UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELOS DE DECISÃO E SAÚDE

MODELAGEM DA OBESIDADE ADULTA NAS NAÇÕES: UMA ANÁLISE VIA MODELOS DE REGRESSÃO BETA E QUANTÍLICA

Saul de Azevêdo Souza

João Pessoa/PB

2017

SAUL DE AZEVÊDO SOUZA

MODELAGEM DA OBESIDADE ADULTA NAS NAÇÕES: UMA ANÁLISE VIA MODELOS DE REGRESSÃO BETA E QUANTÍLICA

Dissertação apresentada ao Programa de Pós-Graduação em Modelos de Decisão e Saúde – Nível Mestrado do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, como requisito regulamentar para a obtenção do título de Mestre.

Linha de pesquisa: Modelos de Decisão

Orientadores:
Profa. Dra. Tatiene Correia de Souza
Profa. Dra. Caliandra Maria Bezerra Luna Lima

João Pessoa/PB

2017

S729m Souza, Saul de Azevêdo.
Modelagem da obesidade adulta nas nações: uma análise via modelos de regressão beta e quantílica / Saul de Azevêdo Souza. – João Pessoa, 2017.
75 f. : il. –

Orientadoras: Tatiene Correia de Souza, Caliandra Maria Bezerra Luna Lima
Dissertação (Mestrado) – UFPB/CCEN

1. Obesidade. 2. Regressão beta. 3. Regressão quantílica. I. Título.

UFPB/BC CDU: 613.25(043)

SAUL DE AZEVÊDO SOUZA

MODELAGEM DA OBESIDADE ADULTA NAS NAÇÕES: UMA ANÁLISE VIA MODELOS DE REGRESSÃO BETA E QUANTÍLICA

João Pessoa, 20 de Fevereiro de 2017.

BANCA EXAMINADORA

Profa. Dra. Tatiene Correia de Souza
Orientador (DE/UFPB)

Profa. Dra. Caliandra Maria B. Luna Lima
Orientador (DFP/UFPB)

Prof. Dr. Luiz Medeiros de A. Lima Filho
Membro Interno (DE/UFPB)

Prof. Dr. Rodrigo pinheiro de Toledo Vianna
Membro Interno (DN/UFPB)

Profa. Dra. Liana Clébia de Morais Pordeus
Membro Externo (DFP/UFPB)

Este trabalho é carinhosamente dedicado
aos meus pais, Sílvio e Maria, e ao meu irmão Sílvio Jr.

AGRADECIMENTOS

Agradeço primeiramente à DEUS, pelo seu cuidado para comigo, guiando e conduzindo-me para o melhor caminho. Por me proteger, ajudar-me a tomar as decisões e enfrentar os obstáculos que a vida me propôs.

Aos meus pais, Sílvio e Maria, pelo apoio, cuidado e dedicação, aos quais dedico todas as minhas conquistas.

Ao meu irmão Sílvio Jr., aos meus avós Lourival e Herenilde, aos meus tios, Sátiro e Sérgio, as minhas tias Sônia e Marluce, e aos meus primos, Sátiro, Samantha, Sarah, Aline e Sílvia, pelo apoio, incentivo e amizade.

Aos meus amigos, Jodavid, Marina, Maizza, Alisson, Aldine, Hemmylly, Alinne e Francisco pela amizade, companheirismo, descontração, incentivo e paciência. Por tornar meus dias mais divertidos e agradáveis.

As professoras Tatiene e Caliandra, pela orientação constante, amizade, confiança, paciência e respeito. Pelos valiosos ensinamentos e direcionamentos para a conclusão desta dissertação.

Aos professores Hemílio, Ulisses e João Agnaldo pela amizade e conselhos. Por engrandecer meus conhecimentos ao longo desses anos.

Aos docentes da Banca Examinadora pelas contribuições e sugestões, desde a qualificação do projeto de pesquisa.

A todos os Professores do DE-UPPB, por contribuírem para minha formação acadêmica.

A todos os funcionários do Departamento de Estatística e do Programa de Pós-Graduação em Modelos de Decisão e Saúde.

A CAPES, pelo apoio financeiro.

"Se eu vi mais longe, foi por estar sobre ombros de gigantes"
Isaac Newton

RESUMO

Nesta dissertação são abordados os modelos de regressão beta com dispersão variável e de regressão quantílica. Para tanto, foi feita uma introdução com objetivo de motivar sua discussão em estudos epidemiológicos, enfatizando a problematização em torno da obesidade. A aplicação destes métodos considerou um conjunto de dados reais, obtidos a partir de fontes de informação pública, referente a obesidade adulta nas nações no ano de 2014. Após a análise descritiva dos dados verificou-se que 50% das nações apresentam valores da proporção de adultos obesos maiores do que 0.20. Além disso, visualizando o mapa da obesidade por nação constatou-se que a maior concentração de países com menores valores de obesidade encontra-se nos continentes da Ásia e África. Por outro lado, as maiores concentrações de obesos encontram-se nos continentes da América e Europa. Ainda, a partir da análise gráfica do *box-plot* foi observado uma possível diferença nas proporções de adultos obesos entre os continentes da América e Europa com os da África e Ásia. Após ajustar os modelos de regressão beta e quantílica verificou-se que as covariáveis consumo médio de álcool em litros por pessoa, porcentagem de atividade física insuficiente e porcentagem da população que vivem em áreas urbanas apresentam efeito positivo sobre a variável resposta. Ou seja, individualmente tais covariáveis tendem a aumentar os valores de obesidade nos países quando as demais covariáveis permanecem constantes. Além disso, a variável expectativa de vida em anos apresentou efeito positivo e foi significativa apenas para o modelo de regressão beta com dispersão variável. Por fim, analisando as medidas de erros de previsão verificou-se que as estimativas oriundas da regressão beta são mais precisas quando avaliado o erro quadrático médio e o erro percentual total. Portanto, para questões de predizer valores referentes a obesidade adulta nas nações em 2014 o modelo de regressão beta com dispersão variável se mostrou mais adequado para tal propósito.

Palavras-chave: Obesidade. Regressão beta. Regressão quantílica.

ABSTRACT

In this dissertation the beta regression models with variable dispersion and quantile regression are discussed. Therefore, an introduction was made with the objective of motivating its discussion in epidemiological studies, emphasizing the problematization around obesity. The application of these methods considered a real data set, obtained from public information sources, referring to adult obesity in the nations in the year 2014. After the descriptive analysis of the data it was verified that 50% of the nations present values of the proportion of obese adults greater than 0.20. In addition, viewing the obesity map by nation showed that the highest concentration of countries with the lowest obesity values is found in the continents of Asia and Africa. On the other hand, the highest concentrations of obese are found in the continents of America and Europe. Also, from the graphical analysis of the box-plot a possible difference in the proportions of obese adults between the continents of America and Europe with those of Africa and Asia was observed. After adjusting the beta and quantile regression models it was verified that the covariates average alcohol consumption in liters per person, percentage of insufficient physical activity and percentage of the population living in urban areas have a positive effect on the response variable. That is, individually such covariables tend to increase obesity values in the countries when the other covariables remain constant. In addition, the life expectancy variable in years presented a positive effect and was significant only for the variable regression beta regression model. Finally, analyzing the measures of prediction errors, it was verified that the estimates from the beta regression are more accurate when the mean square error and the total percentage error were evaluated. Therefore, for questions of predicting values for adult obesity in the nations in 2014, the beta regression model with variable dispersion was more suitable for this purpose.

Keywords:Obesity. Beta regression. Quantile regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALC	Média do consumo em litros de álcool puro por pessoa em um ano
CA	Circunferência Abdominal
CC	Circunferência da Cintura
DCNT	Doenças Crônicas não Transmissíveis
EDUC	Gastos com a educação como porcentagem da despesa total do governo
i.i.d.	Independentes e identicamente distribuídos
IMC	Índice de Massa Corporal
INAT	Porcentagem das pessoas que praticam atividade física insuficiente
INT	Intercepto do modelo
MQO	Mínimos Quadrados Ordinários
n.i.d.	Não identicamente distribuídos
*OB*2014	Proporção de adultos obeso
OECD	*Organization for Economic Co-operation and Development*
OMS	Organização Mundial da Saúde
RCQ	Relação Cintura-Quadril
SBCBM	Sociedade Brasileira de Cirurgias Bariátricas e Metabólicas
URB	Porcentagem da população que vivem em áreas urbanas
VIDA	Expectativa de vida ao nascer
EQM	Erro quadrático médio
EAM	Erro absoluto médio
EPT	Erro percentual total

SUMÁRIO

1 INTRODUÇÃO

A obesidade é considerada uma doença epidêmica de grande repercussão no cenário mundial, recorrente tanto em países desenvolvidos como naqueles em desenvolvimento (GIGANTE et al., 2006; MARIATH et al., 2007). Essa doença pode apresentar origem genética e metabólica, ser influenciada por fenômenos ambientais, sociais, culturais e econômicos, ou ainda estar relacionada a fatores demográficos e ao sedentarismo (PUGLIA, 2004). A OMS (Organização Mundial da Saúde) define a obesidade como a excessiva concentração de gordura que pode prejudicar a saúde do indivíduo. Portanto, o consumo de alimentos altamente energéticos e a falta de atividade física se destacam por facilitar o ganho de calorias e diminuir o gasto de energia corporal ao longo do dia, tornando a balança energética do indivíduo positiva e facilitando o acúmulo de gordura. A obesidade está inserida no grupo de doenças crônicas não transmissíveis (DCNT) estando associada ou sendo fator de risco para outras complicações como diabetes, hipertensão e doenças cardiovasculares (DUNCAN et al., 2012; PINHEIRO; FREITAS; CORSO, 2004). As DCNTs são doenças multifatoriais que se desenvolvem ao longo da vida e são de longa duração. Além disso, elas se apresentam como um sério problema de saúde pública, sendo classificadas como as principais causas de mortes no mundo. Por exemplo, em 2008 elas foram responsáveis por cerca de 63% das mortes no mundo, sendo 80% delas relacionadas a países de baixa e média renda (Ministério da Saúde, 2011).

Alguns autores sugerem que o aumento da obesidade ou sobrepeso no mundo deve-se a transição de uma alimentação saudável para um consumo exagerado de alimentos altamente energéticos, destacando uma dieta mais rica em gorduras animais, açúcares ou lipídeos, favorecendo o aumento da adiposidade. Além disso, a falta de incentivos a práticas de atividade física, muitas vezes ocasionada pelo o avanço tecnológico ou mudanças no estilo de vida, atribuem as pessoas maiores chances de se tornarem obesas, o que por sua vez é fator de risco para muitas doenças como a diabetes (FRANCISCHI et al., 2000; ANDRADE; PEREIRA; SICHIERI, 2003; GIGANTE et al., 2006; MARIATH et al., 2007). Por exemplo, ao estudar a prevalência da obesidade na sociedade moderna Jung (1997) verificou que o risco de se adquirir diabetes mellitus pode aumentar consideravelmente quando o *IMC* do indivíduo ultrapassa $35 kg/m^2$, em cerca de 42 e 93 vezes para homens e mulheres, respectivamente.

A obesidade pode ser diagnosticada a partir de algumas medidas antropométricas, contudo seus valores ou interpretações podem variar de acordo com o sexo e a faixa etária do indivíduo. Por exemplo, pode-se considerar o *IMC* (Índice de Massa Corporal), o *RCQ* (Relação Cintura-Quadril), o *CA* (Circunferência Abdominal) e o *CC* (Circunferência da Cintura). Essas medidas são utilizadas como forma de se estimar o percentual de gordura do indivíduo, podendo também ser um forte indicador para doenças cardiovasculares.

O *IMC* é uma das medidas mais utilizada para se avaliar a concentração de gordura em adultos, sendo definida como a razão entre o peso do indivíduo dado em quilogramas (kg) e sua altura ao quadrado (m^2). Dessa forma, a obesidade pode ser dividida em três níveis, como apresentado em Stol et al. (2011), a saber: grau I com $30.0 \leq IMC \leq 34.9$, grau II com $35.0 \leq IMC \leq 39.9$ e grau III com $IMC \geq 40.0$. Cabrera e Filho (2001) objetivando identificar a prevalência de obesos em uma população idosa utilizaram a medida *RCQ*. Os autores verificaram que os indivíduos que apresentaram *RCQ* maior que o percentil 75 mostraram se mais dispostos a certas doenças, como hipertensão arterial e diabetes mellitus. Burgos et al. (2013) apresentam o *CC* como uma medida antropométrica bastante utilizada em populações adultas para avaliar o acúmulo da gordura visceral. Por fim, Neto et al. (2008) apresentam o *CA* como uma medida capaz de fornecer valores aproximados para a concentração de gordura abdominal e total. Assim, homens e mulheres que apresentarem $CA \geq 102cm$ e $CA \geq 88cm$, respectivamente, serão classificados como obesos.

Apesar da grande relevância dos fatores genéticos no desenvolvimento dessa doença, a obesidade pode ser prevenida ou tratada. A obesidade exerce um forte impacto sobre a economia devido aos enormes gastos no setor de saúde relacionados a essa doença. Por exemplo, os Estados Unidos estão entre os países desenvolvidos com maior proporção de adultos obesos, são cerca de 35% de sua população total. Além disso, os gastos relacionados a essa doença em 2000 chegaram a 11 bilhões de dólares (ARTERBURN; MACIEJEWSKI; TSEVAT, 2005). No Brasil, segundo as estimativas de Bahia et al. (2012), os custos totais em um ano com as doenças associadas ao sobrepeso e a obesidade chegam a 2.1 bilhões de dólares. Além disso, 68.4% desses custo estão envolvidos com as hospitalizações e 31.6% estão envolvidos com os procedimentos ambulatoriais. Os autores alertam ainda que apenas 10% desses custos são provenientes do tratamento contra o sobrepeso e a obesidade. Ou seja, boa parte dos gastos no setor de saúde são devido aos problemas ocasionados por essa doença.

A OECD (*Organization for Economic Co-operation and Development*) tem o objetivo de promover políticas que melhorem a economia e o bem-estar social das pessoas ao redor do mundo. Ela é constituída por 34 países membros, entre eles os desenvolvidos e os em desenvolvimento. Em seu relatório para o ano de 2014 mostrou que nos últimos cinco anos os países Canadá, Inglaterra, Itália, Coreia, Espanha e Estados Unidos apresentaram um crescimento modesto ou praticamente estável do sobrepeso e da obesidade. Por outro lado, os países Austrália, França, México e Suíça, apresentaram um crescimento de 2% a 3%, não havendo nenhum indício da redução ou contenção dessa epidemia entre as nações. Além disso, é estimado que os gastos das nações no setor de saúde relacionado a essa doença variam de 1% a 3%, podendo ser maior quando associado a outras complicações (OECD, 2014).

Segundo a Sociedade Brasileira de Cirurgias Bariátricas e Metabólicas (SBCBM) a primeira opção para redução do excesso de peso é o tratamento clínico. Esse tratamento consiste em dieta, atividade física, medicação e acompanhamento de endocrinologista e nutricionista,

tendo como objetivo conscientizar os indivíduos sobre as práticas para uma melhor qualidade de vida. Contudo, para os casos em que o tratamento clínico não traz resultados satisfatórios, existe a possibilidade do tratamento cirúrgico, cabendo ao médico indicar a opção mais apropriada e segura com base no *IMC*, idade e tempo de doença do paciente. Portanto compreender melhor sobre o tema da obesidade é de fundamental importância para melhorar a qualidade de vida das pessoas e buscar medidas preventivas a fim de evitar fatores agravantes e reduzir o desperdício de dinheiro por falta de planejamento.

Na literatura existe uma ampla quantidade de métodos estatísticos que podem ser utilizados para modelar dados. Contudo, durante a revisão bibliográfica foi visto que os modelos de regressão logística têm bastante destaque nos estudos epidemiológicos em relação aos outros métodos. Tal fato muitas vezes é justificado pelo desconhecimento a respeito das diversas técnicas de análises de regressão. Por exemplo,

- Antiporta et al. (2015) buscando determinar a associação entre obesidade e tempo de residência em áreas urbanas, no Peru, verificaram por meio da análise multivariada que para cada 10 anos de residência nas áreas urbanas os imigrantes rurais-urbanos tinham em média 12% a mais de prevalência da obesidade.

- Martínez-González et al. (1999) com o objetivo de verificar a associação entre as práticas de atividade física no lazer e a obesidade adulta na população europeia utilizaram o modelo de regressão logística. O autores observaram que o tempo de lazer em conjunto com tais práticas apresentavam-se como fator de proteção para a obesidade, com *odds ratio* (*OR*) de 0.52 e intervalo com 95% de confiança dado por $(0.43, 0.64)$. Além disso, os autores verificaram que a atividade física no tempo de lazer era inversamente proporcional ao *IMC*.

- Shelton e Knott (2014) investigaram a contribuição da ingestão de calorias derivadas do álcool para a relação álcool-obesidade, a partir de um Inquérito de Saúde para a Inglaterra em 2006. Os autores por meio da regressão logística verificaram uma associação entre as calorias do álcool e a obesidade. Além disso, eles alertam que as calorias provenientes do álcool podem contribuir significativamente para o aumento da obesidade.

Os modelos de regressão logística são uteis para investigar a influência de determinado conjunto de covariáveis sobre o desfecho em questão. Para tanto, é necessário que a variável resposta, Y, tenha caráter binário. Ou seja, assuma dois valores $Y = 0$ ou $Y = 1$ em que este último é considerado o evento de interesse ou "sucesso". Dessa forma, em situações que a variável resposta não segue distribuição Binominal(n, p), com n repetições do evento e probabilidade de sucesso fixa p, é necessário aplicar alguma transformação nesta variável para que ela possa se adequar ao modelo proposto. O ideal seria ter o conhecimento a respeito dos diferentes tipos de modelos propostos na literatura, tornando mais adequada as análises dos conjuntos de dados. Ou

seja, permitindo que a relação entre a variável resposta e o desfecho se defina da melhor maneira possível. Dessa forma o objetivo desta dissertação consiste em motivar o uso dos modelos de regressão beta e quantílica para modelar dados epidemiológicos com restrição no intervalo (0,1). Neste contexto, pode-se pensar em modelar a prevalência da obesidade adulta nas nações em vez do indivíduo como unidade de interesse. A obesidade se trata de um problema mundial e envolve diversos países, inclusive o Brasil. Portanto, é útil compreender o comportamento de sua distribuição e identificar os fatores relacionados com o aumento dessa doença nos últimos anos.

A respeito da organização desta dissertação, no Capítulo 2, é apresentado os objetivos gerais e específicos abordados neste estudo. No Capítulo 3, são apresentados os modelos de regressão beta e quantílica. No Capítulo 4, é feito um levantamento dos procedimentos metodológicos. No Capítulo 5, é feita uma breve descrição dos dados e as especificações dos modelos propostos. Por fim, no Capítulo 6 são pontuadas as considerações finais.

2 OBJETIVOS

2.1 OBJETIVO GERAL

O objetivo principal desta dissertação é estudar a abordagem dos modelos de regressão beta e quantílica como uma ferramenta de modelagem em estudos epidemiológicos.

2.2 OBJETIVOS ESPECÍFICOS

Como objetivos específicos pode-se citar:

- Utilizar análise descritiva para extrair informações importantes a respeito das variáveis abordadas no estudo.
- Construir mapa com a prevalências da obesidade por nação objetivando visualizar os níveis dessa doença.
- Descrever os modelos de regressão beta e quantílica.
- Identificar quais covariáveis estão relacionadas com o aumento da obesidade a partir da modelagem dos dados.
- Comparar os resultados obtidos por meio dos modelos propostos e identificar o modelo mais adequado para modelar a prevalência da obesidade nas nações.

3 REFERENCIAL TEÓRICO

3.1 MODELO DE REGRESSÃO CLÁSSICO

Em diversas situações práticas, sejam elas observacionais ou experimentais, pesquisadores buscam compreender e explicar os fenômenos ocorridos em diversas áreas da ciência. Para isso, é possível utilizar os modelos de regressão, pois estes permitem expressar a relação existente entre uma variável resposta, Y, e as demais covariáveis independentes, $(X_1,\ldots,X_p)$, abordadas no estudo. O modelo de regressão linear é um dos métodos mais conhecidos, devido a facilidade de interpretação dos seus parâmetros por parte dos pesquisadores, além de encontra-se disponível em diversos programas estatísticos. A estrutura de regressão deste modelo pode ser definida da seguinte maneira

$$Y = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_p X_{pt} + \varepsilon_t, \quad \text{com} \quad t = 1,\ldots,n,$$

em que Y é a variável resposta ou desfecho, $(X_1,\ldots,X_p)$ são as covariáveis independentes as quais se acredita serem responsáveis pela variabilidade de Y e $(\beta_0,\ldots,\beta_p)$ são os parâmetros desconhecidos a serem estimados. Aqui, temos que os erros, ε_t, são uma sequência aleatória, independente e normalmente distribuída com $E(\varepsilon_t) = 0$ e $\text{Var}(\varepsilon_t) = \sigma^2$.

