SOUZA, T.C.; LIMA, C.M.B.L. Modelagem da proporção de obesos nos Estados Unidos utilizando modelo de regressão beta com dispersão variável. *Ciência e Natura*, 38(3), p.1146–1156, 2016.

3. Capítulo 4:

   O artigo referente ao Capítulo 4, intitulado de "Erros de especificação no modelo de regressão beta com dispersão variável" foi submetido para a Revista Brasileira de Biometria, que é uma publicação do Departamento de Estatística da Universidade Federal de Lavras-UFLA, e encontra-se sob avaliação.

Referências Bibliográficas

[1] ABRANTES, M.M.; LAMOUNIER, J.A.; COLOSIMO, E.A. Prevalência de sobrepeso e obesidade em crianças e adolescentes das regiões Sudeste e Nordeste. *Jornal de Pediatria*, 78(4), p.335–340, 2002.

[2] AKAIKE, H. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), p.716–723, 1964.

[3] ALMEIDA JUNIOR, P.M.; SOUZA, T.C. Estimativas de votos da presidente Dilma Roussef nas eleições presidenciais de 2010 sob o âmbito do Bolsa Família. *Ciência e Natura*, 37(1), p.12–22, 2015.

[4] ANDRADE, A.C.G. *Efeitos da especificação incorreta da função de ligação no modelo de regressão beta*. Dissertação (Mestrado em Ciências) - Universidade de São Paulo, São Paulo, 2007.

[5] ARTERBURN, D.; MACIEJEWSKI, M.; TSEVAT, J. Impact of morbid obesity on medical expenditures in adults. *International Journal of Obesity*, 29(3), p.334–339, 2005.

[6] BAYER, F.M.; CRIBARI-NETO, F. Model selection criteria in beta regression with varying dispersion. *Communications in Statistics – Simulation and Computation*, 46(1), p.729–746, 2017.

[7] BASU, S.; MCKEE, M.; GALEA, G.; STUCKLER, D. Relationship of soft drink consumption to global overweight, obesity and diabetes: a cross-national analysis of 75 countries. *American Journal of Public Health*, 103(11), p.2017–2077, 2013.

[8] BRASIL. Ministério da Saúde. *Obesidade e Desnutrição*. Departamento de Nutrição da Faculdade de Ciências da Saúde da Universidade de Brasília (FS/UnB). Área Técnica de Alimentação e Nutrição do Departamento de Atenção Básica da Secretaria de Política de Saúde do Ministério da Saúde (DAB/SPS/MS). Disponível em: $< http://bvsms.saude.gov.br/bvs/publicacoes/obesidade\_desnutricao.pdf >$. Acesso em: março de 2015.

[9] BRASIL. Ministério do Desenvolvimento Social (MDS) - Bolsa Família. Disponível em: $< http://www.mds.gov.br/bolsafamilia >$. Acesso em: março de 2015.

[10] BRASIL. Secretaria de Direitos Humanos da Presidência da República. *SDH/PR apresenta dados sobre alimentação de crianças e adolescentes no Brasil*. Disponível em: $< http://www.sdh.gov.br/noticias/pdf/alimentacao–adequada–estudo–completo >$. Acesso em: março de 2015.

[11] CABRAL, M.J.; VIEIRA, K.A.; SAWAYA, A.L.; FLORÊNCIO, T.M.M.T. Perfil socioeconômico, nutricional e de ingestão alimentar de beneficiários do Programa Bolsa Família. *Estudos Avançados*, 27(78), p.71–87, 2013.

[12] CABRERA, M.; FILHO, W. Obesidade em idosos: prevalência, distribuição e associação com hábitos e co-morbidades. *Arquivos brasileiros de Endocrinologia e Metabologia*, 45(5), p.494–501, 2001.

[13] CANTERLE, D.R.; PALM, B.G.; BAYER, F.M. Efeitos da especificação incorreta das funções de ligação no modelo de regressão beta com dispersão variável. *Revista Brasileira de Biometria*, 33(3), p.378–394, 2015.

[14] CASTANHO, G.K.F.; MARSOLA, F.C.; MCLELLAN, K.C.P.; NICOLA, M.; MORETO, F.; BURINI, R.C. Consumo de frutas, verduras e legumes associado à síndrome metabólica e seus componentes em amostra populacional adulta. *Ciência & Saúde Coletiva*, 18(2), p.385–392, 2013.

[15] CAVALCANTI, C.B.S.; BARROS, M.V.G.; MENÊSES, A.L.; SANTOS, C.M.; AZEVEDO, A.M.P. GUIMARÃES, F.J.D.S.P. Obesidade abdominal em adolescentes: prevalência e associação com atividade física e hábitos alimentares. *Arquivos Brasileiros de Cardiologia*, 94(3), p.371–377, 2010.

[16] COOK, R.D. Detection of influential observations in linear regression. *Technometrics*, 19(1), p.15–18, 1977.

[17] COTTA, R.M.M.; MACHADO, J.C. Programa Bolsa Família e segurança alimentar e nutricional no Brasil: revisão crítica da literatura. *Revista Panamericana de Salud Publica*, 33(1), p.54–60, 2013.

[18] CRIBARI-NETO, F.; LUCENA, S.E.F. Nonnested hypothesis testing in the class of varying dispersion beta regression. *Journal of Applied Statistics*, 42, p.967–985, 2015.

[19] CRIBARI-NETO, F.; PEREIRA, T.L. Avaliação da eficiência de administrações municipais no estado de São Paulo: uma nova abordagem via modelos de regressão beta. *Revista Brasileira de Biometria*, 31(2), p.270–294, 2013.

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

[21] CRIBARI-NETO, F.; SOUZA, T.C. Religious belief and intelligence: Worldwide evidence. *Intelligence*, 41(5), p.482–489, 2013.

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

[23] DANAEI, G.; DING, E.; MOZAFFARIAN, D.; TAYLOR, B.; REHM, J.; MURRAY, C.; EZZATI, M. The preventable causes of death in the United States: comparative risk assesment of dietary, lifestyle and metabolic risk factors. *PLOS Medice*, 6(4), p.1–23, 2009.

[24] DHAROD, J.M.; CROOM, J.E.; SADY, C.G. Food insecurity: its relationship to dietary intake and body wheight among somali refugee women in the United States. *Journal of Nutrition Education and Behavior*, 45(1), p.47–53, 2013.

[25] ESPINHEIRA, P.L. *Regressão Beta*. Tese (Doutorado em Ciências) - Universidade de São Paulo, São Paulo, 2007.

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

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

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

[29] FERRARI, S.L.P.; ESPINHEIRA, P.L.; CRIBARI-NETO, F. Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, 65(3), p.337–351, 2011.

[30] FLEGAL, K.M. The effects of changes in smoking prevalence on obesity prevalence in the United States. *American Journal of Public Health*, 97(8), p.1510–1514, 2007.

[31] GABRIELE, R. *Índice de massa corporal no diagnóstico de transtornos nutricionais em idosos institucionalizados no município de Fortaleza, Ceará*. Dissertação (Mestrado em Saúde Coletiva) - Universidade de Fortaleza, Fortaleza, 2011.