Nesta classe de modelos é usual assumir os seguintes pressupostos:

- $E(\varepsilon_t) = 0, t = 1,...,n$;
- $\text{Var}(\varepsilon_t) = \sigma^2, 0 < \sigma^2 < \infty, t = 1,...,n$;
- $E(\varepsilon_t \varepsilon_s) = 0, \forall\, t \neq s$;
- Os erros $\varepsilon_1,...,\varepsilon_n$ seguem distribuição normal $\mathcal{N}(0,\sigma^2)$.

Para realizar inferências a respeito de β pode-se pensar em utilizar o método dos mínimos quadrados ordinários, pois este apresenta propriedades ótimas quando os erros seguem distribuição normal. Este método em termos matriciais pode ser expresso da seguinte maneira

$$\Sigma_{t=1}^{n} \varepsilon^\top \varepsilon = \Sigma_{t=1}^{n} (Y - X\beta)^\top (Y - X\beta), \quad \text{com} \quad t = 1,\ldots,n,$$

contudo suas estimativas são facilmente influenciadas por observações atípicas uma vez que este método atribui o mesmo peso a cada observação durante o processe de estimação dos parâmetros (DRAPER; SMITH, 1998).

Na literatura os modelos de regressão clássicos são utilizados para explicar diversos fenômenos aleatórios, mesmo quando não é razoável assumir distribuição de probabilidade normal para os dados. Neste contexto, muitas vezes é necessário aplicar transformações na

variável resposta, a exemplo da família de transformações de Box e Cox (1964) (GERACI; JONES, 2015; PAULA, 2004). Ou seja, tal transformação permite alterar o valor observado de $Y \in \mathbb{R}^+$ em

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

em que λ é um parâmetro desconhecido. Além disso, $\forall\, \lambda \neq 0$ a variável transformada z não está definida no conjunto $\mathbb{R}$, mas no intervalo $(-1/\lambda, \infty)$ (GERACI, 2016). Vale ressaltar que a Box-Cox é uma família de transformações de único parâmetro que se aplica a variáveis com restrição no $\mathbb{R}^+$, ou seja, $y > 0$. O objetivo dessa transformação é tornar a variável z aproximadamente normal, linear e com variância constante (cenário homoscedástico). Contudo, raramente isto ocorre para um único valor de λ (BOX; DRAPER, 1987).

Dessa forma, torna-se útil conhecer os diferentes métodos de regressão propostos na literatura, com o objetivo de modelar determinado conjunto de dados sem que haja a necessidade de aplicar transformações na variável resposta. Dito isto, os modelos de regressão beta e quantílica são fortes candidatos para explicar da melhor maneira possível a relação existente entre uma variável resposta, a exemplo das prevalências, e as demais covariáveis ditas serem responsáveis por sua variabilidade.

3.2 MODELO DE REGRESSÃO BETA

O modelo de regressão normal linear é muitas vezes utilizado em situações inapropriadas, a exemplo, quando a variável resposta encontra-se restrita a algum intervalo $[a,b]$, com a e $b \in \mathbb{R}$, ou mesmo quando esta apresenta uma distribuição muito assimétrica dificultando a aproximação normal. Kieschnick e McCullough (2003) estudando a modelagem de variáveis restritas ao intervalo (0,1), foram capazes de identificar sete tipos de modelos utilizados na literatura para analisar dados com esta restrição, a saber: o modelo normal linear, o modelo logito, o modelo normal censurado, o modelo normal não-linear, o modelo baseado na distribuição beta, o modelo baseado na distribuição simplex e o modelo de quasi-verossimilhança. Os autores ainda discutem o uso inapropriado do estimador de mínimos quadrados ordinários neste cenário. Isto ocorre por que tal abordagem é contraria a duas condições: como a variável resposta assume valores restritos ao intervalo $(0,1)$, a média das respostas deve ser não-linear nos parâmetros de regressão e sua variância deve ser heteroscedástica, uma vez que se aproxima de zero à medida que a média se aproxima dos limites do intervalo. Além disso, proporções não estão definidas no conjunto dos reais tornando inadequada a aproximação normal. Por fim, os autores recomendam o uso de regressão baseada na distribuição beta ou um modelo de regressão quasi-verossimilhança (PAPKE; WOOLDRIDGE, 1996) para dados com este tipo de restrição.

Dessa forma, para modelar dados assimétricos e restritos ao intervalo (0,1) Ferrari e Cribari-Neto (2004) propuseram o modelo de regressão beta. Essa classe de modelos assume

que a distribuição de probabilidade da variável resposta é a beta, ou seja, os dados devem estar dispostos como taxas ou proporções, equivalente as prevalências em estudos epidemiológicos. Diferente dos modelos normais lineares seu estimador usual é o de máxima verossimilhança. Dessa forma, é possível estimar o vetor de parâmetros desconhecidos com base na função de verossimilhança. Entretanto esse modelo não pode ser utilizado quando os dados contém valores zeros e/ou uns, ou seja, quando alguma observação equivale aos limites do intervalo (PEREIRA, 2010).

A distribuição de probabilidade beta suposta para a variável reposta, y, apresenta a seguinte densidade de probabilidade

$$f(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

em que os parâmetros α e β assumem valores positivos e $\Gamma(.)$ é a função gama. A média e a variância são respectivamente $\mathrm{E}(y) = \alpha/(\alpha+\beta)$ e $\mathrm{Var}(y) = (\alpha+\beta)/((\alpha+\beta)^2(\alpha+\beta+1))$.

Ferrari e Cribari-Neto (2004) propuseram uma reparametrização para a densidade beta permitindo a modelagem da resposta média através de uma estrutura de regressão que envolvesse também um parâmetro de precisão. Considerando $\phi = \alpha+\beta$ e $\beta = (1-\mu)\phi$ temos que a densidade beta reparametrizada tem a seguinte 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 a variável resposta, y, e a resposta média, μ, estão restritas ao intervalo (0,1). Aqui, $\mathrm{E}(y)$ = μ e $\mathrm{Var}(y) = \mathrm{Var}(\mu)/(1+\phi)$, sendo $\mathrm{Var}(\mu) = \mu(1-\mu)$ a "função de variância", e ϕ pode ser interpretado como o parâmetro de precisão.

Considere $y_1,\ldots,y_n$ variáveis aleatórias independentes, em que cada y_t, $t = 1,\ldots,n$, segue a densidade em (3.1) com média μ_t e parâmetro de precisão ϕ desconhecidos. Assim, temos que a resposta média y_t está relacionada a um preditor linear por meio de uma função de ligação da seguinte maneira

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

em que $\beta = (\beta_1,\ldots,\beta_k)^\top$ é o vetor de parâmetros desconhecidos a ser estimado ($\beta \in \mathbb{R}^k$), $x_{t1},\ldots,x_{tk}$ são observações de k variáveis independentes 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. Dessa forma, a resposta média é obtida aplicando a inversão da função de ligação, $\mu_t = g^{-1}(\eta_t)$, 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, em certas situações esse parâmetro pode variar ao longo das observações como apresentado em Almeida Junior e Souza (2015), Cribari-Neto e Souza (2012), Cribari-Neto e Souza (2013), Espinheira, Ferrari e Cribari-Neto (2008a), Espinheira, Ferrari e Cribari-Neto (2008b), Souza et

al. (2016), Smithson e Verkuilen (2006), Simas, Barreto-Souza e Rocha (2010), Silva e Souza (2014), Souza e Cribari-Neto (2015). Ou seja, o parâmetro de precisão é variável e precisar ser modelado a partir de uma estrutura de regressão similar a da resposta média. Dessa forma, temos que a precisão está relacionada ao preditor linear por meio de uma função de ligação, definida da seguinte maneira

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

em que $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1}, \ldots, z_{tq}$ são observações de q variáveis independentes $(k+q<n)$, 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, $\phi_t = h^{-1}(\vartheta_t)$ e $\mathrm{Var}(\mu_t) = \mu_t(1-\mu_t)/(1+\phi_t)$. Aqui, a variância depende da resposta média, sendo assim um modelo de natureza heteroscedástica.

Existem algumas escolhas possíveis para as funções de ligação $g(\cdot)$ e $h(\cdot)$. Por exemplo, para $g(\cdot)$, referente ao modelo da média, pode-se utilizar a função de ligação logit, $g(\mu) = \log\{\mu/(1-\mu)\}$, ou cloglog, $g(\mu) = \log\{-\log(1-\mu)\}$. Em relação ao modelo da precisão, pode-se utilizar a função $h(\phi) = \log(\phi)$ ou $h(\phi) = \sqrt{\phi}$ para $h(\cdot)$. Maiores detalhes sobre as funções de ligação ver Mccullagh e Nelder (1989).

Para realizar a estimação do vetor de parâmetros desconhecidos, $\beta = (\beta_1, \ldots, \beta_k)^T$, no modelo de regressão beta com dispersão variável utiliza-se o estimador de máxima verossimilhança. Segue de (3.1) que o logaritmo da função de verossimilhança utilizado no processo de estimação de β é

$$l(\beta, \gamma) = \sum_{t=1}^{n} l_t(\mu_t, \phi_t),$$

em que

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

Entretanto, 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, a exemplo, 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, Barreto-Souza e Rocha (2010).

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

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

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 MODELO DE REGRESSÃO QUANTÍLICA

A teoria clássica dos modelos lineares de regressão é fundamentalmente baseada na teoria das esperanças condicionais. Um texto de Mosteller e Tukey (1977), citado em Koenker e Bassett (1982) diz:

> O que a curva de regressão faz é dar um grande resumo das médias das distribuições correspondentes ao conjunto dos x's observados. Nós poderíamos ir mais longe e calcular diversas curvas de regressões diferentes correspondendo aos vários pontos percentuais da distribuição e assim obter uma visão mais completa desse conjunto. Assim como a média dá uma visão incompleta de uma única distribuição, a curva de regressão também dá uma visão incompleta correspondente para um conjunto de distribuições.[1]

Ou seja, de imediato é possível questionar se a esperança condicional ou outra medida de tendência central condicional é capaz de caracterizar adequadamente uma relação estatística entre a variável resposta e as demais covariáveis em estudo. Adicionalmente os modelos clássicos de regressão que utilizam o método dos mínimos quadrados ordinários podem gerar estimativas errôneas quando existem *outliers* nos dados ou quando a distribuição de probabilidade da variável resposta é bastante assimétrica. Isto ocorre por que tal método atribui pesos iguais a cada observação durante o processo de estimação dos parâmetros (DRAPER; SMITH, 1998). Dessa forma, é possível que a relação entre a variável resposta e as demais covariáveis independentes não seja estabelecida de forma correta, levando o pesquisador a interpretação errada dos resultados. Por esse motivo muitas vezes é necessário utilizar outros modelos de regressão mais robustos com objetivo de fornecer resultados mais confiáveis ou que contribuam com maiores informações nas análises. Para tanto, uma alternativa viável seria utilizar o modelo de regressão quantílica proposto por Koenker e Bassett (1978). Esta classe de modelos se trata de uma abordagem mais geral que os modelos clássicos e podem ser caracterizados como modelos de regressão semiparamétricos, pois não exigem nenhuma distribuição de probabilidade para a variável resposta.

Buchinsky (1998) sintetiza de forma simples algumas características importante da regressão quantílica. Primeiro, o modelo pode ser utilizado para caracterizar toda a distribuição condicional da variável resposta dado um vetor de covariáveis independentes. Segundo, os modelos de regressão quantílica tem uma representação de programação linear que torna as estimativas mais fáceis. Terceiro, o estimador soma dos desvios absolutos ponderados é responsável por dar um caráter mais robusto a regressão quantílica, de modo que a estimativa do vetor de coeficientes

[1] Traduzido do texto: "What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions".

não seja sensível a *outliers* na variável resposta. Quarto, quando os erros não são normais, o estimador dos desvios absolutos ponderados pode ser mais eficiente do que o estimador de mínimos quadrados ordinários. Por fim, soluções potencialmente diferentes nos distintos quantis podem ser obtidas por modificar as covariáveis nos diferentes pontos da distribuição condicional da variável resposta. Para maiores detalhes ver Koenker (2005).

Seja Y uma variável aleatória com uma função de distribuição acumulada $\mathrm{F}(y) = Pr(Y \leq y)$, então o quantil de ordem τ da variável Y pode ser definido como

$$Q_\tau(Y) = \mathrm{F}^{-1}(\tau) = \inf\{y|\mathrm{F}(y) \geq \tau\},$$

em que $\tau \in [0,1]$. Por exemplo, para $\tau = 0.5$ tem-se um importante caso em que é estimado a mediana condicional (KOENKER; BASSETT, 1982).

Considere (y_t, x_t), $t = 1, \ldots, n$, uma amostra obtida de uma população qualquer. Dessa forma, é possível relacionar a variável resposta, y, e o vetor de covariáveis, x, por meio da seguinte relação linear

$$Q_\tau(y_t|x_t) = \beta_{0(\tau)} + \beta_{1(\tau)}x_{t1} + \cdots + \beta_{p(\tau)}x_{tp},$$

em que $Q_\tau(y_t|x_t)$ é o quantil condicional de $y_t|x_t$ e $\beta_{0(\tau)}, \ldots, \beta_{p(\tau)}$ são os parâmetros desconhecidos indexados no quantil τ. O quantil de ordem τ de uma amostra ou população é um ponto tal que $100\tau\%$ dos valores amostrais ou populacionais são menores do que ele, com $\tau \in (0,1)$.

Então como apresentado em Koenker e Machado (1999), é possível estimar os parâmetros desconhecidos, indexados no quantil τ, diretamente por meio da solução do problema de minimização definido a seguir

$$\min_{(\beta)\in\mathbb{R}^p} \sum_{t=1}^{n} \rho_{(\tau)}(y_t - (\beta_{0(\tau)} + \beta_{1(\tau)}x_{t1} + \cdots + \beta_{p(\tau)}x_{tp})),$$

em que $\rho_{(\tau)}$ é um termo ponderador. Assim, o τ-ésimo estimador $\hat{\beta}_{(\tau)}$ da regressão quantílica é escolhido de modo a minimizar a soma dos desvios do y observado até um valor ajustado ($\hat{y}$) (NEYMAN; PEARSON, 1928), podendo ser reescrito da seguinte maneira

$$\sum_{t=1}^{n} d(y_t, \hat{y}_t) = \tau(\sum_{y_t \geq \hat{y}_t} |y_t - x_t^T\hat{\beta}_\tau|) + (1-\tau)(\sum_{y_t < \hat{y}_t} |y_t - x_t^T\hat{\beta}_\tau|),$$

em que $d(y_t, \hat{y}_t)$ é a distância a ser minimizada. Aqui, são utilizados os pesos τ se $y \geq \hat{y}$ (apresentando resíduo positivo) e o peso $1-\tau$ caso contrário. Ou seja, o método de estimação da regressão quantílica busca atribuir pesos diferentes a um determinado conjunto de observações dependendo de sua localidade na reta de regressão estimada. Portanto, a estimação dos parâmetros para cada regressão quantílica é baseada nos dados ponderados de toda a amostra em vez de considerar uma porção amostral daquele quantil (NEYMAN; PEARSON, 1928).

Koenker e Machado (1999) propuseram uma medida similar a estatística R^2, bastante utilizada na regressão normal linear, que permite ao investigador avaliar a qualidade do ajuste nos

modelos de regressão quantílica. Para tanto, considere o modelo linear para o quantil condicional de $y|x$

$$Q_\tau(y_t|x_t) = x_t^T \beta_{(\tau)} = x_{t1}^T \beta_{1(\tau)} + x_{t2}^T \beta_{2(\tau)},$$

em que $\widehat{\beta}$ é responsável por minimizar a soma dos desvios absolutos ponderados para o modelo completo

$$\widehat{V}(\tau) = \sum_{t=1}^{n} \rho_\tau(y_t - x_t^T \widehat{\beta}_{(\tau)})$$

e $\widetilde{\beta}$, sob a restrição linear que $\mathcal{H}_0 : \beta_2 = 0$, é responsável por minimizar a soma dos desvios absolutos ponderados para o modelo reduzido

$$\widetilde{V}(\tau) = \sum_{t=1}^{n} \rho_\tau(y_t - x_{t1}^T \widetilde{\beta}_{1(\tau)}).$$

Dessa forma, podemos definir a medida de bondade de ajuste para modelos de regressão quantílica da seguinte maneira

$$R^1(\tau) = 1 - \frac{\widehat{V}(\tau)}{\widetilde{V}(\tau)},$$

como $\widehat{V}(\tau) < \widetilde{V}(\tau)$, temos que $R^1(\tau)$ apresenta valores restritos ao intervalo $(0,1)$, constituindo assim uma medida da qualidade do ajuste para um particular quantil da regressão quantílica.

Portanto o modelo de regressão quantílica permite ao investigador ter uma visão mais completa da relação existente entre a variável resposta e as covariáveis observadas, uma vez que é possível construir um modelo para cada quantil de interesse, possibilitando identificar diferenças existentes entre os coeficientes estimados (NASCIMENTO et al., 2012). Contudo, quando as superfícies das linhas de regressão não são paralelas. Ou seja, quando o efeito das covariáveis não são uniformes ao longo dos quantis temos que os erros podem apresentar alguma forma de heteroscedasticidade (KOENKER, 2005).

Os modelos lineares de heteroscedasticidade apresentam-se como um importante caso na regressão quantílica. Neste cenário, por exemplo, pode-se assumir que a média e a variância são funções lineares das covariáveis. Segundo Koenker e Bassett (1982) partindo do pressuposto de erros independentes e identicamente distribuídos, pode-se definir um modelo geral de heteroscedasticidade sistemática dado por

$$y = \mu(x) + \sigma(x)\varepsilon,$$

em que $\mu(x)$ pode ser pensado como a média condicional do processo de regressão, $\sigma(x)$ como a escala condicional e ε como o termo de erros independente de x provenientes de uma distribuição com função quantílica $Q_\tau(\varepsilon)$.

A literatura disponibiliza algumas maneiras para estimar o erro padrão em modelos de regressão quantílica com a finalidade de construir intervalos de confiança ou realizar testes de hipóteses. Por exemplo, é possível supor que os erros são independentes e identicamente distribuídos (i.i.d.) para obter a estimativa assintótica da matriz de covariâncias definida em Koenker e Bassett (1978). Entretanto ao considerar os erros não identicamente distribuídos (n.i.d.), caracterizando um cenário heteroscedástico, computamos a estimativa assintótica da matriz de covariâncias por meio do estimador *Hubber–Sandwich* (HUBER, 1967; KOENKER, 2005).

Para motivar o uso da regressão quantílica foi realizado uma simulação considerando um cenário em que a variância dos erros não é constante. Ou seja, ela é definida em função da covariável, x, permitindo assim que a variância aumente em conjunto com x, violando assim a suposição de homoscedasticidade dos modelos de regressão. Para tanto, ajustou-se o modelo de regressão

$$Q_\tau(y_t|x_t) = \beta_0 + \beta_{1(\tau)}x_{t1}, \quad t = 1,\ldots,100,$$

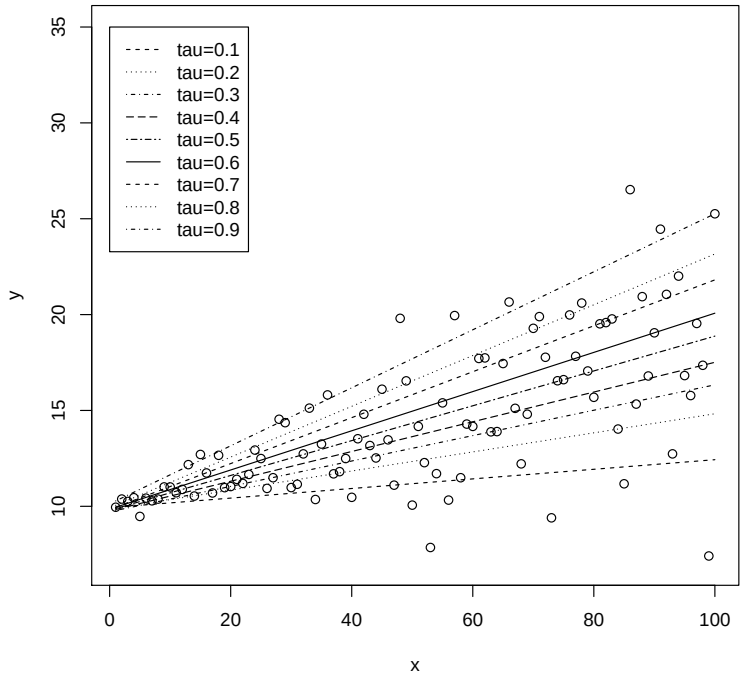
em que x é uma sequência de 1 até 100 com espaçamento 1 e parâmetros $\beta_{0(\tau)}$ e $\beta_{1(\tau)}$ iguais a 10 e 0.09, respectivamente. A Figura 1 apresenta os resultados obtidos após o ajuste do modelo para τ variando de 0.1 até 0.9 com espaçamento 0.1. A partir da análise gráfica podemos observar que as curvas de regressão não são paralelas, ou seja, os coeficientes estão variando em função de τ. Portanto, o efeito da covariável x não é uniforme através dos quantis. Dessa forma, se fosse utilizado os modelos de regressão normal linear não seria possível detectar os diferentes efeitos da covariável x ao longo da distribuição condicional da variável resposta.

Conforme verificado na literatura os modelos clássicos de regressão são bastante uteis quando os dados são normais e não ocorre violação do pressuposto de homoscedasticidade. Isto ocorre por que os testes de hipóteses utilizados para realizar inferência a respeito dos parâmetros desconhecidos estão fortemente relacionados a suposição de normalidade dos dados. Contudo, em muitas situações envolvendo conjunto de dados reais não é possível validar tais pressupostos. Dessa forma, muitos pesquisadores buscam aplicar transformações na variável resposta de modo a adequar os dados ao modelo proposto por obter uma distribuição mais próxima da desejada.

A princípio uma transformação na variável resposta pode ser uma alternativa prática para se conseguir um bom ajuste nos modelos de regressão normal linear. Entretanto, tal classe de modelos não permite obter o valor esperado da variável original por aplicar uma transformação inversa no preditor linear. Por exemplo, considere o caso em que aplica-se a transformação logarítmica na variável resposta, que não segue distribuição normal, obtendo assim a variável $z = \log(y)$ com distribuição de probabilidade mais próxima da desejada. Portanto, é possível obter a esperança condicional de $z|x$ da seguinte maneira

$$\mathbb{E}(z_t|X_t) = \beta_0 + \beta_1 x_t + \varepsilon_t,$$

Figura 1 – Ajustes do modelo de regressão quantílica $Q_\tau(y_t|x_t) = \beta_0 + \beta_{1(\tau)}x_{t1}$, $t = 1,\ldots,100$. Assumindo $\beta_0 = 0.1$, $\beta_1 = 0.9$ e $\tau = 0.1, 0.2, \ldots, 0.9$.



com $t = 1,\ldots,n$. Neste cenário, tem-se que

$$\mathbb{E}(z_t|X_t) = \mathbb{E}(\log(y)|X_t) \neq \log\mathbb{E}(y_t|X_t).$$

Ou seja, não é possível obter o valor esperado da variável original, y, por aplicar a exponencial em $\mathbb{E}(z_t|X_t)$. Por outro lado, a regressão quantílica apresenta uma característica distinta denominada de equivariância a transformações monótonas, como apresentado em Bottai, Cai e McKeown (2010), Hao e Naiman (2007), Mua e Hea (2007), Powell (1986), Santos (2012). Ou seja, caso torne-se necessário aplicar uma função $h(\cdot)$, a exemplo da exponencial ou da logarítmica, não decrescente no conjunto $\mathbb{R}$ na variável resposta y, tem-se que os quantis serão obtidos por aplicar a mesma transformação na função quantílica, $Q_\tau(y|x)$. Por exemplo, se q é o τ-ésimo quantil de y, então $h(q)$ é o τ-ésimo quantil de $h(y)$. Portanto, considere y uma variável reposta restrita ao intervalo $(0,1)$, com distribuição de probabilidade não gausiana, e x um vetor de covariáveis independentes. Dessa forma, é possível estabelecer uma relação funcional entre a variável y e o vetor de covariáveis, x, da seguinte maneira

$$h\{Q_\tau(y_t|x_t)\} = \beta_{0(\tau)} + \beta_{1(\tau)}x_{t1} + \cdots + \beta_{p(\tau)}x_{tp}, \quad t = 1,\ldots,n, \tag{3.2}$$

em que $Q_\tau(y_t|x_t)$ é o quantil condicional de $y|x$. Aqui, $h(\cdot)$ é uma função estritamente monótona e não decrescente no conjunto $\mathbb{R}$. Existem algumas escolhas possíveis para as funções de ligação $h(\cdot)$, a exemplo das ligações probit, loglog e logit (MCCULLAGH; NELDER, 1989). Portanto,

uma vez estimado os coeficientes $\hat{\beta}_{j(\tau)}$ é possível obter os valores de $Q_\tau(y_t|x_t)$ por aplicar uma transformação inversa $h^{-1}(\cdot)$ em 3.2. Por exemplo, pode-se utilizar a função $h(y) = \text{logit}(y)$ para transformar a variável resposta y, restrita ao intervalo $(0,1)$, da seguinte maneira

$$h(y) = \text{logit}(y) = \log\left(\frac{y}{1-y}\right),$$

Dessa forma, a inferência a respeito do τ-ésimo quantil da variável resposta original pode ser obtida por meio da seguinte transformação inversa

$$Q_\tau(y_t|x_t) = \frac{\exp(\beta_{0(\tau)} + \beta_{1(\tau)}x_{t1} + \cdots + \beta_{p(\tau)}x_{tp})}{1 + \exp(\beta_{0(\tau)} + \beta_{1(\tau)}x_{t1} + \cdots + \beta_{p(\tau)}x_{tp})},$$

com $t = 1, \ldots, n$. Assim é possível retornar aos valores da variável original sem perda de interpretação do modelo original.

4 REFERENCIAL METODOLÓGICO

4.1 TIPOLOGIA DA PESQUISA

O estudo é descritivo com abordagem quantitativa e inferencial com foco na análise de regressão. Na pesquisa em questão foram abordados dois métodos de regressão. O primeiro método se baseia na distribuição de probabilidade beta e permite a modelagem de variáveis respostas restritas ao intervalo (0,1), a exemplo da proporção de adultos obesos nas nações no ano de 2014. A segunda abordagem, denominada de regressão quantílica, permite ajustar diferentes modelos de regressão para cada quantil de interesse, contudo é necessário que a variável resposta esteja definida no conjunto dos reais. Desta forma, o estudo visa a comparação dos dois métodos supracitados para a obtenção da melhor abordagem que descreva a relação entre as variáveis explicativas e a variável resposta da melhor maneira possível.

4.2 CENÁRIO DO ESTUDO

O estudo realizado utiliza dados de domínio público referente a obesidade adulta em 78 países no ano de 2014, onde o cálculo da proporção observada considerou a população adulta, com 18 anos ou mais, que apresentasse *IMC* maior que $30kg/m^2$.

4.3 POPULAÇÃO E AMOSTRA

A amostra foi constituída por 78 observações (proporções) referentes às nações ao redor do mundo, das quais 25 (32%) pertencem a África, 11 (14%) pertencem a América, 14 (18%) pertencem a Ásia, 25 (32%) pertencem a Europa e 3 (4%) pertencem a Oceania.

4.4 PROCEDIMENTOS DE OBTENÇÃO DE DADOS

Os dados são oriundos de fontes de informação pública, mais especificamente das bases disponíveis nos endereços eletrônicos:

- http://databank.worldbank.org: o *Word Bank* é composto por cinco instituições que buscam reduzir a pobreza e fornecer assistência técnica e financeira a países em desenvolvimento.
- http://apps.who.int: tal organização funciona em mais de 150 países e trabalha lado a lado com os governos e outros parceiros para garantir o mais alto nível possível de saúde para todas as pessoas.

4.5 PROCEDIMENTOS DE ANÁLISE DOS DADOS

Os dados foram tabulados em uma planilha eletrônica, e após checagem foram transferidos para o *software* `R` (R Core Team, 2013) que é uma plataforma livre e possui diversos métodos estatísticos de análise de dados já implementados. Inicialmente, foi realizado uma análise descritiva dos dados a fim de extrair informações importantes a respeito das covariáveis abordadas no estudo. Adicionalmente, procedeu-se com a construção de mapas referentes as variáveis em estudo com objetivo de visualizar os níveis destas covariáveis nos diferentes países.

Em seguida, são apresentados os procedimentos inferenciais, as medidas da qualidade do ajuste e análise de diagnóstico relacionados aos modelos de regressão beta e quantílica. Vale ressaltar que para tais modelos utilizou-se respectivamente os pacotes `betareg` (CRIBARI-NETO; ZEILEIS, 2010) e `quantreg` (KOENKER, 2005) disponíveis no *software* `R`. Por fim, são comparadas as técnicas abordadas nesse estudo para determinar a equação que melhor relacione a variável resposta com as demais variáveis independentes.

5 RESULTADOS

5.1 DESCRIÇÃO DOS DADOS

A Tabela 1 apresenta uma breve descrição a respeito das variáveis abordadas nesta dissertação. As fontes de informação para obtenção dos dados foram as páginas da web: http://databank.worldbank.org e http://apps.who.int. Adicionalmente na Tabela 2 estão apresentados os países considerados neste estudo com suas respectivas codificações.

Tabela 1 – Descrição das variáveis.

Variáveis	Definição
$OB2014$	Proporção de adultos obesos, 18 anos ou mais, com $IMC \geq 30kg/m^2$ em 2014.
$INAT$	Porcentagem de atividade física insuficiente entre os adultos em 2010.
$EDUC$	Gastos com a educação como porcentagem da despesa total do governo em 2010.
$VIDA$	Expectativa de vida ao nascer em anos no ano de 2014.
ALC	Média do consumo em litros de álcool puro por pessoa em um ano, considerando a população com 15 anos ou mais em 2008.
URB	Porcentagem da população que vivem em áreas urbanas em 2014.

A Tabela 3 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 na modelagem da regressão beta e quantílica. Algumas conclusões podem ser extraídas da Tabela 3. Por exemplo, a proporção de adultos obesos varia de 0.03 até 0.41, com cerca de 25% dessas nações apresentando valores de $OB2014$ superior a 0.26 ou 26%. Em 50% das nações a prevalência das pessoas que praticam atividade física insuficiente é superior a 23.80%, com mínimo de 4.10% e máximo de 63.60%. A menor expectativa de vida ao nascer foi 49 anos e a maior 83 anos, com uma esperança de viver em média 72 anos. Os gastos com a educação como porcentagem da despesa total pelo governo variou de 5.53% até 26.30%, verificando-se ainda que 25% dessas nações apresentam valores de $EDUC$ menores que 11.25%. Considerando a porcentagem da população que vivem em áreas urbanas, temos que 50% dessas nações apresentam valores inferiores a 60.00%, com mínimo de 16.10% e máximo de 100.00%. Além disso, cerca de 25% desses países possuem valores de URB superiores a 74.82%. O consumo médio de álcool por pessoa em litros apresentou valor mínimo de 0.10 e máximo de 15.40, com média de 7.39. O coeficiente de variação é definido como a razão entre o desvio padrão e a média, sendo classificado como uma medida de dispersão. A partir dele é possível verificar que a variável ALC apresenta a maior variabilidade de dados em relação a média, com CV igual a 0.597. Vale ressaltar que um coeficiente de variação igual a zero nos diz que os dados de uma determinada variável são homogêneos. Ou seja, todas as observações equivalem a média.

Tabela 2 – Identificação das nações.

Continente	País	Código	Continente	País	Código
África	Nigéria	NGR	Ásia	Nepal	NEP
África	Mauritânia	MTN	Ásia	Butão	BHU
África	Senegal	SEM	Ásia	Catar	QAT
África	Mali	MLI	Ásia	Camboja	CAM
África	Burkina Faso	BUR	Ásia	Vietnã	VIE
África	Etiópia	ETH	Ásia	Tailândia	THA
África	Chade	CHA	Ásia	Malásia	MAL
África	Serra Leoa	SLE	Ásia	Índia	IND
África	Guiné-Bissau	GBS	Ásia	Sri Lanka	SRI
África	Gana	GHA	Ásia	Singapura	SIN
África	Benim	BEM	Ásia	Indonésia	IHO
África	R. C. Africana	CAF	Europa	Finlândia	FIN
África	Togo	TOG	Europa	Letônia	LAT
África	Camarões	CMR	Europa	Dinamarca	DEN
África	Quênia	KEN	Europa	Lituânia	LTU
África	Ruanda	RWA	Europa	Alemanha	GER
África	Malawi	MAW	Europa	Países Baixos	NED
África	Zimbabwe	ZIM	Europa	Polónia	POL
África	Namíbia	NAM	Europa	Bélgica	BEL
África	Suazilândia	SWZ	Europa	República Checa	CZE
África	África do Sul	SAF	Europa	Áustria	AUT
África	Tunísia	TUN	Europa	Itália	ITA
África	Cabo Verde	CPV	Europa	Eslovênia	SLO
África	São Tomé e Príncipe	STP	Europa	Hungria	HUN
África	Maurícia	MRI	Europa	França	FRA
América	Canadá	CAN	Europa	Romênia	ROM
América	Guatemala	GUA	Europa	Espanha	ESP
América	Brasil	BRA	Europa	Bulgária	BUL
América	Equador	ECU	Europa	Portugal	POR
América	Paraguai	PAR	Europa	Suécia	SWE
América	Estados Unidos	USA	Europa	Noruega	NOR
América	México	MEX	Europa	Estónia	EST
América	Colômbia	COL	Europa	Irlanda	IRL
América	Jamaica	JAM	Europa	Croácia	CRO
América	Barbados	BAR	Europa	Chipre	CYP
América	Chile	CHI	Europa	Sérvia	SRB
Ásia	Mongólia	MGL	Oceania	Ilhas Salomão	SOL
Ásia	Paquistão	PAK	Oceania	Austrália	AUS
Ásia	Líbano	LIB	Oceania	Nova Zelândia	NZL

Rep. C. Africana: República Central Africana.

O país Colômbia, localizado no continente da América do Sul, apresentou a maior proporção de pessoas que praticam atividade física insuficiente. Outros países estão bem próximo

Tabela 3 – Estatística Descritiva das variáveis.

Variáveis	Mínimo	$Q_{1/4}$	Mediana	Média	$Q_{3/4}$	Máximo	CV
*OB*2014	0.03	0.07	0.20	0.17	0.26	0.41	0.568
INAT	4.10	18.40	23.80	24.68	30.65	63.60	0.431
VIDA	48.93	65.06	74.41	71.73	79.94	83.08	0.128
EDUC	5.53	11.25	14.36	14.66	17.50	26.30	0.316
URB	16.10	39.22	60.00	57.35	74.82	100.00	0.401
ALC	0.10	3.92	7.15	7.39	11.25	15.40	0.597

dessa proporção, como Malásia, África do Sul e Mauritânia, sendo o primeiro localizado na Ásia e os dois últimos na África. Os maiores valores para a expectativa de vida foram observados na Espanha e Itália, localizados na Europa, seguidos por Cingapura na Ásia. O continente Europeu se destacou por apresentar o maior consumo de álcool por pessoa. Em ordem decrescente de seus valores temos, Lituânia, Romênia e Hungria. Os países Cingapura, Catar e Bélgica apresentaram as maiores porcentagem de pessoas vivendo em áreas urbanas. Vale ressaltar que os dois primeiro estão localizados na Ásia e o último na Europa. O continente da África se destacou por apresentar os maiores gastos com a educação como porcentagem de despesa total pelo governo, a saber, os países Etiópia, Namíbia e Benin. Por fim, a maior proporção de adultos obesos foi observada em Catar, localizado na Ásia, seguido por Estados Unidos, pertencente a América, enquanto que os menores valores foram observados em Camboja e Nepal, localizados no continente Asiático.

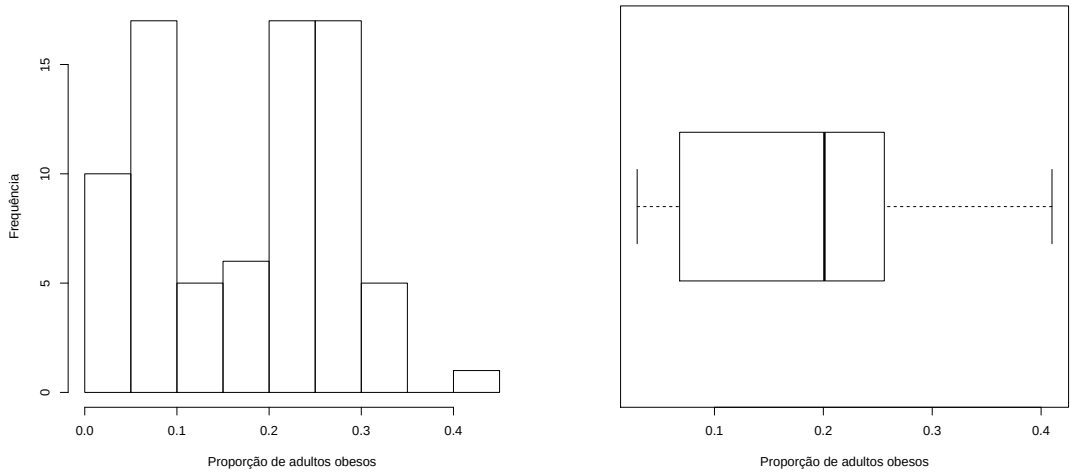
De acordo com a Tabela 4, pode-se observar que *OB*2014 se correlaciona positivamente com a maioria das covariáveis, exceto *EDUC*. Além disso, as maiores correlações lineares com a variável resposta são verificadas para *URB* e *VIDA*. Apesar de haver uma correlação de 0.70 entre elas não ocorreram problemas relacionados a multicolinearidade na análise de regressão mais adiante.

Tabela 4 – Correlação linear entre as variáveis *OB*2014, *INAT*, *VIDA*, *ALC*, *URB* e *EDUC*.

	*OB*2014	*INAT*	*VIDA*	*ALC*	*URB*	*EDUC*
*OB*2014	1.00	0.42	0.68	0.57	0.69	−0.29
INAT	-	1.00	0.23	0.05	0.38	−0.03
VIDA	-	-	1.00	0.47	0.70	−0.29
ALC	-	-	-	1.00	0.45	−0.36
URB	-	-	-	-	1.00	−0.22
EDUC	-	-	-	-	-	1.00

A Figura 2 apresenta o histograma de frequências e o *box-plot* da variável proporção de adultos obesos em 2014. É possível observar que a distribuição da variável resposta é assimétrica, facilmente observada no *box-plot*, já que a mediana está mais próxima do terceiro quartil. Além disso é verificado ausência de *outliers*, ou seja, observações discrepantes que excedem os limites do *box-plot*. Tais limites são definidos a partir das quantidades $Q_{1/4} - 1.5 \times (Q_{3/4} - Q_{1/4})$ e $Q_{3/4} + 1.5 \times (Q_{3/4} - Q_{1/4})$, referente respectivamente aos limites inferiores e superiores.

Figura 2 – Histograma e *box-plot* da variável proporção de adultos obesos nas nações em 2014, respectivamente.



A Figura 3 apresenta os *box-plots* das variáveis abordadas nesta dissertação segundo os continentes África, América, Ásia, Europa e Oceania. Como resultado pode-se observar que a maior concentração de nações com menores valores de $OB2014$ esta no continente Africano e Asiático. Por outro lado, os continentes da América, Europa e Oceania apresentam os maiores valores. Vale ressaltar que não existe interseção entre os *box-plots* da Europa e Oceania com os da África e Ásia, significando uma possível diferença existente entre as proporções de adultos obesos nestes continentes. Observando o *box-plot* de $INAT$ é visto que a Ásia apresenta a maior amplitude de dados. A partir do *box-plot* da variável URB não é possível identificar grandes diferenças entres os valores nos diferentes continentes. O continente Africano apresentou os menores valores de expectativa de vida. Por fim, também não é possível determinar grandes diferenças entre os *box-plots* da variável $EDUC$ devido a presença de interseção entre eles.

As Figuras 4, 5, 6, 7, 8 e 9 referem-se respectivamente aos mapas com os valores de $OB2014$, $INAT$, URB, $EDUC$, $VIDA$ e ALC por nação. Aqui, a ideia é visualizar de forma mais geral a distribuição dos valores destas covariáveis por continente. Por exemplo, na Figura 4 as regiões amareladas referem-se as nações com menores proporções de adultos obesos, ou seja, $OB2014 < 0.1$. Por outro lado, as regiões alaranjadas e avermelhadas referem-se as nações com valores mais altos de $OB2014$. Dessa forma, é possível visualizar que as nações com menores prevalências de adultos obesos estão distribuídas no continente da África e Ásia, enquanto que as maiores estão na América, Europa e Oceania.

Figura 3 – *Box-plot* das variáveis *OB*2014, *INAT*, *URB*, *ALC*, *VIDA* e *EDUC* segundo os continentes da África, América, Ásia, Europa e Oceania.