[32] HOLBEN, D. Position of the american dietetic association: Food insecurity in the United States. *Journal of the American Dietetic Association*, 110(9), p.1368–1377, 2010.

[33] IBASE. Instituto Brasileiro de Análises Sociais e Econômicas. *Repercussões do Programa Bolsa Família na segurança alimentar e nutricional das famílias beneficiadas*. Rio de Janeiro, 2008.

[34] JAGIELSKI, A.; BROWN, A.; HOSSEINI-ARAGHI, M.; THOMAS, N.; TAHERI, S. The association between adiposity, mental well-being and quality of life in extreme obesity. *PLOS One*, 9(3), p.1–8, 2014.

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

[36] KLEIBER, C.; ZELEIS, A. *Applied Econometrics with R*. New York: Springer, 2008.

[37] LIMA, F.E.L.; RABITO, E.I.; DIAS, M.R.M.G. Estado nutricional de população adulta beneficiária do Programa Bolsa Família no município de Curitiba, PR. *Revista Brasileira de Epidemiologia*, 14(2), p.198–206, 2011.

[38] LIMA, L.B. *Um teste de especificação correta para modelos de regressão beta*. Dissertação (Mestrado em Estatística) - Universidade Federal de Pernambuco, Recife, 2007.

[39] LOOSE, L.H.; PALM, B.G.; BAYER, F.M. Avaliação dos estimadores do modelo de regressão beta com dispersão variável: um estudo de simulação. *Revista Eletrônica Matemática e Estatística em Foco*, 2(1), p.14–24, 2014.

[40] MCCULLAGH, P.; NEIDER, J.A. *Generalized Linear Models, 2nd ed.* London: Chapman and Hall, 1989.

[41] MONTEIRO, F.; SCHMIDT, S.T.; COSTA, I.B.; ALMEIDA, C.C.B.; MATUDA, N.S. Bolsa Família: insegurança alimentar e nutricional de crianças menores de cinco anos. *Ciência & Saúde Coletiva*, 19(5), p.1347–1357, 2014.

[42] MOREIRA, M.A.; CABRAL, P.C.; FERREIRA, H.S.; LIRA, P.I.C. Overweight and associated factors in children from northeasten Brazil. *Jornal de Pediatria*, 88(4), p.347–352, 2012.

[43] NEYMAN, J.; PEARSON, E.S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20, p.175–240, 1928.

[44] OGDEN, C.; CARROLL, M.; KIT, B.; FLEGAL, K. Prevalence of obesity among adults: United States, 2011–2012. *Medical Benefits*, 31(1), p.1–8, 2014.

[45] OLIVEIRA, A.A.; SOUZA, T.C. Avaliação da proporção de crianças obesas beneficiadas pelo Programa Bolsa Família nas regiões do Brasil. *Revista Ciências Exatas e Naturais*, 18(1), p.55–80, 2016.

[46] OLIVEIRA, F.C.C.; COTTA, R.M.M.; SANT'ANA, L.F.; PRIORI, S.E.; FRANCESCHINI, S.C.C. Programa Bolsa Família e estado nutricional infantil: desafios estratégicos. *Ciência & Saúde Coletiva*, 16(7), p.3307–3316, 2011.

[47] OLIVEIRA, J.S.C. *Detectando má especificação em regressão beta*. Dissertação (Mestrado em Estatística) - Universidade Federal de Pernambuco, Recife, 2013.

[48] OSPINA, R.; CRIBARI-NETO, F.; VASCONCELOS, K.L.P. Improved point and interval estimation for a beta regression model. *Computational Statistics and Data Analysis*, 51, p.960–981, 2006.

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

[50] PEREIRA, T.L.; CRIBARI-NETO, F. Detecting model misspecification in inflated beta regressions. *Communications in Statistics: Simulation and Computation*, 43, p.631–656, 2014.

[51] PIETILÄINEN, K.; KAPRIO, J.; BORG, P.; PLASQUI, G.; YKIJÄRVINEN, H.; KUJALA, U.; ROSE, R.; WESTERTERP, K.; RISSANEN, A. Physical inactivity and obesity: A vicious circle. *Obesity (Silver Spring)*, 16(2), p.409–414, 2008.

[52] PINTO, E.R.; PEREIRA, L.A.; RESENDE, L.O.; DESTRO FILHO, J.B. Modelos Estatísticos para estimação da área miocárdica sob risco de necrose. *Revista Brasileira de Biometria*, 29(3), p.395–415, 2011.

[53] PRESS, W.; TEUKOLSKY, S.; VETTERLING, W.; FLANNERY, B. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, 1992.

[54] R DEVELOPMENT CORE TEAM R: A Language and Environment for Statistical Computing. *R Fondation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, 2014.

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

[56] REZENDE, F.A.C.; ROSADO, L.E.F.P.; RIBEIRO, R.C.L.; VIDIGAL, F.C.; VASQUES, A.C.J.; BONARD, I.S.; CARVALHO, C.R. Índice de massa corporal e circunferência abdominal: associação com fatores de risco cardiovascular. *Arquivos Brasileiros de Cardiologia*, 87(6), p.728–734, 2006.

[57] SALDIVA, S.R.D.M.; SILVA, L.F.F.; SALDIVA, P.H.N. Avaliação antropométrica e consumo alimentar em crianças menores de cinco anos residentes em um município da região do semiárido nordestino com cobertura parcial do programa bolsa família. *Revista de Nutrição*, 23(2), p.221–229, 2010.

[58] SANT'ANNA, A.M.O.; CATEN, C.S. Modelagem da fração de não-conformes em processos industriais. *Pesquisa Operacional*, 30(1), p.53–72, 2010.

[59] SEGALL-CORRÊA, A.M.; MARIN-LEON, L.; HELITO, H.; PÉREZ-ESCAMILLA, R.; SANTOS, L.M.P.; PAES-SOUZA, R. Transferência de renda e segurança alimentar no Brasil: análise dos dados nacionais. *Revista de Nutrição*, 21(Suplemento), p.39s–51s, 2008.

[60] SCHWARZ, G. Estimating the dimension of the model. *Annals of Statistics*, 6(2), p.461–464, 1978.

[61] SILVA, C.R.; SOUZA, T.C. Modelagem da taxa de analfabetismo no estado da Paraíba via modelo de regressão beta. *Revista Brasileira de Biometria*, 32(3), p.345–359, 2014.

[62] SILVA, D.A.S. Sobrepeso e obesidade em crianças de cinco a dez anos de idade beneficiárias do programa bolsa família no estado de Sergipe, Brasil. *Revista Paulista de Pediatria*, 29(4), p.529–535, 2011.

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

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

[65] SOUZA, S.A.; OLIVEIRA, A.A.; SOUZA, T.C.; LIMA, C.M.B.L. Modelagem da proporção de obesos nos Estados Unidos utilizando modelo de regressão beta com dispersão variável. *Ciência e Natura*, 38(3), p.1146–1156, 2016.

[66] SOUZA, T.C.; CRIBARI-NETO, F. Uma estimativa do impacto eleitoral do Programa Bolsa Família. *Revista Brasileira de Biometria*, 31(1), p.79–103, 2013.

[67] SOUZA, T.C.; CRIBARI-NETO, F. Intelligence, religiosity and homosexuality non-acceptance: Empirical evidence. *Intelligence*, 52, p.63–70, 2015.

[68] SOUZA, T.C.; PEREIRA, T.L.; CRIBARI-NETO, F.; LIMA, V.M.C. Testing inference in inflated beta regressions under model misspecification. *Communications in Statistics - Simulation and Computation*, 45, p.625–642, 2016.

[69] STOL, A.; GUGELMIN, G.; LAMPA-JUNIOR, V.M.; FRIGULHA, C.; SELBACH, R.A. Complicações e óbitos nas operações para tratar a obesidade mórbida. *Arquivos Brasileiros de Cirurgia Digestiva*, 24(4), p.282–284, 2011.

[70] WALD, A. Test of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3), p.426–482, 1943.

[71] WEI, B.; HU, Y.; FUNG, W. Generalized leverage and its aplications. *Scandinavian Journal of Statistics*, 25(1), p.25–37, 1998.

[72] WOLF, M.R.; FILHO, A.A.B. Estado nutricional dos beneficiários do Programa Bolsa Família no Brasil - uma revisão sistemática. *Ciência & Saúde Coletiva*, 19(5), p.1331–1338, 2014.

[73] WORLD HEALTH ORGANIZATION. *Childhood overweight and obesity on the rise.* Disponível em: $< http://www.who.int/dietphysicalactivity/childhood/en/ >$. Acesso em: março de 2015.

[74] WORLD HEALTH ORGANIZATION. *Obesity and overweight.* Disponível em: $< http://www.who.int/mediacentre/factsheets/fs311/en/ >$. Acesso em: novembro de 2015.

[75] WORLD HEALTH ORGANIZATION. *Obesity: Situation and trends.* Disponível em: $< http://www.who.int/gho/ncd/risk\_factors/obesity\_text/en/ >$. Acesso em: novembro de 2015.

[76] ZHANG, Q.; LAMICHHANE, R.; WANG, Y. Associations between U.S. adult obesity and state and county economic conditions in the recession. *Journal of Clinical Medicine*, 3(1), p.153–166, 2014.