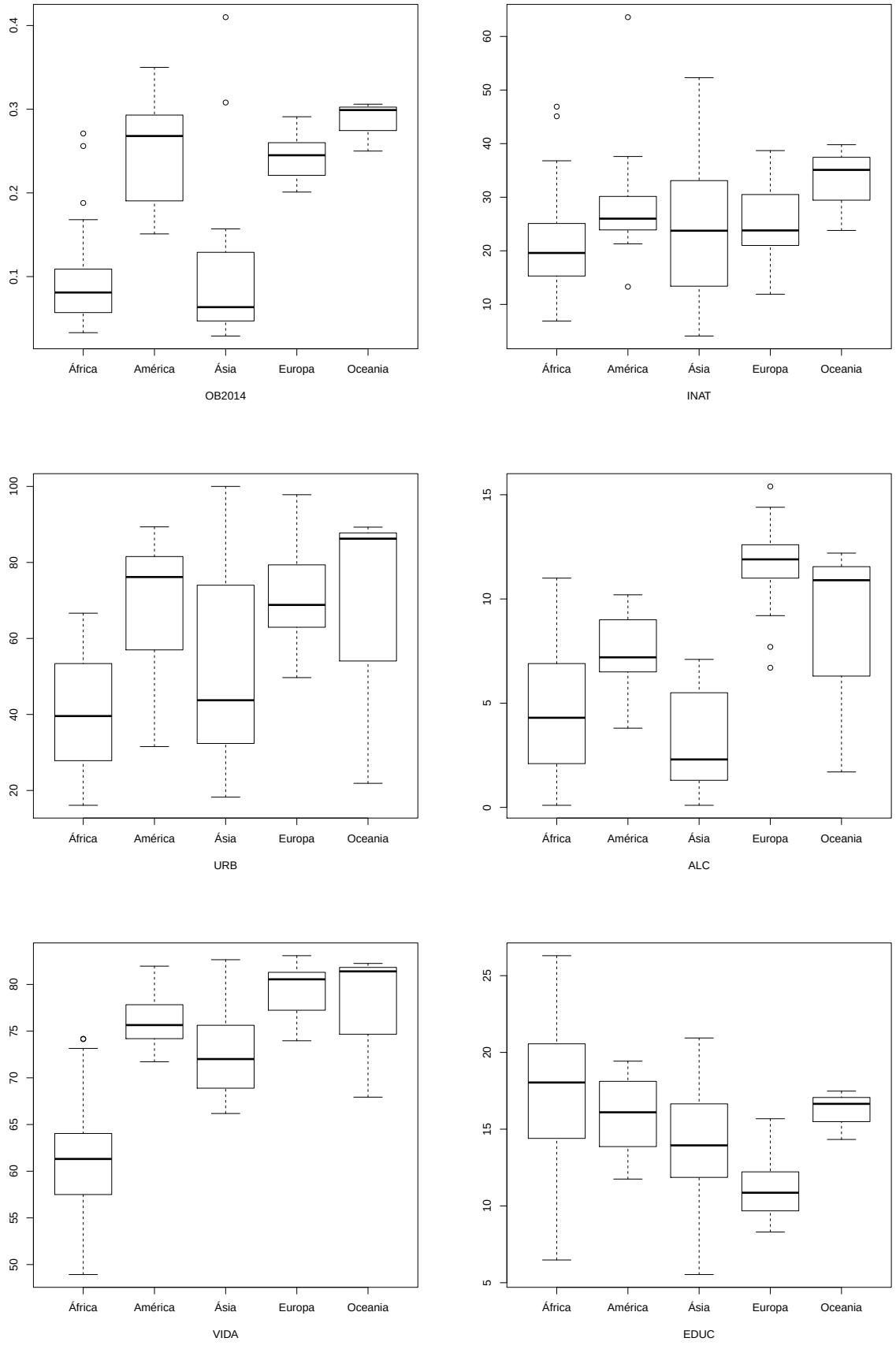


Figura 4 – Mapa da proporção de adultos obesos nas nações em 2014.

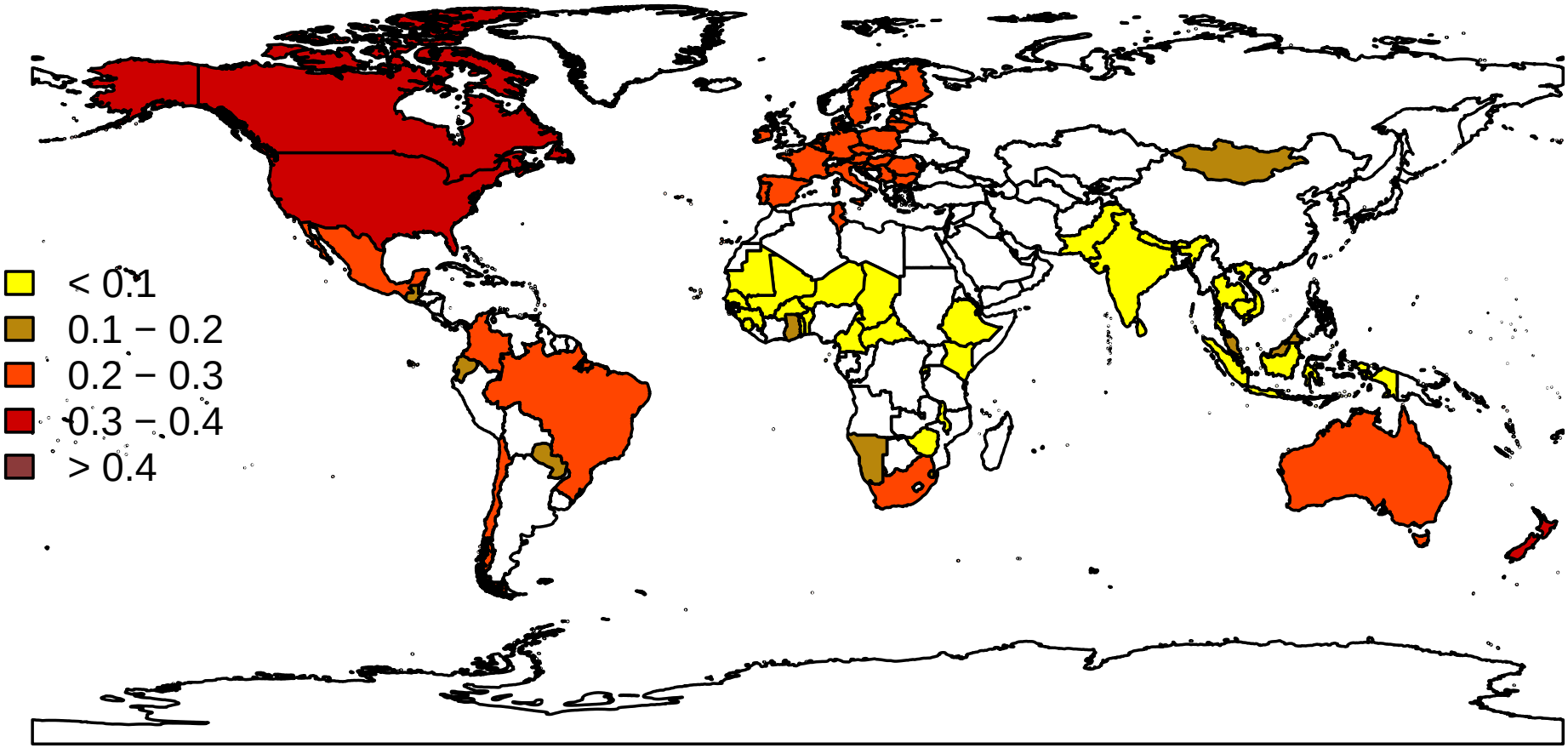


Figura 5 – Mapa da porcentagem de atividade física insuficiente entre os adultos nas nações em 2010.

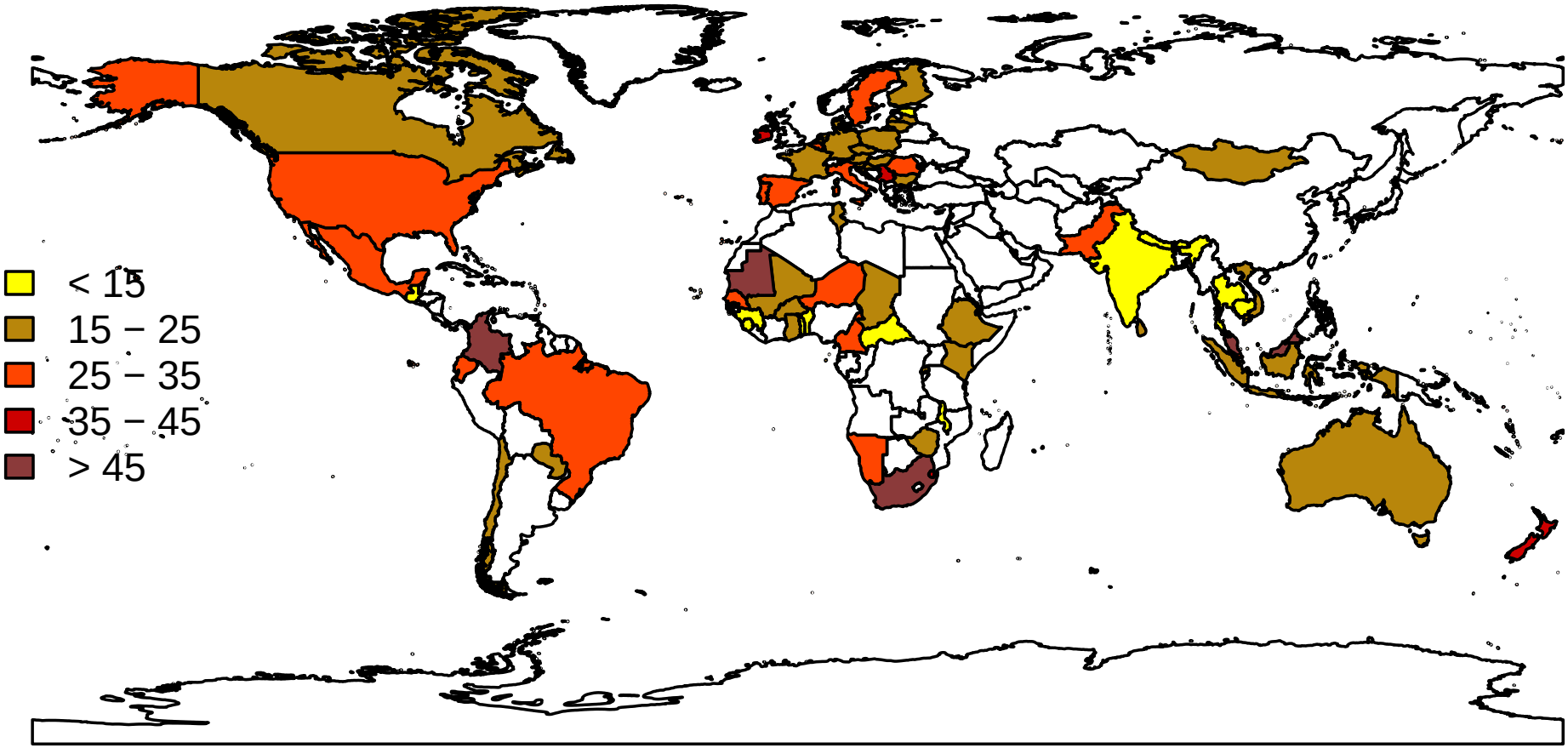


Figura 6 – Mapa da porcentagem da população que vivem em áreas urbanas nas nações em 2014.

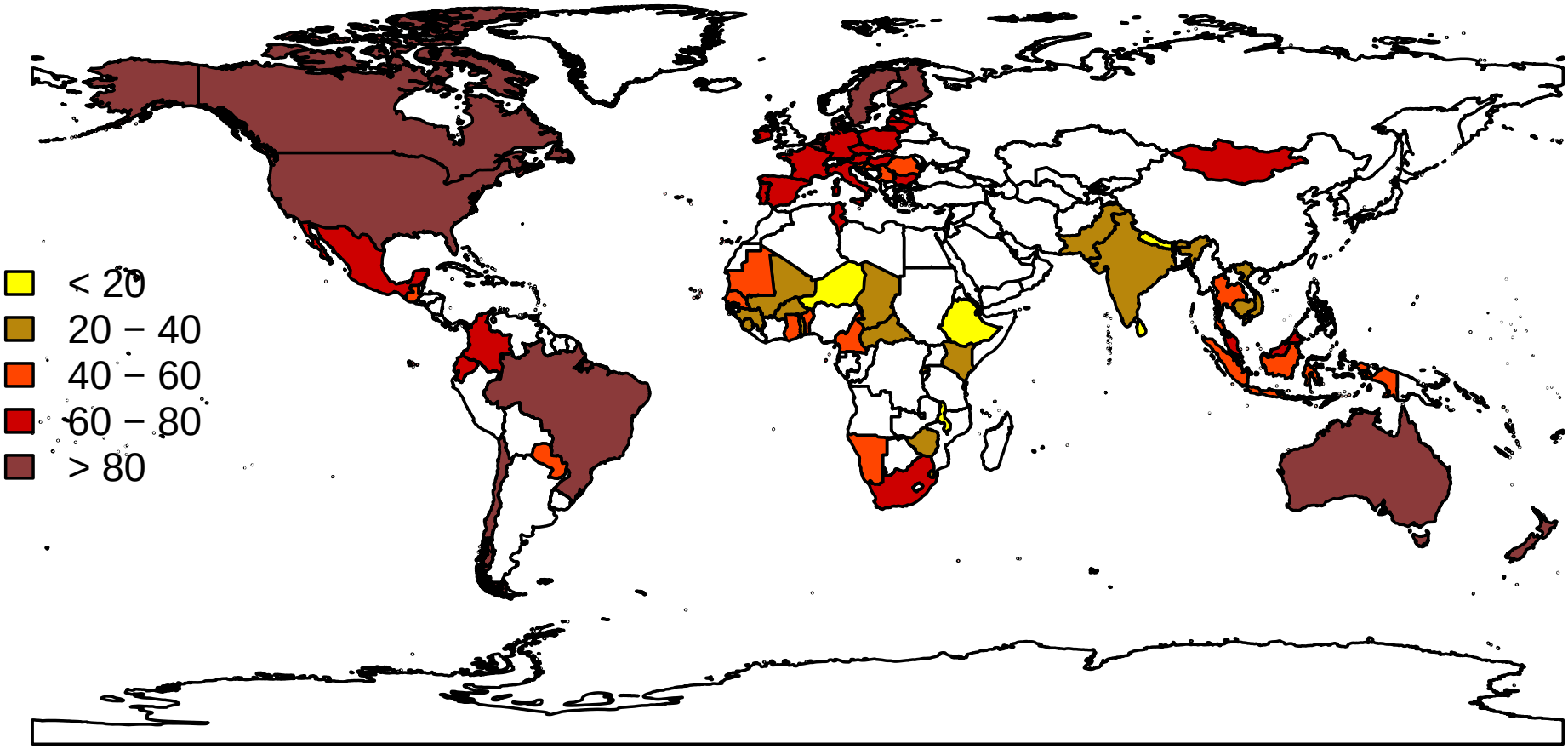


Figura 7 – Mapa dos gastos com a educação como porcentagem da despesa total do governo nas nações em 2010.

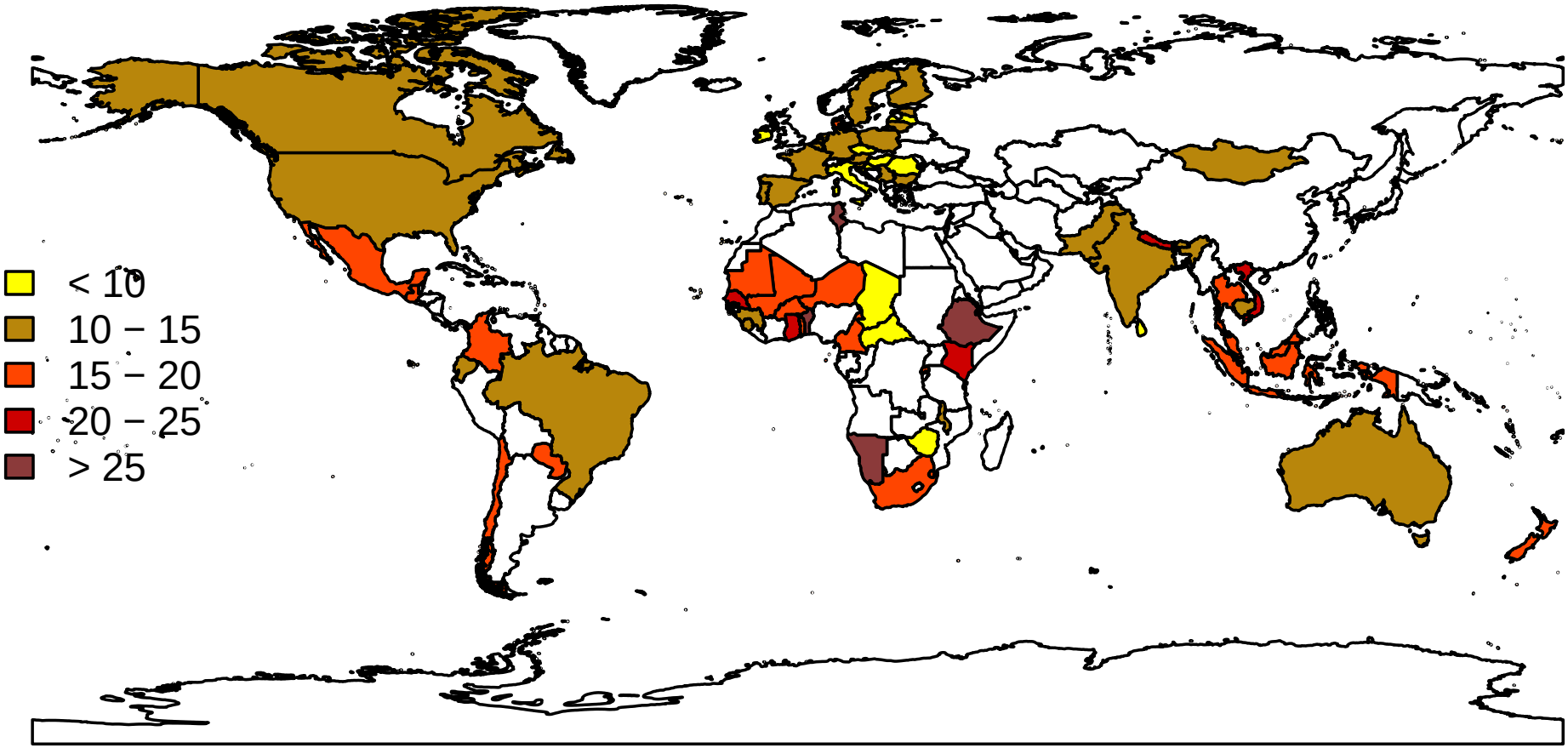


Figura 8 – Mapa da expectativa de vida em anos nas nações em 2014.

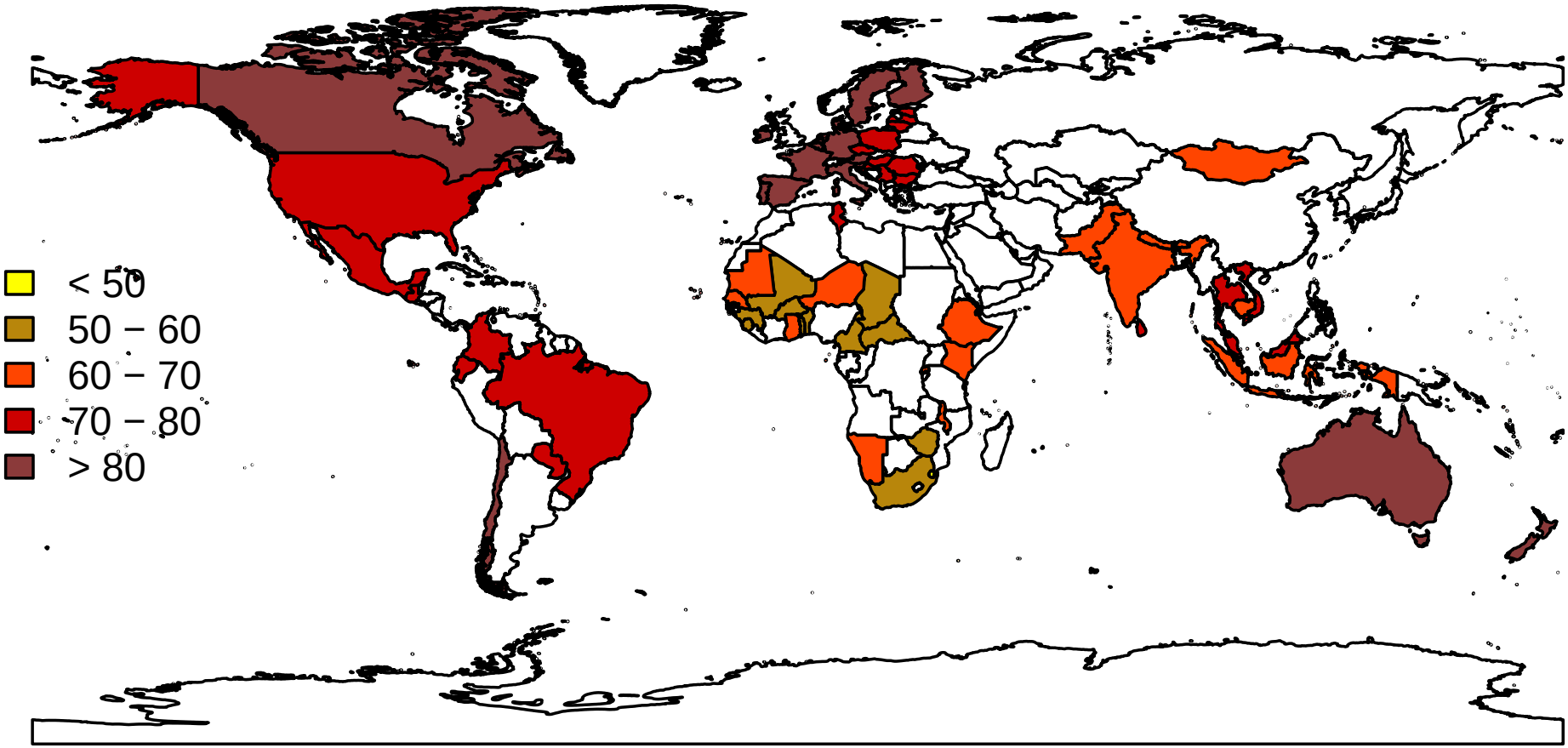
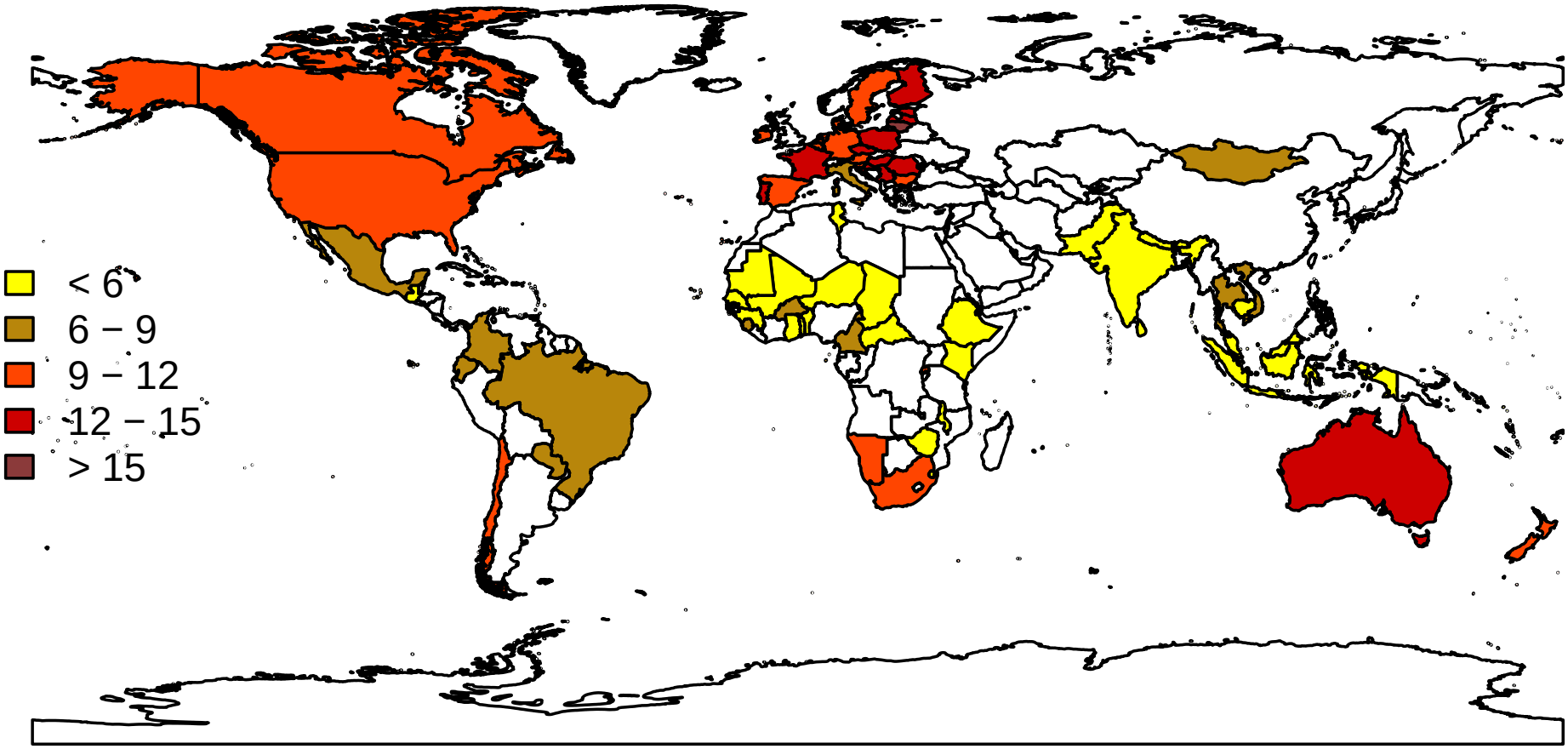


Figura 9 – Mapa do consumo médio em litros de álcool puro por pessoa em um ano, considerando a população com 15 anos ou mais em 2008.



5.2 ESPECIFICAÇÃO DO MODELO DE REGRESSÃO BETA

Nesta seção será apresentado os resultado obtidos a partir do modelo de regressão beta utilizado na modelagem de respostas com restrição no intervalo $(0,1)$. Para isso, será considerado o conjunto de dados referente a obesidade adulta nas nações que totalizam 78 observações.

Inicialmente, ao ajustarmos o modelo de regressão beta, é fundamental se questionar a respeito da dispersão dos dados. Como apresentado por Smithson e Verkuilen (2006), modelos de regressão com dispersão variável necessitam de uma estrutura para modelar a precisão dos parâmetros de modo a melhorar os resultados inferenciais. Para tanto, foi utilizado o teste da razão de verossimilhanças (ALMEIDA JUNIOR; SOUZA, 2015; NEYMAN; PEARSON, 1928; SILVA; SOUZA, 2014; SOUZA et al., 2016) com objetivo de verificar a hipótese nula de precisão fixa, isto é, $\mathcal{H}_0 : \phi_1 = \phi = \phi_n = \phi$. Como resultado, obteve-se um p-valor< 0.0001 (valor obtido a partir dos dados amostrais e que reflete a probabilidade de rejeitar a hipótese nula dado que ela é verdadeira). Ou seja, considerando o nível de significância de 5% rejeitamos a hipótese nula de precisão fixa. Portanto, é necessário uma estrutura de regressão para modelar a precisão dos dados.

O modelo de regressão beta com dispersão variável encontra-se apresentado a seguir

$$\begin{aligned} \text{loglog}(\mu_t) &= \beta_0 + \beta_1 INAT_t + \beta_2 URB_t + \beta_3 ALC_t + \beta_4 VIDA_t \\ \log(\phi_t) &= \gamma_0 + \gamma_1 VIDA_t + \gamma_2 EDUC_t + \gamma_3 ALC_t, \end{aligned}$$

com $t = 1, \ldots, 78$. Neste modelo temos que o parâmetro de precisão varia com as observações, havendo assim uma estrutura heteroscedástica. Dito isto, podemos definir uma medida do grau de heteroscedasticidade da seguinte maneira

$$\lambda = max(\phi_t)/min(\phi_t),$$

em que λ é a razão entre a máxima e a mínima precisão, obtendo-se um valor de $\lambda = 68.3$. Sob um cenário homoscedástico temos que o valor de λ é igual a 1. Entretanto mesmo que a dispersão dos dados seja fixa, a variância da variável resposta não é constante uma vez que seu valor depende das médias desconhecidas que variam com a estrutura de regressão.

A Tabela 5 apresenta as estimativas, os erros padrões e os p-valores utilizados para determinar a significância das estimativas do modelo proposto. Aqui, o modelo de regressão beta com dispersão variável utiliza as funções de ligação loglog e log para relacionar o preditor linear respectivamente a resposta média e a precisão. Como apresentado em Cribari-Neto e Zeileis (2010) é possível utilizar o teste de Wald (WALD, 1943) para verificar a hipótese nula de que $\beta_j = 0$ com $j = 1, \ldots, p$, ou seja, a variável associada ao parâmetro β_j não apresenta efeito significativo. Dessa forma, considerando o nível nominal de 5% temos que as variáveis atividade física insuficiente ($INAT$), pessoas vivendo em áreas urbanas (URB), consumo de álcool (ALC) e expectativa de vida ($VIDA$) são relevantes para explicar a proporção de adultos

obesos nas nações, uma vez que apresentaram p-valor< 0.05. Além disso, pode-se destacar que tais covariáveis apresentam efeito positivo no sentido de aumentar a proporção de adultos obesos nas nações. Ou seja, tal resultado se mostra coerente com os obtidos na análise descritiva por meio das correlações lineares com a variável resposta apresentadas na Tabela 4. O efeito positivo da variável $INAT$ pode ser justificado pela diminuição da perda de calorias ao longo do dia proporcionada pelas práticas de atividade física insuficiente. Por outro lado, o efeito positivo da variável URB pode estar ligado a dificuldade de se realizar refeições em casa devido ao crescente problema na rede de transporte urbano provocado pelo crescimento da urbanização. Dessa forma, a correria da vida moderna incentiva o consumo de refeições fora do domicílio, com destaque para o *fast-food* com suas ofertas de alimentos altamente energéticos (ANJOS, 2006). Além disso, a modernização e as mudanças no estilo de vida, devido ao avanço tecnológico, tornam as pessoas mais sedentárias e atribuem a elas maiores chances de se tornarem obesas. O efeito positivo da variável ALC pode ser entendido a partir da enorme quantidade de calorias ingeridas por meio do consumo do álcool, podendo contribuir para o aumento da obesidade nos países. O processo de envelhecimento das pessoas traz diversas mudanças no corpo como a diminuição do metabolismo e o ganho de peso (SOUZA; SCHROEDER; LIBERALI, 2007). Dessa forma, o efeito positivo da variável $VIDA$ pode estar relacionada ao processo de envelhecimento uma vez que quanto maior a expectativa de vida nas nações maior será a proporção de pessoas em idade mais avançada.

Por exemplo, para as nações com as covariáveis $INAT$, URB e ALC fixadas na mediana e apresentando uma expectativa de vida de 74 anos, de acordo com o modelo ajustado, estima-se a proporção média de adultos obesos como

$$\text{loglog}(\mu_t) = -2.009 + 0.009 \times 23.80 + 0.005 \times 60 + 0.027 \times 7.15 + 0.01 \times 74.41$$

Contudo, como a função de ligação utilizada foi a loglog a função inversa aplicada ao preditor linear a fim de obter o valor esperado da variável resposta é

$$\mu_t = \exp(-\exp(2.009 - 0.009 \times 23.80 - 0.005 \times 60 - 0.027 \times 7.15 - 0.01 \times 74.41))$$

$$\mu_t = 0.17$$

Ou seja, para as nações com 23.80% de atividade física insuficiente, 60% da população vivendo em áreas urbanas, consumo médio de álcool de 7.15 litros por pessoa e expectativa de vida de 74 anos é esperado uma proporção de adultos obesos em torno de 0.17 ou 17%.

Em relação a modelagem da precisão, Tabela 5, temos que as covariáveis expectativa de vida ($VIDA$), gastos com a educação pelo governo ($EDUC$) e consumo de álcool (ALC) foram estatisticamente relevantes ao nível de significância de 5%. Vale ressaltar que quanto maior for os valores de $VIDA$ e $EDUC$ nas nações menor será a precisão dos dados, consequentemente a dispersão aumenta. Por outro lado quanto maior for os valores de ALC maior será a precisão,

ou seja, o aumento da precisão significará uma menor dispersão dos dados tornando a resposta média mais precisa.

Tabela 5 – Estimativa dos coeficientes, erro padrão e p-valor do modelo de regressão beta com dispersão variável, considerando as funções de ligação loglog e log para modelar a média e a precisão, respectivamente.

Função de Ligação	Variáveis	Parâmetros	Estimativa	Erro padrão	p-valor
loglog(μ)	*INT*	β_0	-2.009	0.124	< 0.001
	INAT	β_1	0.009	0.002	< 0.001
	URB	β_2	0.005	0.001	< 0.001
	ALC	β_3	0.027	0.005	< 0.001
	VIDA	β_4	0.010	0.002	< 0.001
log(ϕ)	*INT*	γ_0	9.458	1.546	< 0.001
	VIDA	γ_1	-0.059	0.020	< 0.001
	EDUC	γ_2	-0.133	0.036	< 0.001
	ALC	γ_3	0.099	0.044	0.023

INT: refere-se ao intercepto do modelo.

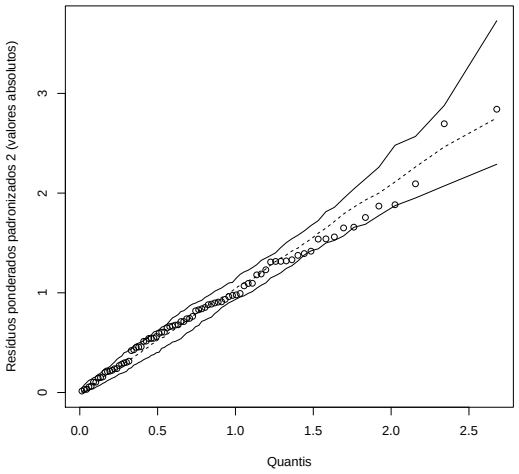
Para verificar-se a qualidade do ajuste do modelo utilizou-se o coeficiente de determinação ajustado (pseudo-R^2) e o teste *RESET*. O pseudo-R^2 é uma medida global da variação explicada e análogo ao coeficiente de determinação utilizado nos modelos de regressão linear. Essa medida é definida como o quadrado do coeficiente de correlação amostral entre $\hat{\eta}$ e $g(y)$ e foi proposta por Ferrari e Cribari-Neto (2004) para modelos de regressão beta. Dessa forma, com um pseudo-$R^2 = 0.69$, é dito que as covariáveis são capazes de explicar cerca de 69% da variabilidade total da proporção de adultos obesos nas nações. Além disso, temos que está medida apresenta valores restritos ao intervalo $(0,1)$, ou seja, quanto mais próximo de um melhor a qualidade do ajuste ou o poder explicativo do modelo. Por outro lado, para testar a correta especificação do modelo utilizou-se o teste *RESET* para modelos de regressão beta (LIMA, 2007; SOUZA et al., 2016; SOUZA; CRIBARI-NETO, 2015). O mecanismo do teste consiste em adicionar como covariável o preditor linear estimado elevado a segunda potência, $\hat{\eta}^2$, ao submodelo da média. A ideia por trás do teste é que se esta covariável têm algum poder em explicar a variável resposta, então rejeitamos a hipótese nula de ausência de erros de especificação. Ou seja, o modelo proposto apresenta forma funcional correta e não ocorrem omissões de variáveis (RAMSEY, 1969). Portanto com um p-valor$= 0.075$ não temos evidências suficientes para rejeitar a hipótese nula de que o modelo está bem especificado ao nível de significância de 5%.

A análise de diagnóstico se apresenta como uma importante etapa na construção dos modelos de regressão, pois é nela que o investigador analisa algumas medidas da qualidade do ajuste além das suposições feitas ao modelo, tais como adequação da distribuição de probabilidade suposta para a variável resposta, aleatoriedade dos resíduos e análise das medidas de

influência (SOUZA et al., 2016). Neste artigo utilizou-se os resíduos ponderados padronizados apresentados em Espinheira, Ferrari e Cribari-Neto (2008b) para construção e posterior análise dos gráficos.

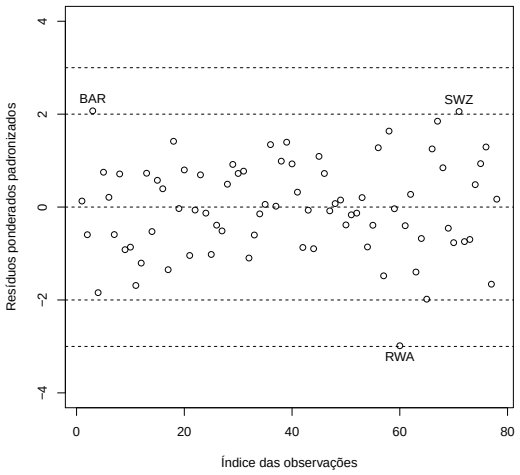
O gráfico de probabilidade normal com envelope simulado é uma técnica que permite ao investigador identificar desvios na suposição do modelo e possíveis observações discrepantes. Na Figura 10 verifica-se que as observações encontram-se distribuídas de forma aleatória dentro dos limites do envelope e próximo a linha central, apresentando uma quantidade reduzida de observações que excedem levemente esses limites. Portanto não temos evidências suficientes para discorda da adequação do modelo. Por outro lado, a Figura 11 apresenta o gráfico dos resíduos ponderados padronizados versus a ordem das observações. A partir dela é possível determinar se os resíduos apresentam algum tipo de tendência ou se estão distribuídos de forma aleatória em torno do zero. Dessa forma, é possível contabilizar um número reduzido de observações que ultrapassaram os limites estabelecidos $[-2, 2]$, utilizados para classificar os resíduos como demasiadamente grandes. As observações referem-se especificamente as nações: Barbados (BAR), Suazilândia (SWZ) e Ruanda (RWA).

Figura 10 – Gráfico da probabilidade normal com envelope simulado.



As medidas de influência auxiliam o investigador a identificar possíveis observações discrepantes, a exemplo, dos *outliers* que estão relacionados a variável resposta e dos pontos de alavanca que estão relacionados as variáveis independentes. Dessa forma é útil identificar e verificar o impacto que cada uma dessas observações podem ocasionar nas estimativas dos parâmetros, uma vez que elas tem o potencial para desviar a reta de regressão. 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, Ferrari e Cribari-Neto (2008a) 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, Hu e

Figura 11 – Resíduos ponderados padronizados versus os índices das observações.

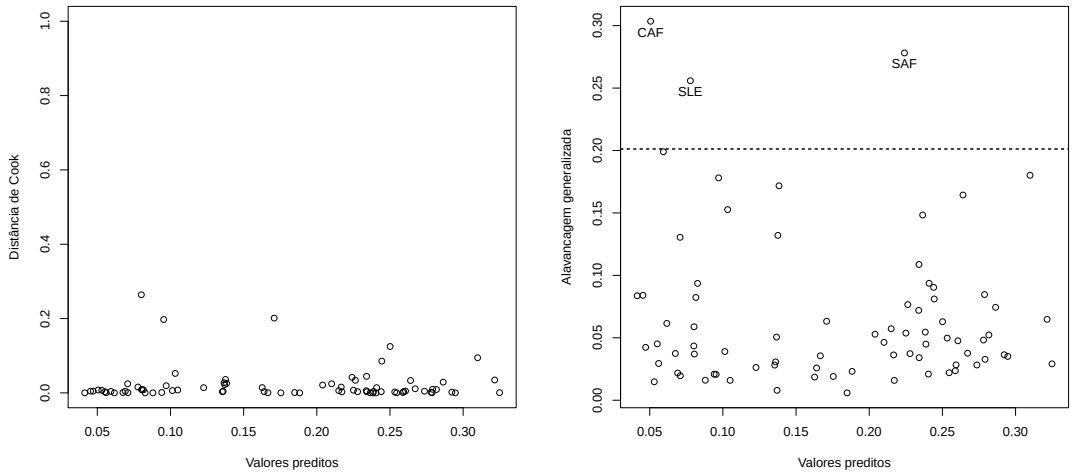


Fung (1998) é definida em modelos de regressão como uma medida da importância individual das observações. Ferrari, Espinheira e Cribari-Neto (2011) propuseram a alavancagem generalizada para modelos de regressão beta com dispersão variável.

A Figura 12 refere-se ao gráfico da distância de Cook versus os valores preditos e da alavancagem generalizada versus os valores preditos. Esses gráficos permitem identificar observações que possam interferir no processo de estimação dos parâmetros, produzindo resultados distorcidos ou estimativas errôneas. No gráfico da distância de Cook não foi possível identificar nenhuma observação como influente, ou seja, que ultrapassasse o limite estabelecido pelo quantil da distribuição F com p e $n-p$ graus de liberdade. Por outro lado, no gráfico da alavancagem generalizada foi possível identificar três observações com valores maiores que a quantidade $3p/n$, classificando as observações como pontos de alavanca, a saber: República Central Africana (CAF), Serra Leoa (SLE) e África do Sul (SAF).