Apêndice

Neste apêndice apresentamos os *scripts* do R utilizados nas aplicações e nas simulações de Monte Carlo (disponíveis pelo e-mail: andreoliveira53@hotmail.com).

```
#============================================================================
#REGRESSÃO BETA - APLICAÇÃO PRÁTICA - DADOS OBESIDADE NOS EUA EM 2014 --------
#============================================================================

#Limpar o R
rm(list=ls())

#Ler dados
dadosEUA = read.table("DadosEUA.txt", header=T)

#Nome das variáveis
names(dadosEUA)

#Matriz de dados
dataS = data.frame(dadosEUA)

#Liberar pacotes
library(betareg)
library(gamlss)
library(lmtest)

#Excluir linhas com NA
dados = na.omit(dadosEUA)
dados

#"Soltar" os dados no R
attach(dados)

#Tamanho amostral
n = length(OB2014)

#Renomeando as linhas
n1 = seq(1,50,1)
rownames(dados) = n1

#----------------------------------------------------------------------------
#ANÁLISE DESCRITIVA

#Nome das variáveis
names(dados)

#Transformação da variável resposta para proporção
OB2014P = OB2014/100
OB2014P
```

```
#Histograma da variável resposta
#postscript("hist.ps")
hist(OB2014P, xlab="Proporção de adultos obesos", ylab="Frequência",main="")
lines(density(OB2014P))
#dev.off()

#Boxplot da variável resposta
#postscript("boxplot.ps")
boxplot(OB2014P,horizontal=T)
#dev.off()

#Vetor de Variáveis
X = cbind(OB2014P, InFISICA2014, DIABETES2014, POP2014, GDP2014, MRENDA2014,
DESEMP2014, BEMESTAR2014, DESC2014, HIPER2013, FRUITS2011, VEGETABLES2011,
FARMERS2012, InALIMENTAR2013, FUMO2012, NUTRI2009)
X

#Correlação entre as variáveis explicativas
cor(X)

#Correlação entre as variáveis explicativas e a variável resposta
cor(X,OB2014P)

#Análise descritiva
summary(X)

#Identificando pontos de máximo e mínimo
#Máximo OB2014P
ifelse(OB2014P==0.3590,99999,0)
dados$ESTADO[4]
#ARKANSAS

#Mínimo OB2014P
ifelse(OB2014P==0.2130,99999,0)
dados$ESTADO[6]
#COLORADO

#Máximo InFISICA2014
ifelse(InFISICA2014==31.60,99999,0)
dados$ESTADO[24]
#MISSISSIPI

#Mínimo InFISICA2014
ifelse(InFISICA2014==16.40,99999,0)
dados$ESTADO[6]
#COLORADO

#Máximo VEGETABLES2011
ifelse(VEGETABLES2011==32.50,99999,0)
dados$ESTADO[18]
#LOUISIANA

#Mínimo VEGETABLES2011
ifelse(VEGETABLES2011==15.30,99999,0)
dados$ESTADO[37]
#OREGON
```

```
#Máximo FUMO2012
ifelse(FUMO2012==28.30,99999,0)
dados$ESTADO[17]
#KENTUCKY

#Mínimo FUMO2012
ifelse(FUMO2012==10.60,99999,0)
dados$ESTADO[44]
#UTAH

#Máximo DESEMP2014
ifelse(DESEMP2014==20.70,99999,0)
dados$ESTADO[28]
#NEVADA

#Mínimo DESEMP2014
ifelse(DESEMP2014==9.00,99999,0)
dados$ESTADO[34]
#NORTHDAKOTA

#Máximo InALIMENTAR2013
ifelse(InALIMENTAR2013==22.70,99999,0)
dados$ESTADO[24]
#MISSISSIPI

#Mínimo InALIMENTAR2013
ifelse(InALIMENTAR2013==7.80,99999,0)
dados$ESTADO[34]
#NORTHDAKOTA

#Máximo BEMESTAR2014
ifelse(BEMESTAR2014==64.70,99999,0)
dados$ESTADO[2]
#ALASKA

#Mínimo BEMESTAR2014
ifelse(BEMESTAR2014==59.00,99999,0)
dados$ESTADO[48]
#WESTVIRGINIA

#Máximo DESC2014
ifelse(DESC2014==24.40,99999,0)
dados$ESTADO[43]
#TEXAS

#Mínimo DESC2014
ifelse(DESC2014==4.60,99999,0)
dados$ESTADO[21]
#MASSACHUSETTS


#-------------------------------------------------------------------------
#AJUSTANDO O MODELO DE REGRESSÃO BETA

#STEPWISE - CRITÉRIO DE SELEÇÃO DE MODELOS BIC
mod = gamlss(OB2014P~InFISICA2014+NUTRI2009+VEGETABLES2011+FUMO2012+POP2014
+GDP2014+DESEMP2014+InALIMENTAR2013+BEMESTAR2014, family="BE", mu.link="loglog",
k = log(n))
```

```
mod5 <-stepGAIC(mod, what="mu", scope=~(InFISICA2014+NUTRI2009+VEGETABLES2011
+FUMO2012+POP2014+GDP2014+DESEMP2014+InALIMENTAR2013+BEMESTAR2014)^2,
mu.link="loglog",k = log(n))
mod5$anova

#Ajuste do modelo com precisão fixa
ajuste8 =  betareg(OB2014P ~ InFISICA2014+VEGETABLES2011+FUMO2012+DESEMP2014
+InALIMENTAR2013+BEMESTAR2014+FUMO2012:InALIMENTAR2013,link="cloglog")
summary(ajuste8)

#Ajuste do modelo com precisão variável (funções cloglog|log)
ajuste8.a =  betareg(OB2014P ~ InFISICA2014+VEGETABLES2011+FUMO2012+DESEMP2014
+InALIMENTAR2013+BEMESTAR2014+FUMO2012:InALIMENTAR2013|InALIMENTAR2013+DESC2014
+VEGETABLES2011, link="cloglog",link.phi="log")
summary(ajuste8.a)

#Teste da Razão de Verossimilhanças (Precisão Fixa vs. Precisão Variável)
lrtest(ajuste8,ajuste8.a)

#Teste RESET (preditor linear elevado a segunda potência)
lrtest(ajuste8.a, . ~ . + I(predict(ajuste8.a, type = "link")^2))

#-------------------------------------------------------------------------
#GRÁFICOS - Resíduos Ponderados Padronizados

#Resíduos vs. Número de Observações
#postscript("GraficoResNumObs.eps" ,width = 5.5, height = 5.5)
plot(ajuste8.a, which = 1:1, type="sweighted", sub.caption = "",
caption = "", main = "", ylim = c(-3,3), ann = 0)
title(xlab = "Índice da Observações", ylab = "Resíduos Ponderados Padronizados")
abline(h = 0, lty=2); abline(h = 2, lty=2); abline(h = -2, lty=2)
cutI=-2;cutS=2
res6=residuals(ajuste8.a,type="sweighted")
idI=which(res6<cutI);idS=which(res6>cutS)
text(idS,res6[idS],rownames(dados)[idS],pos=3)
text(idI,res6[idI],rownames(dados)[idI],pos=1)
#dev.off()

#Envelope
#postscript("GraficoEnvelope.eps" ,width = 5.5, height = 5.5)
plot(ajuste8.a, which = 5:5, type="sweighted", sub.caption = "", caption = "",
main = "", ann = 0)
title(xlab = "Quantis ", ylab = "Resíduos Ponderados Padronizados
(valores absolutos)")
#dev.off()

#Distância de Cook
#postscript("Cook.eps" ,width = 5.5, height = 5.5)
val_pred=predict(ajuste8.a)
alavan=gleverage(ajuste8.a)
CD = cooks.distance(ajuste8.a)
plot(val_pred,cooks.distance(ajuste8.a),xlab="Valores Preditos",
ylab="Distância de Cook",ylim=c(0,1.0))
abline(qf(0.5,8,length(OB2014P)-8),0,lty=2)
#Identificando observações destacadas no gráfico