Após identificar as observações de alta alavancagem é possível avaliar as variações percentuais nas estimativas dos parâmetros quando excluímos tais observações. Esta análise é importante pois permite ao investigador avaliar o potencial das observações influentes em desviar a reta de regressão estimada, ocasionando estimativas distorcidas dos parâmetros. Para tanto, foram construídos quatro possíveis cenários e seus resultados encontram-se disponíveis na Tabela 6. O país República Central Africana (observação 14) se destaca por apresentar a segunda menor expectativa de vida, aproximadamente 51 anos, e o segundo menor gasto com a educação em porcentagem da despesa total do governo, em torno de 6.48%. A exclusão dessa observação causa um maior impacto para as estimativas de $\hat{\beta}_1$ e $\hat{\gamma}_1$, reduzindo seus valores em aproximadamente 6.93% e 9.34%. O país Serra Leoa (observação 64) destaca-se por apresentar a terceira menor expectativa de vida, perdendo para Suazilândia e República Central Africana. Com a exclusão dessa observação as estimativas de $\hat{\beta}_4$ e $\hat{\gamma}_1$ sofrem as maiores variações

Figura 12 – Gráfico de distância de Cook e da alavancagem generalizada.



percentuais com uma redução de seus valores respectivamente em 8.16% e 5.95%. A África do Sul (observação 68) apresenta a terceira maior taxa de atividade física insuficiente, com cerca de 46.9%, além de uma das menores expectativas de vida, em torno de 57 anos. A exclusão dessa observação influencia mais as estimativas de $\hat{\beta}_1$, $\hat{\beta}_4$ e $\hat{\gamma}_1$, causando as respectivas variações percentuais 13.40%, 9.78% e 7.22%.

A exclusão das três observações simultaneamente causa uma variação considerável em $\hat{\beta}_1$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ e $\hat{\gamma}_3$, reduzindo seus valores em 23.38%, 11.60%, 18.78% e 12.66%, respectivamente. O λ representa o grau de heterogeneidade da precisão dos dados, sendo definido como $\max(\phi_t)/\min(\phi_t)$. Nas suas estimativas para os diferentes casos, ocorre redução de seus valores, refletindo assim a intensidade de não-constância da precisão.

Tabela 6 – Variações percentuais nas estimativas dos parâmetros ao se retirar observações influentes. Proporção de adultos obesos nas nações em 2014.

Cenários	14 (CAF)	64 (SLE)	68 (SAF)	14, 64, 68
β_0	-2.65	-2.81	1.26	-5.16
β_1	-6.93	-4.73	-13.40	-23.38
β_2	4.40	2.92	-3.30	5.61
β_3	-1.90	3.03	-4.49	-3.08
β_4	-6.01	-8.16	9.78	-8.00
γ_0	-6.14	-3.70	-4.52	-11.60
γ_1	-9.34	-5.95	-7.22	-18.78
γ_2	-6.74	-3.58	-3.51	-9.36
γ_3	5.21	3.01	6.74	12.66
λ	-38.26	-15.55	-16.18	-47.77

Como apresentado em Cribari-Neto e Souza (2013) é possível estimar o impacto de uma determinada covariável, a exemplo, da porcentagem de atividade física insuficiente sobre a

proporção de adultos obesos nas nações da seguinte maneira

$$\frac{\partial \mathbb{E}(y_t)}{\partial INAT_t} = \frac{\partial \mu_t}{\partial INAT_t}, \tag{5.1}$$

em que

$$\mu_t = g^{-1}(\beta_0 + \beta_1 INAT_t + \beta_2 URB_t + \beta_3 ALC_t + \beta_4 VIDA_t).$$

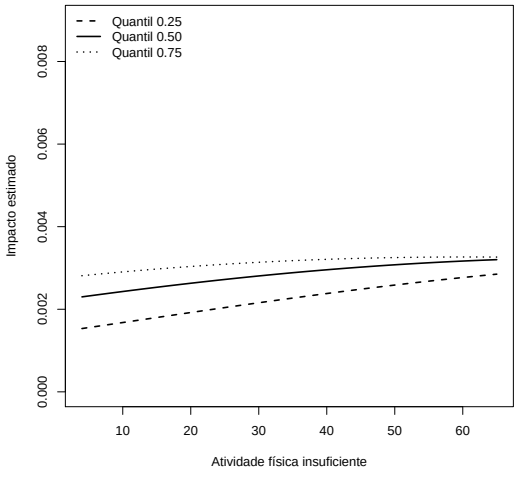
Como utilizou-se a função de ligação loglog para a média, temos que o impacto definido em 5.1 pode ser escrito como

$$\begin{aligned} \frac{\partial \mathbb{E}(y_t)}{\partial INAT_t} &= exp(-exp(-(\beta_0 + \beta_1 INAT_t + \beta_2 URB_t + \beta_3 ALC_t + \beta_4 VIDA_t))) \\ &\times (-exp(-(\beta_0 + \beta_1 INAT_t + \beta_2 URB_t + \beta_3 ALC_t + \beta_4 VIDA_t))) \times (-\beta_1), \end{aligned}$$

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

Com o objetivo de estimar as curvas de impacto para descrever o efeito da atividade física insuficiente sobre a proporção de adultos obesos nas nações considerou-se três situações, como apresentado na Figura 13. Ou seja, em que as covariáveis URB, ALC e $VIDA$ estão fixadas no primeiro, segundo e terceiro quartis. Dessa forma, é possível variar os valores de $INAT$ para determinar o aumento provocado na resposta média. Como resultado observa-se que o impacto é positivo e cresce lentamente quando aumenta-se os valores da atividade física insuficiente. Além disso, não existem grandes diferenças entre as curvas nos quantis 0.50 e 0.75 e elas diminuem a medida que aumenta-se os valores de $INAT$.

Figura 13 – Impacto da atividade física insuficiente sobre a proporção de obesos nas nações em 2014.



5.3 ESPECIFICAÇÃO DO MODELO DE REGRESSÃO QUANTÍLICA

Nesta seção é utilizado o modelo de regressão quantílica para explicar a relação existente entre a proporção de adultos obesos nas nações e as covariáveis exibidas na Tabela 1. O modelo de

regressão quantílica proposto por Koenker e Bassett (1978) é caracterizado como uma abordagem mais geral que os modelos clássicos, pois permite ajustar retas de regressão para diversos quantis de interesse e tratar dos problemas relacionados a assimetria nos dados (como a presença de *outliers* que podem prejudicar o processo de estimação dos parâmetros). Além disso, a regressão quantílica apresenta uma propriedade básica denominada de equivariância a transformações monótonas, como apresentada em Bottai, Cai e McKeown (2010), Hao e Naiman (2007), Santos (2012), permitindo assim a modelagem de resposta restritas ao intervalo $(0,1)$, por meio do uso de uma função monótona.

Como verificado na análise descritiva a variável resposta está restrita ao intervalo $(0,1)$, dessa forma torna-se necessário aplicar algum tipo de transformação. Na literatura existem várias possibilidades de transformação, contudo a logit é a mais utilizada. Assim, a variável transformada pode ser definida da seguinte maneira

$$h(y) = \text{logit}(OB2014) = \log\left(\frac{OB2014}{1-OB2014}\right),$$

em que $h(\cdot)$ é a função logit. Essa transformação permite que a variável resposta possa assumir valores no conjunto $\mathbb{R}$, possibilitando assim o uso da regressão quantílica. O uso dessa abordagem é denominada de regressão quantílica logística como apresentado em Bottai, Cai e McKeown (2010), Luca e Boccuzzo (2014), Feizi, Aliyari e Roohafza (2012). Esta propriedade da regressão quantílica não é aplicável a regressão das médias. Ou seja, não é possível ajustar um modelo de regressão normal linear com a variável $h(y)$ e esperar que a mesma interpretação seja obtida em y, após aplicar a função inversa h^{-1}. Os procedimentos computacionais foram realizados utilizando o pacote `quantreg` (KOENKER, 2005; KOENKER, 2008) do *software* estatístico `R` (R Core Team, 2013), que é uma plataforma livre e apresenta em suas bibliotecas diversos métodos estatísticos já implementados.

A seguir está apresentado o modelo de regressão quantílica utilizado para obter informações a respeito do τ-ésimo quantil condicional do logit da proporção de adultos obesos nas nações.

$$Q_\tau(\text{logit}(OB2014_t)|x_t) = \beta_{0(\tau)} + \beta_{1(\tau)}INAT_t + \beta_{2(\tau)}URB_t + \beta_{3(\tau)}ALC_t,$$

com $t = 1,\ldots,78$ e parâmetros indexados no quantil τ, com $\tau \in (0,1)$.

A Tabela 7 apresenta as estimativas dos coeficientes após ajustar o modelo de regressão quantílica para $\tau = 0.15$, $\tau = 0.25$, $\tau = 0.50$, $\tau = 0.75$, $\tau = 0.85$ e $\tau = 0.90$, além da regressão normal linear referenciada por MQO (Mínimos Quadrados Ordinários). Vale ressaltar que na literatura não existe um valor fixo de τ para ser avaliado. Dessa forma a escolha desses quantis busca analisar o efeito das covariáveis em diferentes pontos na cauda inferior, superior e no quantil mediano da distribuição condicional da variável resposta. Aqui, o p-valor é obtido assumindo que os resíduos são não identicamente distribuídos tornando as análises mais robustas. A partir dos resultados é possível observar que a variável porcentagem de indivíduos que praticam

atividade física insuficiente $(INAT)$ apresenta efeito positivo para todos os quantis avaliados. Ou seja, quanto menor for os valores de $INAT$ menor será a prevalência de adultos obesos nas nações (mantendo as demais covariáveis constantes). Isto ocorre por que as práticas de atividade física apresentam-se como uma fator de proteção para a obesidade, reduzindo o acúmulo de calorias ao longo do dia. Entretanto em alguns pontos da distribuição condicional da variável resposta está covariável não demonstrou ser relevante dado o p-valor> 0.05. As estimativas dos parâmetros relacionados a variável porcentagem de pessoas vivendo em áreas urbanas (URB) revela um efeito positivo sobre a resposta. Este resultado pode ser devido as mudanças no estilo de vida proporcionada pela crescente urbanização, uma vez que a correria da vida moderna incentiva o consumo de refeições fora do domicílio. Ou seja, as pessoas tornam-se mais propensas a ingerir alimentos mais calóricos, contribuindo para o acúmulo de gordura (ANJOS, 2006). Contudo esta covariável não demonstrou efeito significativo nos quantis 0.85 e 0.90 dado o p-valor> 0.05. A variável consumo médio de álcool por pessoa(ALC) também apresentou efeito positivo sobre a prevalência de obesos nas nações. Este resultado pode ser explicado a partir da ingestão de calorias provenientes do álcool que pode contribuir para o aumento da obesidade. Além disso, esta covariável se mostrou relevante para todos os quantis analisados e seu efeito parece diminuir em função dos quantis.

Analisando o ajuste proveniente da regressão normal linear, Tabela 7, é verificado que o efeito médio destas covariáveis é significativo para explicar a variabilidade da proporção de adultos obesos nos países. Além disso, comparando as estimativas de MQO com as estimativas no quantil mediano é possível observar algumas diferenças. Por exemplo, a estimativa de β_1 via MQO é quase duas vezes maior que a estimativa no quantil mediano para a mesma covariável. Contudo, a significância destas diferenças serão determinadas a partir de análise gráfica mais adiante.

Tabela 7 – Estimativa dos coeficientes e p-valor referente ao τ-ésimo quantil, $\tau = 0.15, 0.25, 0.50, 0.75, 0.85, 0.90$, e MQO. p-valor em parêntese e erro padrão assumido ser não identicamente distribuído (n.i.d).

		Quantis						
Variáveis	Parâmetros	$\tau = 0.15$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.85$	$\tau = 0.90$	MQO
INT	β_0	-4.375	-4.222	-3.680	-3.774	-2.727	-2.617	-3.816
		< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
$INAT$	β_1	0.013	0.014	0.010	0.022	0.028	0.037	0.021
		0.385	0.001	0.118	0.077	0.011	< 0.001	< 0.001
URB	β_2	0.023	0.019	0.017	0.022	0.011	0.006	0.017
		< 0.001	< 0.001	< 0.001	< 0.001	0.062	0.133	< 0.001
ALC	β_3	0.076	0.096	0.091	0.058	0.033	0.043	0.075
		0.013	< 0.001	< 0.001	0.013	0.077	0.008	< 0.001

INT: refere-se ao intercepto do modelo.

Dessa forma, para países com 23.80% de atividade física insuficiente, 60% da população

vivendo em áreas urbanas e consumo médio de álcool equivalente a 7.15 litros por pessoa, de acordo com o modelo ajustado, estima-se que o quantil de ordem 0.50 do logit da proporção de adultos obesos é

$$Q_{0.50}(\text{logit}(OB2014)|x) = -3.680 + 0.010 \times 23.80 + 0.017 \times 60 + 0.091 \times 7.15$$

Contudo, pela propriedade de equivariância a transformações monótonas temos que

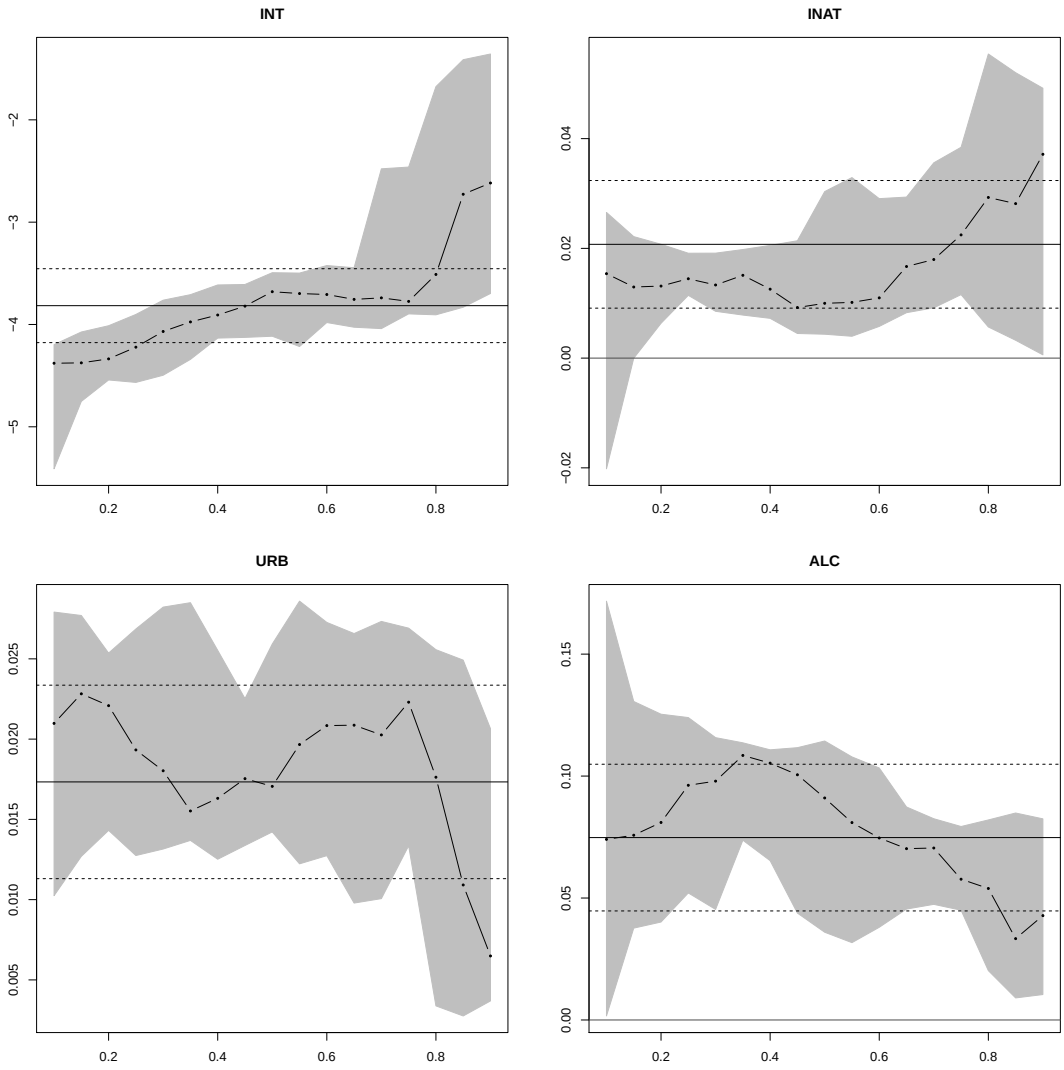
$$Q_{0.50}(OB2014/x) = \frac{\exp(-3.680 + 0.010 \times 23.80 + 0.017 \times 60 + 0.091 \times 7.15)}{1 + \exp(-3.680 + 0.010 \times 23.80 + 0.017 \times 60 + 0.091 \times 7.15)} \approx 0.15$$

Ou seja, para as nações com 23.80% de atividade física insuficiente, 60% da população vivendo em áreas urbanas e consumo médio de álcool equivalente a 7.15 litros por pessoa é esperado uma proporção de adultos obesos em torno de 0.15 ou 15%.

A Figura 14 permite visualizar as estimativas dos parâmetros da regressão quantílica e compará-las com as que seriam obtidas caso se ajustasse um modelo de regressão normal linear em função dos quantis. Ou seja, é possível visualizar de forma mais geral o comportamento das estimativas de cada parâmetro em diferentes quantis da distribuição condicional da variável resposta. Para comparação, as linhas pontilhadas são as estimativas e o intervalo com 95% de confiança do estimador de mínimos quadrados ordinários (MQO). Por outro lado, a área hachurada refere-se ao intervalo com 95% de confiança das estimativas da regressão quantílica. Dessa forma, para obter um gráfico mais detalhado definiu-se um conjunto de quantis mais amplo. Como resultado é verificado que as estimativas do intercepto, denotado por INT, crescem gradativamente em função dos quantis. Além disso, $\hat{\beta}_{0(\tau)}$ demonstra ser significativo para todos os quantis avaliados, uma vez que o valor zero não está contido no intervalo com 95% de confiança da regressão quantílica. As estimativas dos coeficientes relacionado a $INAT$ oscilam dentro do intervalo de confiança via MQO em boa parte dos quantis, especificamente nos quantis inferiores e mediano. É observado também que suas estimativas crescem lentamente para $\tau > 0.50$. As estimativas relacionadas a variável URB apresentam um comportamento oscilatório com um decrescimento acentuado a partir do quantil 0.75. Além disso, a maior parte de suas estimativas, especificamente para $\tau < 0.75$, está contida no intervalo via MQO. As estimativas relacionadas a variável ALC assumem um padrão decrescente para o quantis maiores que 0.40. É visto também que em diversos pontos as estimativas da regressão quantílica se mantém contidas dentro do intervalo de confiança de mínimos quadrados ordinários, não demonstrado haver diferença estatística entre as estimativas obtidas por meio das duas abordagens.

Após estimar os parâmetros do modelos um ponto importante é verificar se as diferenças observadas nos diferentes quantis é realmente significativa. Para isso, pode-se utilizar o teste de Wald (BUCHINSKY, 1998; KOENKER; BASSETT, 1982; NEYMAN; PEARSON, 1928) cuja hipótese nula é $\mathcal{H}_0 : \beta_{2(\tau)} = \beta_{2(\theta)}$ e $\beta_{3(\tau)} = \beta_{3(\theta)}$ e...e $\beta_{j(\tau)} = \beta_{j(\theta)}$ versus a alternativa $\mathcal{H}_1 : \beta_{2(\tau)} \neq \beta_{2(\theta)}$ ou $\beta_{3(\tau)} \neq \beta_{3(\theta)}$ ou...ou $\beta_{j(\tau)} \neq \beta_{j(\theta)}$, com $\theta \neq \tau$ e $\theta, \tau \in (0,1)$. Ou seja, é

Figura 14 – Estimativas (linhas contínuas) e intervalo de confiança de 95% (área hachurada) para os coeficientes de regressão considerando um conjunto denso de quantis, $\tau = 0.05, 0.10, \ldots, 0,90$. A linha horizontal em zero é marcada como referência. As linhas pontilhadas referem-se ao intervalo de confiança de mínimos quadrados ordinários.



possível testar de forma simultânea múltiplos coeficientes e verificar se a função do τ-ésimo e θ-ésimo quantil condicional são diferentes uma da outra. Além disso, este teste permite verificar a real necessidade de se utilizar a regressão quantílica visto que não é preciso ajustar diferentes modelos se o efeito das covariáveis é uniforme ao longo dos quantis (SANTOS, 2012). Como resultado obteve-se um p-valor< 0.001 ao testar todos os parâmetros de regressão nos quantis $\tau = (0.15, 0.25, 0.50, 0.75, 0.90)$ de forma simultânea. Portanto, a rejeição da hipótese nula sugere que pelo menos um dos coeficientes é estatisticamente diferente dos demais, justificando o uso da regressão quantílica.

É possível ainda aplicar o teste de Wald individualmente para cada parâmetro e suas hipóteses serão definidas da seguinte maneira $\mathcal{H}_0 : \beta_{j(\tau)} = \beta_{j(\theta)}$ versus a alternativa $\mathcal{H}_1 : \beta_{j(\tau)} \neq$

$\beta_{j(\theta)}$. Ou seja, a hipótese nula sugere que a variável associada ao parâmetro β_j apresenta efeito uniforme nos diferentes quantis especificados. A Tabela 8 contém os p-valores obtidos após comparar a estimativa atual (por exemplo, no quantil 0.15) contra as demais estimativas referente ao primeiro quartil, ao mediano, ao terceiro quartil e a posição no quantil 0.90. Como resultado, é observado que o efeito da variável *URB* no quantil 0.90 é significativamente diferente daqueles nos quantis 0.15, 0.25, 0.50 e 0.75, considerando o nível nominal de 5%. Por outro lado, O efeito das variáveis *INAT* e *ALC* nos quantis 0.25 e 0.50 mostraram diferenças significativas quando comparadas com as estimativas no quantil 0.90.

Tabela 8 – Teste de igualdade dos parâmetros de regressão. Diferença das estimativas nos quantis 0.15, 0.25, 0.50, 0.75, 0.85 e 0.90.

		τ/p-valor					
τ	Variáveis	0.15	0.25	0.50	0.75	0.85	0.90
0.15	*INAT*	-	0.905	0.826	0.577	0.368	0.118
	URB	-	0.422	0.247	0.938	0.101	0.012*
	ALC	-	0.457	0.575	0.587	0.196	0.305
0.25	*INAT*	-	-	0.409	0.504	0.200	0.001*
	URB	-	-	0.415	0.569	0.138	0.004*
	ALC	-	-	0.729	0.094	0.001*	0.003*
0.50	*INAT*	-	-	-	0.231	0.071	< 0.001*
	URB	-	-	-	0.271	0.273	0.024*
	ALC	-	-	-	0.082	0.002*	0.008*
0.75	*INAT*	-	-	-	-	0.530	0.172
	URB	-	-	-	-	0.011*	0.001*
	ALC	-	-	-	-	0.149	0.462
0.85	*INAT*	-	-	-	-	-	0.215
	URB	-	-	-	-	-	0.225
	ALC	-	-	-	-	-	0.436

*: p-valor menor que o nível de significância de 5%.

Portanto, o efeito destas covariáveis sobre a proporção de adultos obesos é positivo e pode variar dependendo da posição observada na distribuição condicional da variável resposta. Especificamente, quando comparado o efeito exercido nos quantis inferiores e mediano com o efeito no quantil superior, ou seja, na posição 0.90.

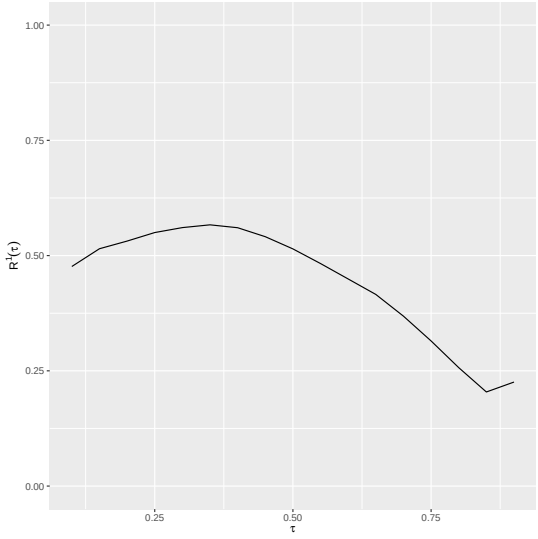
Com o objetivo de avaliar a qualidade do ajuste dos modelos utilizou-se o teste da falta de ajuste para verificar a hipótese nula de linearidade do modelo contra a hipótese alternativa de não linearidade (HE; ZHU, 2003). Outra maneira de verificar a qualidade do ajuste é a partir de uma medida similar ao pseudo-R^2, introduzida por Koenker e Machado (1999), denotada por $R^1(\tau)$. Na Tabela 9 pode-se verificar que apenas o modelo ajustado no quantil mediano rejeitou a hipótese de linearidade, uma vez que seu p-valor é menor do que o nível de significância de 5%. Em relação a medida da bondade de ajuste, $R^1(\tau)$, é possível notar um decrescimento dos seus valores ao longo dos quantis, nos levando a entender que a contribuição das variáveis independentes na

distribuição condicional da variável resposta não é a mesma. Ou seja, suas contribuições são maiores na cauda inferior da distribuição. Vale ressaltar que a maior contribuição é avaliada no quantil 0.25 apresentando um valor de 0.55. Através da Figura 15 é possível ter uma visão mais geral da contribuição destas covariáveis para explicar a variação do logit da proporção de adultos obesos.

Tabela 9 – Medida da bondade de ajuste e p-valor referente ao teste da falta de ajuste.

τ-ésimo quantil	Medida da bondade de ajuste	p-valor
0.15	0.52	0.075
0.25	0.55	0.143
0.50	0.52	0.004
0.75	0.37	0.088
0.85	0.20	0.045
0.90	0.23	0.354

Figura 15 – Medida da bondade de ajuste, $R^1(\tau)$, para o modelos de regressão quantílica.



Para avaliar a qualidade do ajuste por meio da análise dos resíduos pode-se utilizar o gráfico de envelope simulado exibido na Figura 16. Como apresentado em Santos (2012) o uso dessa técnica requer a suposição de alguma distribuição probabilidade para a variável resposta. Um resultado bastante conhecido na literatura diz que o estimador de mínimos quadrados ordinários coincide com o estimador de máxima verossimilhança quando os erros seguem distribuição normal padrão. De forma similar, para modelos de regressão quantílica tem-se que o estimador dos desvios absolutos ponderados coincide com o estimador de máxima verossimilhança quando erros seguem distribuição Laplace Assimétrica, $\text{LA}(\mu, \sigma, \tau)$, com parâmetros μ, σ e τ (YU; ZHANG, 2005). Sua distribuição acumulada pode ser definida como

$$F(y;\mu,\sigma,\tau) = \begin{cases} \tau \exp\left(\dfrac{(1-\tau)}{\sigma}(y-\mu)\right), & \text{se} \quad y \leq \mu, \\ 1-(1-\tau)\exp\left(-\dfrac{\tau}{\sigma}(y-\mu)\right), & \text{se} \quad y > \mu, \end{cases}$$

em que μ é o parâmetro de locação, σ é o parâmetro de escala e τ é o parâmetro de assimetria. Além disso, é possível considerar os resíduos quantílicos propostos por Dunn e Smyth (1996), para a construção dos gráficos, uma vez que estes convergem para a distribuição normal padrão quandos os parâmetros β e σ são consistentemente estimados. Tais resíduos podem ser definidos como

$$r_{q,t} = \Phi^{-1}\{F(y_t;\hat{\mu}_t,\hat{\sigma},\tau)\},$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição normal padrão, $F(y;\mu,\sigma,\tau)$ é a função de distribuição acumulada da Laplace Assimétrica e $r_{q,t}$ são os resíduos quantílicos. A Figura 16 apresenta o gráfico do envelope simulado considerando os resíduos quantílicos. A partir dessa análise gráfica é possível verificar a qualidade do ajuste dos modelos de regressão quantílica. Como resultado é verificado que os ajuste com base nos quantis 0.25 e 0.50 apresentaram uma melhor adequação uma vez que as as observações ou resíduos encontram-se distribuídos dentro das bandas assintóticas do envelope simulado.

Por fim, é possível computar o efeito marginal de uma determinada covariável por meio de suas respectivas derivadas parciais (KOENKER, 2005; GERACI, 2016). Ou seja, considere a função do τ-ésimo quantil condicional dada por

$$Q_\tau(h(Y)|X) = X\beta_{(\tau)},$$

em que Y é a variável resposta, X é a matriz das covariáveis independentes e $h(\cdot)$ é uma transformação não decrescente no conjunto $\mathbb{R}$ com inversa h^{-1}. Dessa forma, o efeito marginal pode ser obtido da seguinte maneira

$$\frac{\partial Q_{(\tau)}(Y|X)}{\partial x_j} = \frac{\partial h^{-1}\{Q_{(\tau)}(h(Y)|X)\}}{\partial x_j},$$

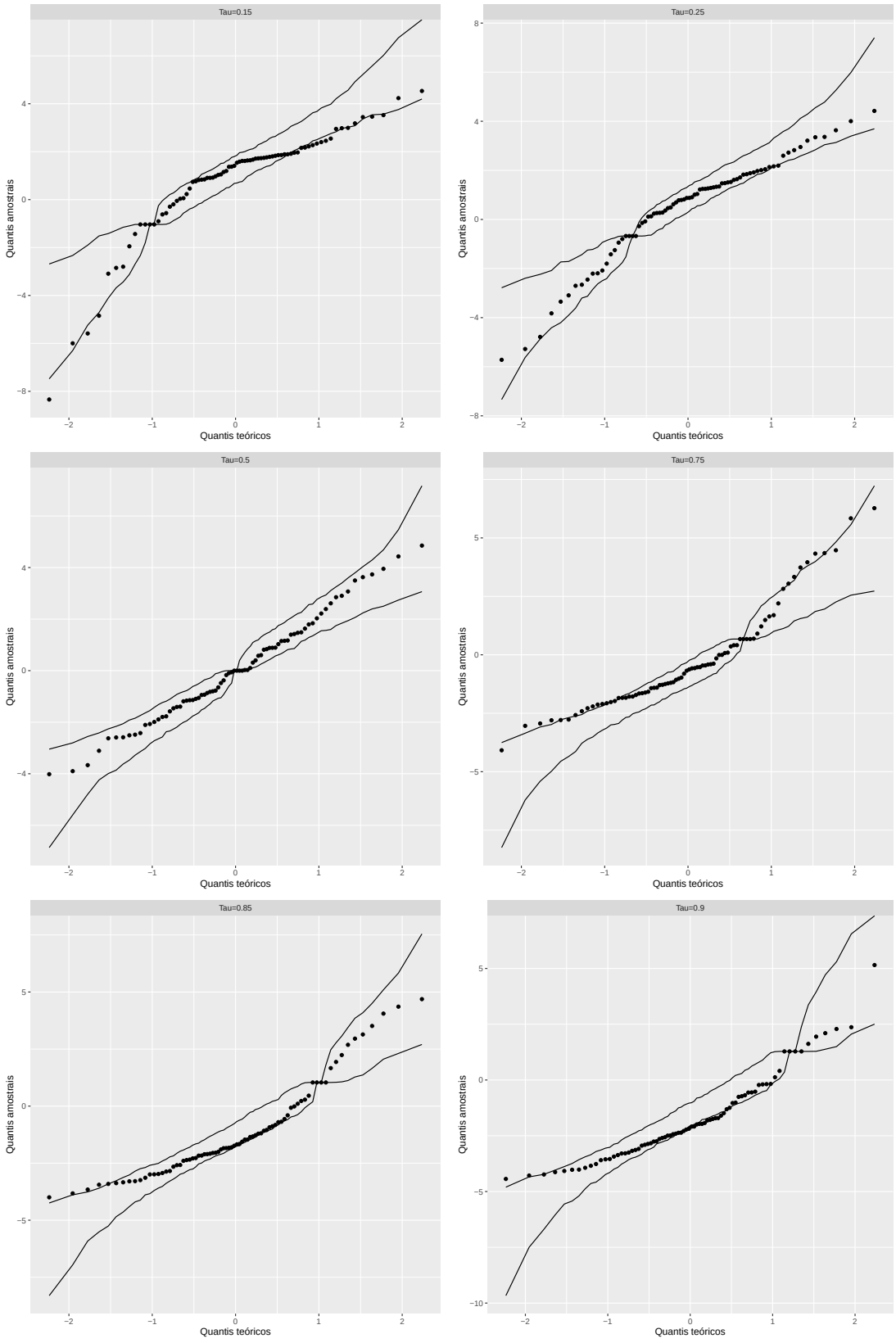
em que x_j é a j-ésima covariável com relação ao qual o efeito marginal deve ser calculado. Portanto, como a transformação logit foi utilizada para modelar a proporção de adultos obesos a inversa, h^{-1}, é definida como

$$h^{-1}\{Q_{(\tau)}(\text{logit}(Y)|X)\} = \frac{exp(\beta_{0(\tau)} + \beta_{1(\tau)}INAT_t + \beta_{2(\tau)}URB_t + \beta_{3(\tau)}ALC_t)}{1 + exp(\beta_{0(\tau)} + \beta_{1(\tau)}INAT_t + \beta_{2(\tau)}URB_t + \beta_{3(\tau)}ALC_t)}.$$

Então o efeito marginal da variável $INAT$ pode ser obtido da seguinte maneira

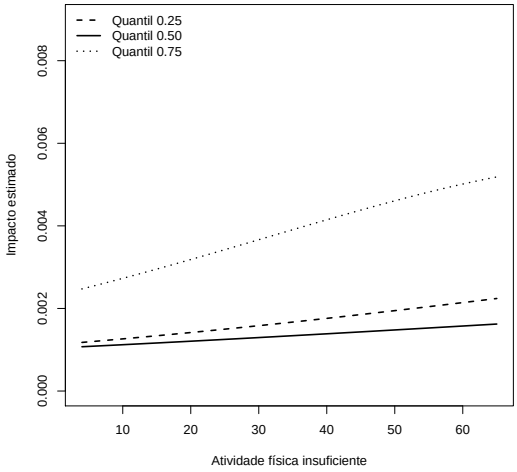
$$\frac{\partial h^{-1}\{Q_{(\tau)}(\text{logit}(Y)|X)\}}{\partial INAT_j} = \frac{\beta_{1(\tau)} \times (exp(\beta_{0(\tau)} + \beta_{1(\tau)}INAT_t + \beta_{2(\tau)}URB_t + \beta_{3(\tau)}ALC_t))}{(1 + exp(\beta_{0(\tau)} + \beta_{1(\tau)}INAT_t + \beta_{2(\tau)}URB_t + \beta_{3(\tau)}ALC_t))^2}.$$

Figura 16 – Gráfico de envelope simulado para modelos de regressão quantílica considerando os resíduos quantílicos e os quantis condicionais 0.15, 0.25, 0.50, 0.75, 0.85 e 0.90 de *OB*2014.



Com o objetivo de estimar as curvas de impacto ou efeito marginal da atividade física insuficiente sobre a proporção de adultos obesos nas nações considerou-se três situações, como apresentado na Figura 17. Ou seja, os ajustes referentes aos quantis 0.25, 0.50 e 0.75, além disso as covariáveis URB e ALC estão fixadas na mediana. Dessa forma, é possível variar os valores de $INAT$ para determinar o aumento provocado na variável resposta. Como resultado, observa-se que o impacto é positivo e cresce lentamente a medida que aumenta-se os valores de $INAT$. A curva estimada no quantil 0.25 apresenta um impacto maior que no quantil 0.50, além disso o maior impacto é observado no quantil 0.75. Por fim, é observado que não existem grandes diferenças entre as curvas nos quantis 0.25 e 0.50.

Figura 17 – Efeito marginal da atividade física insuficiente sobre a proporção de adultos obesos considerando os ajustes nos quantis 0.25, 0.50 e 0.75.



5.4 COMPARAÇÃO ENTRE OS MODELOS DE REGRESSÃO BETA E QUANTÍLICA

Comparando-se os resultados obtidos para o ajuste da regressão beta com dispersão variável com os resultados obtidos para o ajuste da regressão quantílica é possível verificar algumas semelhanças e diferenças. Por exemplo, as covariáveis porcentagem de indivíduos que praticam atividade física insuficiente $(INAT)$, porcentagem da população que vivem em áreas urbanas (URB) e consumo médio de álcool por pessoa em litros (ALC) continuaram relevantes para os dois modelos. Por outro lado, a variável expectativa de vida em anos ($VIDA$) deixou de ser relevante no ajuste da regressão quantílica. Além disso, constatou-se que todas as covariáveis significativas para explicar a obesidade adulta influenciam positivamente a mesma. Contudo, a intensidade com a qual tais covariáveis influenciam tal prevalência pode ser obtidas por meio das curvas de impacto ou efeitos marginais, apresentadas nesta dissertação. Em relação ao impacto estimado sobre a variável resposta é verificado que em ambos os modelos $INAT$ causa um efeito positivo e crescente a medida que aumenta-se os valores desta covariável. Contudo, no modelo

referente a regressão quantílica a curva estimada no quantil 0.25 apresenta efeito maior do que a curva no quantil 0.50.

Para comparação do melhor modelo utilizou-se medidas para avaliar a precisão das estimativas oriundas do método selecionado. As medidas erro quadrático médio (EQM), erro absoluto médio (EAM) e erro percentual total (EPT) buscam comparar os valores preditos pelos modelos com os valores observados (GOMES, 2003) e podem ser definidas como

$$EQM = \frac{1}{n}\sum_{j=1}^{n} e_j^2,$$

$$EAM = \frac{1}{n}\sum_{j=1}^{n} |e_j|,$$

$$EPT = \left(\frac{\sum_{j=1}^{n} e_j}{\sum_{j=1}^{n} y_j}\right) \times 100,$$

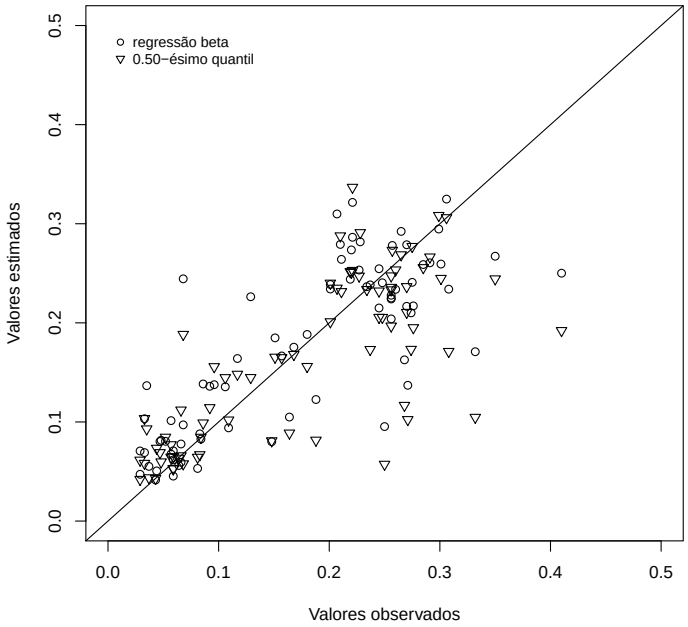
onde $\hat{y}_j$ é a estimativa de y_j, $e_j = y_j - \hat{y}_j$ e n é o número de observações. Assim, o modelo que apresentar o menor erro será consequentemente o modelo escolhido para se obter as melhores estimativas. A Tabela 10 apresenta as medidas de erro obtidas a partir dos valores ajustados dos modelos de regressão beta e quantílica. Como resultado, verifica-se que o modelo de regressão beta apresentou os menores valores em duas das medidas. Por outro lado, o ajuste no quantil de ordem 0.50 apresentou o segundo menor erro ao se avaliar as três medidas. Adicionalmente, a Figura 18 apresenta o gráfico dos valores observados versus os valores estimados o que permiti a comparação entre os ajustes obtidos a partir dos modelos de regressão beta e quantílica para $\tau = 0.50$. A partir da análise gráfica não é possível verificar grandes diferenças entre os ajustes uma vez que as estimativas de ambos os modelos se mostram praticamente equidistantes da reta.

Tabela 10 – Variações percentuais nas estimativas dos parâmetros ao se retirar observações influentes. Proporção de adultos obesos nas nações em 2014.

Modelos	EQM	EAM	EPT
Beta	**0.0035**	0.0438	**−0.7600**
Quantil 0.15	0.0073	0.0614	31.1843
Quantil 0.25	0.0058	0.0514	22.8598
Quantil 0.50	0.0044	**0.0435**	7.3693
Quantil 0.75	0.0064	0.0569	−16.9205
Quantil 0.85	0.0109	0.0840	−42.8779
Quantil 0.90	0.0172	0.1043	−58.6039

Por fim, tem-se que o ajuste obtido por meio do modelo de regressão beta com dispersão variável se mostrou mais adequado para explicar a obesidade adulta nas nações. Isto devido a melhor precisão de suas estimativas e ao maior poder de explicação proporcionado por suas covariáveis (medida a partir do pseudo-R^2). Por outro lado, o ajuste relacionado a regressão quantílica apresentou algumas violações quando avaliado o gráfico dos resíduos e o teste de linearidade para alguns quantis.

Figura 18 – Gráfico dos valores observados versus os valores estimados da variável *OB*2014. Considerando os modelos obtidos para a regressão beta e o quantil de ordem 0.50 da regressão quantílica.



6 CONSIDERAÇÕES FINAIS

A base de dados utilizada nesta dissertação é constituída por 78 observações, correspondente as nações no mundo, das quais 25 (32%) pertencem a África, 11 (14%) pertencem a América, 14 (18%) pertencem a Ásia, 25 (32%) pertencem a Europa e 3 (4%) pertencem a Oceania. De acordo com a realização do mesmo foi visto que 50% das nações apresentam valores de obesidade maiores que 0.20. Além disso, a expectativa de vida média delas oscila em torno de 72 anos. Vale ressaltar que os valores de atividade física insuficiente são maiores que 23.80% em 50% dos países. A partir da análise do *box-plot* foi observado uma possível diferença nas proporções de adultos obesos entre os continentes da América e Europa com os da África e Ásia. Além disso, a construção do mapa da obesidade em 2014 permitiu visualizar que os países com menores valores de obesidade estão concentrados na Ásia e África, enquanto que os países com maiores valores encontram-se na América e Europa.