lim=(qf(0.5,8,length(OB2014P)-8))
id=which(CD>0.2)
text(val_pred[id],CD[id],id,pos=1)
```

```
#dev.off()

#Alavancagem
#postscript("loglogelogAlav.eps" ,width = 5.5, height = 5.5)
alavan=gleverage(ajuste8.a)
val_pred=predict(ajuste8.a)
plot(val_pred,alavan,xlab="Valores preditos",ylab="Alavancagem Generalizada",
main="")
abline(3*mean(alavan),0,lty=2)
lim=3*mean(alavan)
id=which(alavan>lim)
text(val_pred[id],alavan[id],id,pos=1)
#dev.off()

#-------------------------------------------------------------------------
#NOVO AJUSTE SEM OS PONTOS DE ALAVANCA

x1 = c(18)
ajusteEUA1 = update(ajuste8.a, subset = -x1)
summary(ajusteEUA1)

x2 = c(34)
ajusteEUA2 = update(ajuste8.a, subset = -x2)
summary(ajusteEUA2)

x3 = c(41)
ajusteEUA3 = update(ajuste8.a, subset = -x3)
summary(ajusteEUA3)

x4 = c(18,34,41)
ajusteEUA4 = update(ajuste8.a, subset = -x4)
summary(ajusteEUA4)

#-------------------------------------------------------------------------
#LAMBDA (Grau de Precisão do modelo)
phihat = predict(ajuste8.a, type = "precision")
lambdacomp = max(phihat)/min(phihat)

#-------------------------------------------------------------------------
#RETIRANDO OS PONTOS DE ALAVANCA - MUDANÇA PERCENTUAL (%)

#RETIRANDO A OBSERVAÇÃO 18
#MudancaSEMobs18 =
#((coeficienteAjusteSEMobservacao18/coeficienteAjusteTODASobservacoes)-1)*100

#MÉDIA
MudancaSEMobs18B0 =  ((-1.592987/-1.473664) - 1)*100
MudancaSEMobs18B1 =  ((0.013895/0.012672) - 1)*100
MudancaSEMobs18B2 =  ((0.014604/0.011673) - 1)*100
MudancaSEMobs18B3 =  ((0.048719/0.055583) - 1)*100
MudancaSEMobs18B4 =  ((-0.012758/-0.014370) - 1)*100
MudancaSEMobs18B5 =  ((0.059994/0.069723) - 1)*100
MudancaSEMobs18B6 =  ((-0.015615/-0.018014) - 1)*100
MudancaSEMobs18B7 =  ((-0.002752/-0.003221) - 1)*100

#PRECISÃO
MudancaSEMobs18G0 =  ((2.29971/3.67941) - 1)*100
MudancaSEMobs18G1 =  ((-0.38942/-0.50974) - 1)*100
```

```
MudancaSEMobs18G2 =  ((0.15537/0.19656) - 1)*100
MudancaSEMobs18G3 =  ((0.37538/0.37035) - 1)*100

#LAMBDA
phihatSEMobs18 = predict(ajusteEUA1, type = "precision")
lambdaSEMobs18 = max(phihatSEMobs18)/min(phihatSEMobs18)
MudancaSEMobs18Lbd = ((lambdaSEMobs18/lambdacomp)-1)*100
MudancaSEMobs18Lbd

#RETIRANDO A OBSERVAÇÃO 34
#MudancaSEMobs34 =
#((coeficienteAjusteSEMobservacao34/coeficienteAjusteTODASobservacoes)-1)*100

#MÉDIA
MudancaSEMobs34B0 =  ((-1.720321/-1.473664) - 1)*100
MudancaSEMobs34B1 =  ((0.013905/0.012672) - 1)*100
MudancaSEMobs34B2 =  ((0.009691/0.011673) - 1)*100
MudancaSEMobs34B3 =  ((0.036140/0.055583) - 1)*100
MudancaSEMobs34B4 =  ((-0.012447/-0.014370) - 1)*100
MudancaSEMobs34B5 =  ((0.047769/0.069723) - 1)*100
MudancaSEMobs34B6 =  ((-0.008796/-0.018014) - 1)*100
MudancaSEMobs34B7 =  ((-0.001944/-0.003221) - 1)*100

#PRECISÃO
MudancaSEMobs34G0 =  ((0.86989/3.67941) - 1)*100
MudancaSEMobs34G1 =  ((-0.24186/-0.50974) - 1)*100
MudancaSEMobs34G2 =  ((0.17665/0.19656) - 1)*100
MudancaSEMobs34G3 =  ((0.33078/0.37035) - 1)*100

#LAMBDA
phihatSEMobs34 = predict(ajusteEUA2, type = "precision")
lambdaSEMobs34 = max(phihatSEMobs34)/min(phihatSEMobs34)
MudancaSEMobs34Lbd = ((lambdaSEMobs34/lambdacomp)-1)*100
MudancaSEMobs34Lbd

#RETIRANDO A OBSERVAÇÃO 41
#MudancaSEMobs41 =
#((coeficienteAjusteSEMobservacao41/coeficienteAjusteTODASobservacoes)-1)*100

#MÉDIA
MudancaSEMobs41B0 =  ((-1.301983/-1.473664) - 1)*100
MudancaSEMobs41B1 =  ((0.012875/0.012672) - 1)*100
MudancaSEMobs41B2 =  ((0.011628/0.011673) - 1)*100
MudancaSEMobs41B3 =  ((0.054196/0.055583) - 1)*100
MudancaSEMobs41B4 =  ((-0.015190/-0.014370) - 1)*100
MudancaSEMobs41B5 =  ((0.069034/0.069723) - 1)*100
MudancaSEMobs41B6 =  ((-0.020304/-0.018014) - 1)*100
MudancaSEMobs41B7 =  ((-0.003166/-0.003221) - 1)*100

#PRECISÃO
MudancaSEMobs41G0 =  ((3.64765/3.67941) - 1)*100
MudancaSEMobs41G1 =  ((-0.45095/-0.50974) - 1)*100
MudancaSEMobs41G2 =  ((0.17521/0.19656) - 1)*100
MudancaSEMobs41G3 =  ((0.34462/0.37035) - 1)*100

#LAMBDA
phihatSEMobs41 = predict(ajusteEUA3, type = "precision")
lambdaSEMobs41 = max(phihatSEMobs41)/min(phihatSEMobs41)
```

```
MudancaSEMobs41Lbd = ((lambdaSEMobs41/lambdacomp)-1)*100
MudancaSEMobs41Lbd

#RETIRANDO AS OBSERVAÇÕES 18,34,41
#MudancaSEM3obs =
#((coeficienteAjusteSEM3observacoes/coeficienteAjusteTODASobservacoes)-1)*100

#MÉDIA
MudancaSEM3obsB0 =  ((-1.627612/-1.473664) - 1)*100
MudancaSEM3obsB1 =  ((0.014651/0.012672) - 1)*100
MudancaSEM3obsB2 =  ((0.010102/0.011673) - 1)*100
MudancaSEM3obsB3 =  ((0.042204/0.055583) - 1)*100
MudancaSEM3obsB4 =  ((-0.013144/-0.014370) - 1)*100
MudancaSEM3obsB5 =  ((0.057551/0.069723) - 1)*100
MudancaSEM3obsB6 =  ((-0.012714/-0.018014) - 1)*100
MudancaSEM3obsB7 =  ((-0.002395/-0.003221) - 1)*100

#PRECISÃO
MudancaSEM3obsG0 =  ((1.40997/3.67941) - 1)*100
MudancaSEM3obsG1 =  ((-0.12858/-0.50974) - 1)*100
MudancaSEM3obsG2 =  ((0.12769/0.19656) - 1)*100
MudancaSEM3obsG3 =  ((0.25833/0.37035) - 1)*100

#LAMBDA
phihatSEM3obs = predict(ajusteEUA4, type = "precision")
lambdaSEM3obs = max(phihatSEM3obs)/min(phihatSEM3obs)
MudancaSEM3obsLbd = ((lambdaSEM3obs/lambdacomp)-1)*100
MudancaSEM3obsLbd

#-------------------------------------------------------------------------------
#CURVAS DE IMPACTO - FUNÇÃO DE LIGAÇÃO CLOGLOG

#Renomeando a iteração do ajuste
INT = (FUMO2012*InALIMENTAR2013)

#Ajuste
ajuste8.a=betareg(OB2014P~InFISICA2014+VEGETABLES2011+FUMO2012+DESEMP2014
+InALIMENTAR2013+BEMESTAR2014+INT|InALIMENTAR2013+DESC2014+VEGETABLES2011,
link="cloglog",link.phi="log")
summary(ajuste8.a)

#Fixar semente
set.seed(53)

#Gerando sequência de 0 a 100 - Variável InFISICA2014
xx1= seq(0, 100, length=100)

#Fixando as variáveis no primeiro quartil
impacto_25 = exp(-exp((ajuste8.a$coef$mean[1]+ajuste8.a$coef$mean[2]*xx1
+ajuste8.a$coef$mean[3]*quantile(VEGETABLES2011, 0.25)+ajuste8.a$coef$mean[4]
*quantile(FUMO2012, 0.25)+ajuste8.a$coef$mean[5]*quantile(DESEMP2014, 0.25)
+ajuste8.a$coef$mean[6]*quantile(InALIMENTAR2013, 0.25)+ajuste8.a$coef$mean[7]
*quantile(BEMESTAR2014, 0.25)+ajuste8.a$coef$mean[8]*quantile(INT, 0.25))))
*(exp((ajuste8.a$coef$mean[1]+ajuste8.a$coef$mean[2]*xx1+ajuste8.a$coef$mean[3]
*quantile(VEGETABLES2011, 0.25)+ajuste8.a$coef$mean[4]*quantile(FUMO2012, 0.25)
+ajuste8.a$coef$mean[5]*quantile(DESEMP2014, 0.25)+ajuste8.a$coef$mean[6]
*quantile(InALIMENTAR2013, 0.25)+ajuste8.a$coef$mean[7]*quantile(BEMESTAR2014, 0.25)
+ajuste8.a$coef$mean[8]*quantile(INT, 0.25)))*((ajuste8.a$coef$mean[2])))
```

```
#Fixando as variáveis na mediana
impacto_50 = exp(-exp((ajuste8.a$coef$mean[1]+ajuste8.a$coef$mean[2]*xx1
+ajuste8.a$coef$mean[3]*quantile(VEGETABLES2011, 0.5)+ajuste8.a$coef$mean[4]
*quantile(FUMO2012, 0.5)+ajuste8.a$coef$mean[5]*quantile(DESEMP2014, 0.5)
+ajuste8.a$coef$mean[6]*quantile(InALIMENTAR2013, 0.5)+ajuste8.a$coef$mean[7]
*quantile(BEMESTAR2014, 0.5)+ajuste8.a$coef$mean[8]*quantile(INT, 0.5))))
*(exp((ajuste8.a$coef$mean[1]+ajuste8.a$coef$mean[2]*xx1+ajuste8.a$coef$mean[3]
*quantile(VEGETABLES2011, 0.5)+ajuste8.a$coef$mean[4]*quantile(FUMO2012, 0.5)
+ajuste8.a$coef$mean[5]*quantile(DESEMP2014, 0.5)+ajuste8.a$coef$mean[6]
*quantile(InALIMENTAR2013, 0.5)+ajuste8.a$coef$mean[7]*quantile(BEMESTAR2014, 0.5)
+ajuste8.a$coef$mean[8]*quantile(INT, 0.5)))*((ajuste8.a$coef$mean[2])))

#Fixando as variáveis no terceiro quartil
impacto_75 = exp(-exp((ajuste8.a$coef$mean[1]+ajuste8.a$coef$mean[2]*xx1
+ajuste8.a$coef$mean[3]*quantile(VEGETABLES2011, 0.75)+ajuste8.a$coef$mean[4]
*quantile(FUMO2012, 0.75)+ajuste8.a$coef$mean[5]*quantile(DESEMP2014, 0.75)
+ajuste8.a$coef$mean[6]*quantile(InALIMENTAR2013, 0.75)+ajuste8.a$coef$mean[7]
*quantile(BEMESTAR2014, 0.75)+ajuste8.a$coef$mean[8]*quantile(INT, 0.75))))
*(exp((ajuste8.a$coef$mean[1]+ajuste8.a$coef$mean[2]*xx1+ajuste8.a$coef$mean[3]
*quantile(VEGETABLES2011, 0.75)+ajuste8.a$coef$mean[4]*quantile(FUMO2012,0.75)
+ajuste8.a$coef$mean[5]*quantile(DESEMP2014, 0.75)+ajuste8.a$coef$mean[6]
*quantile(InALIMENTAR2013, 0.75)+ajuste8.a$coef$mean[7]*quantile(BEMESTAR2014, 0.75)
+ajuste8.a$coef$mean[8]*quantile(INT, 0.75)))*((ajuste8.a$coef$mean[2])))

plot(xx1, impacto_25, type="l",  ylab="Impacto estimado",
xlab="Inatividade física", lwd=2)
lines(xx1, impacto_50,lty=2, type="l", lwd=2)
lines(xx1, impacto_75,lty=3, type="l", lwd=2)
legend("topleft", c("Primeiro Quartil","Mediana","Terceiro Quartil"),
lty = c(1,2,3), bty="n", lwd=2)

#=============================================================================


#=============================================================================
#REGRESSÃO BETA - SIMULAÇÃO DE MONTE CARLO ----------------------------------
#=============================================================================

#Limpar o R
rm(list=ls())

#Pacote
library(betareg)

#Fixar semente
set.seed(53)

#Tamanho amostral
n = 25

#Réplicas de Monte Carlo
R = 10000

#-----------------------------------------------------------------------------
#Coeficientes para a estrutura de regressão da média (mu)
beta=c(-1.9,1.5,0.0)
b2=beta[3]
```

```
b1=beta[2]
b0=beta[1]
#-------------------------------------------------------------------------

#Intercepto
x0=rep(1,n)

#Números gerados aleatoriamente a partir da distribuição uniforme (0,1)
x1=runif(n)
x2=runif(n)

#Gerando variáveis adicionais (para estimação)
x3 = (x2^2)
x4 = rt(n,5)
x5 = runif(n)

#Número de Réplicas da amostra
j=1

#Réplicas da amostra
x0=rep(x0,j)
x1=rep(x1,j)
x2=rep(x2,j)
x3=rep(x3,j)
x4=rep(x4,j)
x5=rep(x5,j)

#Matriz de regressores da média (mu) - (valores das variáveis explicativas)
X=cbind(x0,x1,x2)
X

#Número de linhas da Matriz de regressores
len=nrow(X)
len

#Preditor Linear (Correto)
eta1=X%*%beta
eta1

#=========================================================================
#Função de Ligação LOGIT - Valores para mu
mu = (exp(eta1))/(1+(exp(eta1)))
summary(mu)
#=========================================================================

#-------------------------------------------------------------------------
#Coeficientes para a estrutura de regressão para o parâmetro de precisão (phi)
lambda=c(1.0,7.9)
G1 = lambda[1]
G2 = lambda[2]
#-------------------------------------------------------------------------

#Intercepto
z0 =  rep(1,n)
z0

#Números gerados a partir da distribuição uniforme (0,1)
z1=runif(n)
```

```
z2=runif(n)

#Gerando variáveis adicionais - (estimação)
z3 = (z2^2)
z4 = rt(n,3)

#Réplicas da Amostra
z0=rep(z0,j)
z1=rep(z1,j)
z2=rep(z2,j)
z3=rep(z3,j)
z4=rep(z4,j)

#Matriz de regressores do parâmetro da estrutura de regressão da precisão (phi)
Z=cbind(z0,z2)
Z

#Preditor Linear (phi)
eta2=Z%*%lambda
eta2

#============================================================================
#Função de Ligação SQRT
phi = (eta2^2)
summary(phi)

#Razão entre os valores máximo e mínimo de phi
razao = max(phi)/min(phi)
razao
#============================================================================

#Obtendo os valores de p e de q
p = mu*phi
q = (1-mu)*phi

#Valores Tabelados (Distribuição Normal)
VC10 = qnorm(0.95)
VC5 = qnorm(0.975)
VC1 = qnorm(0.99)

#============================================================================
#SUBSTITUIR A PARTIR DAQUI - LOOP 1 OU LOOP 2!
#Scrip acima igual para ambos os loop's!!
#Script abaixo específico para cálculo de taxas de rejeição e taxas de cobertura!
#(Fazer programas separados para cada Loop!)
#============================================================================

#LOOP 1-------------------------------------------------------------------
#Testes - Beta2 (vetores de zeros)
testC = rep(0,R)
testI2 = rep(0,R)
testI3 = rep(0,R)
testI4 = rep(0,R)
testI5 = rep(0,R)
testI6 = rep(0,R)
testI7 = rep(0,R)

#Cobertura - Beta2 (vetores de zeros)
```

```
CobBeta2 = rep(0,R)
CobBeta2I2 = rep(0,R)
CobBeta2I3 = rep(0,R)
CobBeta2I4 = rep(0,R)
CobBeta2I5 = rep(0,R)
CobBeta2I6 = rep(0,R)
CobBeta2I7 = rep(0,R)

#-------------------------------------------------------------------------
#Teste antes do Loop - Gerando y (verificando os limites da variável resposta)
y = rbeta(len, p, q)
y
summary(y)

#Variância de y
vary = ((mu*(1-mu))/(1+phi))
summary(vary)
#-------------------------------------------------------------------------

#=========================================================================
#>>>>>>>>>>>>>>>TAXAS DE REJEIÇÃO E TAXAS DE COBERTURA<<<<<<<<<<<<<<<<<<<<<<
#=========================================================================

#Loop de Monte Carlo----------------------------------------------------

for(i in 1:R){

y = rbeta(len, p, q)

#Modelo corretamente especificado (geração = estimação)
fit1 <- betareg(y~x1+x2|z2, link="logit", link.phi="sqrt")

#Erros (preditor da média (x3 = x2^2); func. de lig.(mu e phi))
fit2 <- betareg(y~x1+x3|z2, link="cloglog", link.phi="log")

#Erros (preditor da precisão (z1 = runif); func. de lig. da precisão)
fit3 <- betareg(y~x1+x2|z1, link="logit", link.phi="log")

#Erros (preditores (mu e phi)(x3 = x2^2, z1 = runif); func. de lig. (mu e phi))
fit4 <- betareg(y~x1+x3|z1, link="cloglog", link.phi="log")

#Erros (preditor da precisão (z1 = runif); func. de lig. média)
fit5 <- betareg(y~x1+x2|z1, link="cloglog", link.phi="sqrt")

#Erros nos preditores (mu e phi) (x3 = x2^2, z1 = runif)
fit6 <- betareg(y~x1+x3|z1, link="logit", link.phi="sqrt")

#Precisão Fixa
fit7 <- betareg(y~x1+x2, link="logit", link.phi="sqrt")

#-------------------------------------------------------------------------
#beta2 - TESTES E TAXAS DE COBERTURA-------------------------------------
#-------------------------------------------------------------------------

coefC = fit1$coef$mean[3]
epC =  sqrt(vcov(fit1)[3,3])
testC[i] = (coefC-b2)/epC
```

```
ICinfbeta2 = coefC - VC5*epC
ICsupbeta2 = coefC + VC5*epC

CobBeta2[i] = (ifelse(((beta[3]>ICinfbeta2)&(beta[3]<ICsupbeta2)),1,0))

coefI2 = fit2$coef$mean[3]
epI2 =  sqrt(vcov(fit2)[3,3])
testI2[i] = (coefI2-b2)/epI2

ICinfbeta2I2 = coefI2 - VC5*epI2
ICsupbeta2I2 = coefI2 + VC5*epI2

CobBeta2I2[i] = (ifelse(((beta[3]>ICinfbeta2I2)&(beta[3]<ICsupbeta2I2)),1,0))

coefI3 = fit3$coef$mean[3]
epI3 =  sqrt(vcov(fit3)[3,3])
testI3[i] = (coefI3-b2)/epI3

ICinfbeta2I3 = coefI3 - VC5*epI3
ICsupbeta2I3 = coefI3 + VC5*epI3

CobBeta2I3[i] = (ifelse(((beta[3]>ICinfbeta2I3)&(beta[3]<ICsupbeta2I3)),1,0))

coefI4 = fit4$coef$mean[3]
epI4 =  sqrt(vcov(fit4)[3,3])
testI4[i] = (coefI4-b2)/epI4

ICinfbeta2I4 = coefI4 - VC5*epI4
ICsupbeta2I4 = coefI4 + VC5*epI4

CobBeta2I4[i] = (ifelse(((beta[3]>ICinfbeta2I4)&(beta[3]<ICsupbeta2I4)),1,0))

coefI5 = fit5$coef$mean[3]
epI5 =  sqrt(vcov(fit5)[3,3])
testI5[i] = (coefI5-b2)/epI5

ICinfbeta2I5 = coefI5 - VC5*epI5
ICsupbeta2I5 = coefI5 + VC5*epI5

CobBeta2I5[i] = (ifelse(((beta[3]>ICinfbeta2I5)&(beta[3]<ICsupbeta2I5)),1,0))

coefI6 = fit6$coef$mean[3]
epI6 =  sqrt(vcov(fit6)[3,3])
testI6[i] = (coefI6-b2)/epI6

ICinfbeta2I6 = coefI6 - VC5*epI6
ICsupbeta2I6 = coefI6 + VC5*epI6

CobBeta2I6[i] = (ifelse(((beta[3]>ICinfbeta2I6 )&(beta[3]<ICsupbeta2I6)),1,0))

coefI7 = fit7$coef$mean[3]
epI7 =  sqrt(vcov(fit7)[3,3])
testI7[i] = (coefI7-b2)/epI7

ICinfbeta2I7 = coefI7 - VC5*epI7
ICsupbeta2I7 = coefI7 + VC5*epI7

CobBeta2I7[i] = (ifelse(((beta[3]>ICinfbeta2I7 )&(beta[3]<ICsupbeta2I7)),1, 0))
```

```
}
#-------------------------------------------------------------------------------
#>>>>>>>>>>!!!FIM DO LOOP!!!<<<<<<<<<<<<<==================================
#-------------------------------------------------------------------------------

taxa10 = ifelse( (abs(testC) > VC10),1,0)
taxa5 = ifelse( (abs(testC) > VC5),1,0)
taxa1 = ifelse( (abs(testC) > VC1),1,0)

taxa10I2 = ifelse( (abs(testI2) > VC10),1,0)
taxa5I2 = ifelse( (abs(testI2) > VC5),1,0)
taxa1I2 = ifelse( (abs(testI2) > VC1),1,0)

taxa10I3 = ifelse( (abs(testI3) > VC10),1,0)
taxa5I3 = ifelse( (abs(testI3) > VC5),1,0)
taxa1I3 = ifelse( (abs(testI3) > VC1),1,0)

taxa10I4 = ifelse( (abs(testI4) > VC10),1,0)
taxa5I4 = ifelse( (abs(testI4) > VC5),1,0)
taxa1I4 = ifelse( (abs(testI4) > VC1),1,0)

taxa10I5 = ifelse( (abs(testI5) > VC10),1,0)
taxa5I5 = ifelse( (abs(testI5) > VC5),1,0)
taxa1I5 = ifelse( (abs(testI5) > VC1),1,0)

taxa10I6 = ifelse( (abs(testI6) > VC10),1,0)
taxa5I6 = ifelse( (abs(testI6) > VC5),1,0)
taxa1I6 = ifelse( (abs(testI6) > VC1),1,0)

taxa10I7 = ifelse( (abs(testI7) > VC10),1,0)
taxa5I7 = ifelse( (abs(testI7) > VC5),1,0)
taxa1I7 = ifelse( (abs(testI7) > VC1),1,0)

#-------------------------------------------------------------------------------

#Tamanho dos testes (Modelo Corretamente Especificado)
tam10 = (sum(taxa10)/R)*100
tam5 = (sum(taxa5)/R)*100
tam1 = (sum(taxa1)/R)*100

#Tamanho dos testes (Modelos com Erros de Especificação)
tam10I2 = (sum(taxa10I2)/R)*100
tam5I2 = (sum(taxa5I2)/R)*100
tam1I2 = (sum(taxa1I2)/R)*100

tam10I3 = (sum(taxa10I3)/R)*100
tam5I3 = (sum(taxa5I3)/R)*100
tam1I3 = (sum(taxa1I3)/R)*100

tam10I4 = (sum(taxa10I4)/R)*100
tam5I4 = (sum(taxa5I4)/R)*100
tam1I4 = (sum(taxa1I4)/R)*100

tam10I5 = (sum(taxa10I5)/R)*100
tam5I5 = (sum(taxa5I5)/R)*100
tam1I5 = (sum(taxa1I5)/R)*100

tam10I6 = (sum(taxa10I6)/R)*100
```

```
tam5I6 = (sum(taxa5I6)/R)*100
tam1I6 = (sum(taxa1I6)/R)*100

tam10I7 = (sum(taxa10I7)/R)*100
tam5I7 = (sum(taxa5I7)/R)*100
tam1I7 = (sum(taxa1I7)/R)*100

#-------------------------------------------------------------------------------
#RESUMO (TAMANHO DOS TESTES E TAXAS DE COBERTURA)-----------------------------
#-------------------------------------------------------------------------------

#TAMANHO DOS TESTES
#10%
tam10
tam10I2
tam10I3
tam10I4
tam10I5
tam10I6
tam10I7

#5%
tam5
tam5I2
tam5I3
tam5I4
tam5I5
tam5I6
tam5I7

#1%
tam1
tam1I2
tam1I3
tam1I4
tam1I5
tam1I6
tam1I7

#TAXAS DE COBERTURA
#Taxas de cobertura - Beta2
(sum(CobBeta2)/R)*100
(sum(CobBeta2I2)/R)*100
(sum(CobBeta2I3)/R)*100
(sum(CobBeta2I4)/R)*100
(sum(CobBeta2I5)/R)*100
(sum(CobBeta2I6)/R)*100
(sum(CobBeta2I7)/R)*100


#LOOP 2-----------------------------------------------------------------------
#==============================================================================
#SUBSTITUIR A PARTIR DAQUI - LOOP 2!
#Script abaixo específico para cálculo de víes relativo e EQM!
#Mesma geração de dados anterior!! (Fazer programas separados para cada Loop!!)
#==============================================================================

#Vetores de zeros
```

```
VRmu = rep(0,R)
VRmuF2 = rep(0,R)
VRmuF3 = rep(0,R)
VRmuF4 = rep(0,R)
VRmuF5 = rep(0,R)
VRmuF6 = rep(0,R)
VRmuF7 = rep(0,R)

varF1 = rep(0,R)
varF2 = rep(0,R)
varF3 = rep(0,R)
varF4 = rep(0,R)
varF5 = rep(0,R)
varF6 = rep(0,R)
varF7 = rep(0,R)

#-------------------------------------------------------------------------------
#Teste antes do Loop - Gerando y (verificando os limites da variável resposta)
y = rbeta(len, p, q)
y
summary(y)

#Variância de y
vary = ((mu*(1-mu))/(1+phi))
summary(vary)
#-------------------------------------------------------------------------------

#===============================================================================
#>>>>>>>>>>>>>>>>LOOP PARA CALCULAR VÍES E ERRO QUADRÁTICO MÉDIO<<<<<<<<<<<<<<<
#===============================================================================

#Loop de Monte Carlo ----------------------------------------------------------

for(i in 1:R){
y = rbeta(len, p, q)

#-------------------------------------------------------------------------------
#VIÉS RELATIVO E ERRO QUADRÁTICO MÉDIO----------------------------------------
#-------------------------------------------------------------------------------

#FIT1 - Modelo corretamente especificado
fit1 <- betareg(y ~x1+x2|z2, link="logit", link.phi="sqrt")

muLogit = predict(fit1, type="response")
VRmu[i] = (sum(abs((muLogit-mu)/mu)))/length(muLogit)
varF1[i] = var(muLogit)

#-------------------------------------------------------------------------------

#FIT2 - Erros (preditor da média (x3 = x2^2); func. de lig.(mu e phi))
fit2 <- betareg(y ~x1+x3|z2, link="cloglog", link.phi="log")

muLogitF2 = predict(fit2, type="response")
VRmuF2[i] = (sum(abs((muLogitF2-mu)/mu)))/length(muLogitF2)
varF2[i] = var(muLogitF2)

#-------------------------------------------------------------------------------
```

```
#FIT3 - Erros (preditor da precisão (z1 = runif); func. de lig. da precisão)
fit3 <- betareg(y ~x1+x2|z1, link="logit", link.phi="log")

muLogitF3 = predict(fit3, type="response")
VRmuF3[i] = (sum(abs((muLogitF3-mu)/mu)))/length(muLogitF3)
varF3[i] = var(muLogitF3)

#-------------------------------------------------------------------------

#FIT4 - Erros (preditores (mu e phi)(x3 = x2^2, z1 = runif);
#func. de lig. (mu e phi))
fit4 <- betareg(y ~x1+x3|z1, link="cloglog", link.phi="log")

muLogitF4 = predict(fit4, type="response")
VRmuF4[i] = (sum(abs((muLogitF4-mu)/mu)))/length(muLogitF4)
varF4[i] = var(muLogitF4)

#-------------------------------------------------------------------------

#FIT5 - Erros (preditor da precisão (z1 = runif); func. de lig. média)
fit5 <- betareg(y ~x1+x2|z1, link="cloglog", link.phi="sqrt")

muLogitF5 = predict(fit5, type="response")
VRmuF5[i] = (sum(abs((muLogitF5-mu)/mu)))/length(muLogitF5)
varF5[i] = var(muLogitF5)

#-------------------------------------------------------------------------

#FIT 6 - Erros nos preditores (mu e phi) (x3 = x2^2, z1 = runif)
fit6 <- betareg(y ~x1+x3|z1, link="logit", link.phi="sqrt")

muLogitF6 = predict(fit6, type="response")
VRmuF6[i] = (sum(abs((muLogitF6-mu)/mu)))/length(muLogitF6)
varF6[i] = var(muLogitF6)

#-------------------------------------------------------------------------

#FIT 7 - Precisão Fixa
fit7 <- betareg(y ~x1+x2, link="logit", link.phi="sqrt")

muLogitF7 = predict(fit7, type="response")
VRmuF7[i] = (sum(abs((muLogitF7-mu)/mu)))/length(muLogitF7)
varF7[i] = var(muLogitF7)

}
#-------------------------------------------------------------------------
#>>>>>>>>>>!!!FIM DO LOOP!!!<<<<<<<<<<<<<==================================
#-------------------------------------------------------------------------

#VIÉS  RELATIVO MÉDIO
(sum(VRmu)/R)
(sum(VRmuF2)/R)
(sum(VRmuF3)/R)
(sum(VRmuF4)/R)
(sum(VRmuF5)/R)
(sum(VRmuF6)/R)
(sum(VRmuF7)/R)
```

```
#ERRO QUADRÁTICO MÉDIO
vF1 = (sum(varF1))/R
vF2 = (sum(varF2))/R
vF3 = (sum(varF3))/R
vF4 = (sum(varF4))/R
vF5 = (sum(varF5))/R
vF6 = (sum(varF6))/R
vF7 = (sum(varF7))/R

viesF1 = (sum(VRmu)/R)
viesF2 = (sum(VRmuF2)/R)
viesF3 = (sum(VRmuF3)/R)
viesF4 = (sum(VRmuF4)/R)
viesF5 = (sum(VRmuF5)/R)
viesF6 = (sum(VRmuF6)/R)
viesF7 = (sum(VRmuF7)/R)

EQMF1 = vF1+(viesF1^2)
EQMF2 = vF2+(viesF2^2)
EQMF3 = vF3+(viesF3^2)
EQMF4 = vF4+(viesF4^2)
EQMF5 = vF5+(viesF5^2)
EQMF6 = vF6+(viesF6^2)
EQMF7 = vF7+(viesF7^2)

EQMF1
EQMF2
EQMF3
EQMF4
EQMF5
EQMF6
EQMF7

#=============================================================================
```