O modelo de regressão beta utilizado definiu que as covariáveis porcentagem de atividade física insuficiente, porcentagem da população que vivem em áreas urbanas, expectativa de vida em anos e o consumo médio de álcool por pessoa em um ano produzem um efeito positivo sobre a obesidade. Ou seja, elas tendem a aumentar os valores da proporção de adultos obesos quando aumentamos cada uma indidualmente enquanto que as demais permanecem constantes.

Através da análise de regressão quantílica foi possível constatar que as covariáveis porcentagem de atividade física insuficiente, consumo médio de álcool por pessoa em um ano e porcentagem da população que vivem em áreas urbanas tendem a aumentar as proporções de adultos obesos nas nações, umas vez que o efeito individual destas covariáveis é também positivo. Adicionalmente, ao ajustar modelos para diferentes quantis observou-se que em alguns pontos da distribuição condicional da variável resposta o impacto destas covariáveis são estatisticamente diferentes.

Por fim, comparando os resultados obtidos a partir das duas abordagens foi possível verificar algumas semelhanças e diferenças. Por exemplo, em ambos os modelos as covariáveis que demostraram ter efeito significativo para explicar a proporção de adultos obesos tendem individualmente a aumentar os valores da variável resposta. Contudo, foi visto que a variável expectativa de vida foi significativa apenas no modelo de regressão beta com dispersão variável. Vale ressaltar que a contribuição das covariáveis para explicar a obesidade nas nações foi maior no modelo de regressão beta. Além disso, analisando as medidas de erros de previsão verificou-se que as estimativas oriundas da regressão beta são mais precisas quando avaliado o erro quadrático médio e o erro percentual total. Portanto, para questões de predizer valores referentes a obesidade adulta nas nações em 2014 o modelo de regressão beta com dispersão variável se mostrou mais adequado para tal propósito.

REFERÊNCIAS

ALMEIDA JUNIOR, P.; SOUZA, T. Estimativas de votos da presidente Dilma Roussef nas eleições presidenciais de 2010 sob o âmbito do bolsa família. *Ciência e Natura*, v. 37, n. 1, p. 12–22, 2015. Citado 3 vezes nas páginas 19, 20 e 40.

ANDRADE, R.; PEREIRA, R.; SICHIERI, R. Consumo alimentar de adolescentes com e sem sobrepeso do município do Rio de Janeiro. *Cadernos de Saúde Pública*, v. 19, n. 5, p. 1485–1495, 2003. Citado na página 12.

ANJOS, L. *Obesidade e saúde pública*. [S.l.]: FIOCRUZ, 2006. Citado 2 vezes nas páginas 41 e 48.

ANTIPORTA, D. et al. Length of urban residence and obesity among within-country rural-to-urban Andean migrants. *Public Health Nutrition*, v. 19, n. 7, p. 1270–1278, 2015. Citado na página 14.

ARTERBURN, D.; MACIEJEWSKI, M.; TSEVAT, J. Impact of morbid obesity on medical expenditures in adults. *International Journal of Obesity*, v. 29, n. 3, p. 334–339, 2005. Citado na página 13.

BAHIA, L. et al. The costs of overweight and obesity-related diseases in the Brazilian public health system: Cross-sectional study. *BioMed Central Public Health*, v. 12, p. 440, 2012. Citado na página 13.

BOTTAI, M.; CAI, B.; MCKEOWN, R. Logistic quantile regression for bounded outcomes. *Statistics in Medicine*, v. 29, n. 2, p. 309–317, 2010. Citado 2 vezes nas páginas 25 e 47.

BOX, G.; COX, D. An analysis of transformations. *Journal of the Royal Statistical Society*, v. 26, n. 2, p. 211–252, 1964. Citado na página 18.

BOX, G.; DRAPER, N. *Empirical Model-Building and Response Surfaces*. [S.l.]: Wiley, 1987. Citado na página 18.

BUCHINSKY, M. Recent advances in quantile regression models: A pratical guideline for empirical research. *The Journal of Humam Resources*, v. 33, n. 1, p. 88–126, 1998. Citado 2 vezes nas páginas 21 e 49.

BURGOS, M. et al. Associação entre medidas antropométricas e fatores de risco cardiovascular em crianças e adolescentes. *Arquivos Brasileiros de Cardiologia*, v. 101, n. 4, p. 288–296, 2013. Citado na página 13.

CABRERA, M.; FILHO, W. Obesidade em idosos: Prevalência, distribuição e associação com hábitos e co-morbidades. *Arquivos brasileiros de Endocrinologia e Metabologia*, v. 45, n. 5, p. 494–501, 2001. Citado na página 13.

COOK, R. Detection of influential observation in linear regression. *Technometrics*, v. 19, n. 1, p. 15–18, 1977. Citado na página 43.

CRIBARI-NETO, F.; SOUZA, T. Testing inference in variable dispersion beta regressions. *Journal os Statistical Computation and Simulation*, v. 82, n. 12, p. 1827–1843, 2012. Citado 2 vezes nas páginas 19 e 20.

CRIBARI-NETO, F.; SOUZA, T. Religious belief and intelligence: Worldwide evidence. *Intelligence*, v. 41, n. 5, p. 482–489, 2013. Citado 3 vezes nas páginas 19, 20 e 45.

CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in R. *Journal of Statistical Software*, v. 34, n. 2, p. 1–24, 2010. Citado 2 vezes nas páginas 28 e 40.

DRAPER, N.; SMITH, H. *Applied Regression Analysis*. [S.l.]: Editora Wily-Interscience, 1998. Citado 2 vezes nas páginas 17 e 21.

DUNCAN, B. et al. Doenças crônicas não transmissíveis no Brasil: Prioridade para enfrentamento e investigação. *Revista Saúde Pública*, v. 46, p. 126–134, 2012. Citado na página 12.

DUNN, P.; SMYTH, G. Randomized quantile residuals. *Journal of Computacional and Graphical Statistics*, v. 5, n. 3, p. 236–244, 1996. Citado na página 53.

ESPINHEIRA, P.; FERRARI, S.; CRIBARI-NETO, F. Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, v. 52, n. 9, p. 4417–4431, 2008. Citado 3 vezes nas páginas 19, 20 e 43.

ESPINHEIRA, P.; FERRARI, S.; CRIBARI-NETO, F. On beta regression residuals. *Journal of Applied Statistics*, v. 35, n. 4, p. 407–419, 2008. Citado 3 vezes nas páginas 19, 20 e 43.

FEIZI, A.; ALIYARI, R.; ROOHAFZA, H. Association of perceived stress with stressful life events, lifestyle and sociodemographic factors: A large-scale community-based study using logistic quantile regression. *Computational and Mathematical Methods in Medicine*, v. 2012, p. 1–12, 2012. Citado na página 47.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modeling rates and proportions. *Journal os Applied Statistics*, v. 31, n. 7, p. 799–815, 2004. Citado 3 vezes nas páginas 18, 19 e 42.

FERRARI, S.; ESPINHEIRA, P.; CRIBARI-NETO, F. Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, v. 65, n. 3, p. 337–351, 2011. Citado na página 44.

FRANCISCHI, R. et al. Obesidade: Atualização sobre sua etiologia, morbidade e tratamento. *Revista de Nutrição*, v. 13, n. 1, p. 17–28, 2000. Citado na página 12.

GERACI, M. *Qtools: A Useful Package for Quantiles*. [S.l.], 2016. Disponível em: <https://cran.r-project.org/web/packages/Qtools/vignettes/Qtools.pdf>. Citado 2 vezes nas páginas 18 e 53.

GERACI, M.; JONES, M. Improved transformation-based quantile regression. *The Canadian Journal of Statistics*, v. 43, n. 1, p. 118–132, 2015. Citado na página 18.

GIGANTE, D. et al. Obesidade da população adulta de Pelotas, Rio Grande do Sul, Brasil e associação com nível sócio-econômico. *Cadernos de Saúde Pública*, v. 22, n. 9, p. 1873–1879, 2006. Citado na página 12.

GOMES, A. *Modelagem e previsão da arrecadação do imposto de renda no Brasil*. Dissertação (Mestrado) — Modelagem e previsão da arrecadação do imposto de renda do Brasil, 2003. Citado na página 56.

HAO, L.; NAIMAN, D. *Quantile regression*. [S.l.]: Sage publications, 2007. Citado 2 vezes nas páginas 25 e 47.

HE, X.; ZHU, L. A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, v. 98, n. 464, p. 1013–1022, 2003. Citado na página 51.

HUBER, P. The behavior of maximum likelihood estimation under non-standard conditions. *In Proceedings of the Fifth Berkeley Symposium on Mathematics Statistics and Probability*, p. 221–233, 1967. Citado na página 24.

JUNG, R. Obesity as a disease. *British Medical Bulletin*, v. 53, n. 2, p. 307–321, 1997. Citado na página 12.

KIESCHNICK, R.; MCCULLOUGH, B. Regression analysis of variates observed on (0,1): Percentages, proportions and fractions. *Statistical Modelling*, v. 3, n. 3, p. 193–213, 2003. Citado na página 18.

KOENKER, R. *Quantile regression*. [S.l.]: Cambridge University Press, 2005. Citado 6 vezes nas páginas 22, 23, 24, 28, 47 e 53.

KOENKER, R. Censored quantile regression redux. *Journal of Statistical Software*, v. 27, n. 6, 2008. Citado na página 47.

KOENKER, R.; BASSETT, G. Regression quantiles. *Econometrica*, v. 46, n. 1, p. 33–50, 1978. Citado 3 vezes nas páginas 21, 24 e 47.

KOENKER, R.; BASSETT, G. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, v. 50, n. 1, p. 43–61, 1982. Citado 4 vezes nas páginas 21, 22, 23 e 49.

KOENKER, R.; MACHADO, J. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, v. 94, n. 448, p. 1296–1310, 1999. Citado 2 vezes nas páginas 22 e 51.

LIMA, L. *Um teste de especificação correta para modelos de regressão beta*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2007. Citado na página 42.

LUCA, F. D.; BOCCUZZO, G. What do healthcare workers know about sudden infant death syndrome?: The results of the Italian campaign 'Genitoripiù'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, v. 177, n. 1, p. 63–82, 2014. Citado na página 47.

MARIATH, A. et al. Obesidade e fatores de risco para o desenvolvimento de doenças crônicas não transmissíveis entre usuários de unidade de alimentação e nutrição. *Cadernos de Saúde Pública*, v. 23, n. 4, p. 897–905, 2007. Citado na página 12.

MARTíNEZ-GONZáLEZ, M. et al. Physical inactivity, sedentary lifestyle and obesity in the European Union. *International Journal of Obesity*, v. 23, n. 11, p. 1192–1201, 1999. Citado na página 14.

MCCULLAGH, P.; NELDER, J. *Generalized linear models*. [S.l.]: London: Chapman and Hall, 1989. Citado 2 vezes nas páginas 20 e 25.

Ministério da Saúde. *Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas Não Transmissíveis (DCNT) no Brasil 2011-2022*. [S.l.], 2011. Citado na página 12.

MOSTELLER, F.; TUKEY, J. *Data analysis and regression: A second course in statistics*. [S.l.]: Addison-Wesley, 1977. Citado na página 21.

MUA, Y.; HEA, X. Power transformation toward a linear regression quantile. *Journal of the American Statistical Association*, v. 102, n. 477, p. 37–41, 2007. Citado na página 25.

NASCIMENTO, A. et al. Eficiência técnica da atividade leiteira em Minas Gerais: Uma aplicação de regressão quantílica. *Revista Brasileira de Zootecnia*, v. 41, n. 3, p. 783–789, 2012. Citado na página 23.

NETO, A. et al. Estado nutricional alterado e sua associação com perfil lipídico e hábitos de vida em idosos hipertensos. *SciELO Brazil*, v. 58, n. 4, p. 350–356, 2008. Citado na página 13.

NEYMAN, J.; PEARSON, E. On the use and interpretation of certain teste criteria for purposes of statistical inference. *Biometrika*, v. 20, p. 175–240, 1928. Citado 3 vezes nas páginas 22, 40 e 49.

OECD. *Obesity Update 2014 - OECD*. [S.l.], 2014. Disponível em: <http://www.oecd.org/health/Obesity-Update-2014.pdf>. Citado na página 13.

PAPKE, L.; WOOLDRIDGE, J. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, v. 11, n. 6, p. 619–632, 1996. Citado na página 18.

PAULA, G. *Modelos de Regressão com Apoio Computacional*. [S.l.], 2004. Disponível em: <https://www.ime.usp.br/~giapaula/texto_2013.pdf>. Citado na página 18.

PEREIRA, T. *Regressão beta inflacionada: Inferência e aplicações*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2010. Citado na página 19.

PINHEIRO, A.; FREITAS, S.; CORSO, A. Uma abordagem epidemiológica da obesidade. *Revista de Nutrição*, v. 17, n. 4, p. 523–533, 2004. Citado na página 12.

POWELL, J. Censored regression quantiles. *Journal of Econometrics*, v. 32, p. 143–155, 1986. Citado na página 25.

PRESS, W. et al. *Numerical recipes in C: The art of scientific computing*. [S.l.]: Cambridge University Press, 1992. Citado na página 20.

PUGLIA, C. Indicações para o tratamento operatório da obesidade mórbida. *Revista da Associação Médica Brasileira*, v. 50, n. 2, p. 118–118, 2004. Citado na página 12.

R Core Team. *R: A language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>. Citado 2 vezes nas páginas 28 e 47.

RAMSEY, J. B. Tests for specification erros in classical linear least squares regression analysis. *Journal of the Royal Statistical Society*, v. 31, n. 2, p. 350–371, 1969. Citado na página 42.

SANTOS, B. *Modelos de regressão quantílica*. Dissertação (Mestrado) — Universidade de São Paulo, 2012. Citado 4 vezes nas páginas 25, 47, 50 e 52.

SHELTON, N.; KNOTT, C. Association between alcohol calorie intake and overweight and obesity in english adults. *American journal of public health*, v. 104, n. 4, p. 629–631, 2014. Citado na página 14.

SILVA, C.; SOUZA, T. Modelagem da taxa de analfabetismo no estado da Paraíba via modelo de regressão beta. *Revista Brasileira de Biometria*, v. 32, n. 3, p. 345–359, 2014. Citado 3 vezes nas páginas 19, 20 e 40.

SIMAS, A.; BARRETO-SOUZA, W.; ROCHA, A. Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, v. 54, n. 2, p. 348–366, 2010. Citado 2 vezes nas páginas 19 e 20.

SMITHSON, M.; VERKUILEN, J. A better lemon-squeezer? Maximum likelihood regression with beta-distribuited dependent variables. *Psychological Methods*, v. 11, n. 1, p. 54–71, 2006. Citado 3 vezes nas páginas 19, 20 e 40.

SOUZA, F.; SCHROEDER, P.; LIBERALI, R. Obesidade e envelhecimento. *Revista Brasileira de Obesidade, Nutrição e Emagrecimento*, v. 1, n. 2, p. 24–35, 2007. Citado na página 41.

SOUZA, S. et al. Modelagem da proporção de obesos nos Estados Unidos utilizando modelo de regressão beta com dispersão variável. *Ciência e Natura*, v. 38, n. 3, p. 1146–1156, 2016. Citado 5 vezes nas páginas 19, 20, 40, 42 e 43.

SOUZA, T.; CRIBARI-NETO, F. Intelligence, religiosity and homosexuality non-acceptance: Empirical evidence. *Intelligence (Norwood)*, v. 52, p. 63–70, 2015. Citado 3 vezes nas páginas 19, 20 e 42.

STOL, A. et al. Complicações e óbitos nas operações para tratar a obesidade mórbida. *Arquivos Brasileiros de Cirurgia Digestiva*, v. 24, n. 4, p. 282–284, 2011. Citado na página 13.

WALD, A. Test of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, v. 54, n. 3, p. 426–482, 1943. Citado na página 40.

WEI, B.; HU, Y.; FUNG, W. Generalized leverage and its applications. *Scandinavian Journal of Statistics*, v. 25, n. 1, p. 25–37, 1998. Citado na página 44.

YU, K.; ZHANG, J. A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics—Theory and Methods*, v. 34, p. 1867–1879, 2005. Citado na página 52.

Apêndice A – *Script* utilizado no *software* R

```
#Fixando semente
set.seed(2)
#Pacotes
library(betareg)
library(quantreg)
library(VGAM)
library(ggplot2)
library(lmtest)
library(Qtools)
#################################################################
####Dados
#################################################################
dados= read.table("dados_dissert.txt", header=T)
attach(dados);names(dados);str(dados)
#################################################################
####Análise descritiva
#################################################################
summary(y)
hist(y,ylab="Frequência",main="",xlab="Proporção de adultos obesos")
boxplot(y,horizontal = TRUE, xlab="Proporção de adultos obesos")
boxplot(y~continente, xlab="OB2014")
boxplot(inat~continente, xlab="INAT")
boxplot(urb~continente, xlab="URB")
boxplot(vida~continente, xlab="VIDA")
boxplot(educ~continente, xlab="EDUC")
X=data.frame(inat,vida,gdp,alc,urb,pop,educ)
cor(y,X);cor(X)
####Modelo de regressão beta com dispersão variável
#################################################################
####Ajustando modelo
#################################################################
#Dispersão fixa
modelo1=  betareg(y ~inat+urb+alc+vida,
link="loglog")
summary(modelo1)
#Dispersão variável
modelo2=  betareg(y ~inat+urb+alc+vida|vida+educ+alc,
link="loglog", link.phi="log")
```

```
summary(modelo2)
#####################################################################
####Teste da razão de verossimilhanças
#####################################################################
lrtest(modelo1,modelo2)
#####################################################################
####Teste RESET
#####################################################################
#Predito linear estimado ao quadrado
lrtest(modelo2, . ~ . + I(predict(modelo2, type = "link")^2))
#Predito linear estimado ao cubo
lrtest(modelo2, . ~ . + I(predict(modelo2, type = "link")^3))
#####################################################################
####Análise de diagnóstico
#####################################################################
####Tipo do resíduo
res6=residuals(modelo2,type="sweighted")#Resíduos Ponderados Padronizados
####Gráfico de probabilidade normal com envelope simulado
plot(modelo2, which = 5:5, type="sweighted", sub.caption = "", caption = "",
main = "", ann = 0,lty=2)
title(xlab = "Quantis ", ylab = "Resíduos Ponderados Padronizados
(valores absolutos)")
####Aleatoriedade dos resíduos
plot(res6,main="",ylab="Resíduos ponderados padronizados",
xlab="Índice das observações",ylim=c(-4,4))
abline(0,0,lty=2)
abline(-2,0,lty=2);abline(2,0,lty=2);cutI=-2;cutS=2
abline(-3,0,lty=2);abline(3,0,lty=2)
idI=which(res6<cutI);idS=which(res6>cutS)
text(idS,res6[idS],codigo[idS],pos=3)
text(idI,res6[idI],codigo[idI],pos=1)
####Detectando pontos de influência: Distância de Cook
val_pred=predict(modelo2)
plot(val_pred,cooks.distance(modelo2),xlab="Valores preditos",
ylab="Distância de Cook",ylim=c(0,1))
####Detectando pontos de alavanca
alavan=gleverage(modelo2)
plot(val_pred,alavan,xlab="Valores preditos",ylab="Alavancagem generalizada")
abline(3*mean(alavan),0,lty=2)
lim=3*mean(alavan)
id=which(alavan>lim)
```

```
text(val_pred[id],alavan[id],codigo[id],pos=1)
abline(3*mean(alavan),0,lty=2)
################################################################
####Estimando o impacto da inatividade física
################################################################
modelo2=  betareg(y ~inat+urb+alc+vida|vida+educ+alc,
link="loglog", link.phi="log")
####Definindo grade de valores para atividade física insuficiente
inat.g=seq(4, 65, length=100)
####Fixando as demais covariáveis no 1º quartil
a1=modelo2$coef$mean[1]#intercepto
a2=modelo2$coef$mean[2]*inat.g
a3=modelo2$coef$mean[3]*quantile(urb, 0.25)
a4=modelo2$coef$mean[4]*quantile(alc, 0.25)
a5=modelo2$coef$mean[5]*quantile(vida, 0.25)
impactod1 = exp(-exp(-(a1+a2+a3+a4+a5)))*
(-exp(-(a1+a2+a3+a4+a5))*
(-(modelo2$coef$mean[2])))
####Fixando as demais covariáveis no 2º quartil (Mediana)
a1=modelo2$coef$mean[1]#intercepto
a2=modelo2$coef$mean[2]*inat.g
a3=modelo2$coef$mean[3]*quantile(urb, 0.50)
a4=modelo2$coef$mean[4]*quantile(alc, 0.50)
a5=modelo2$coef$mean[5]*quantile(vida, 0.50)
impactod2 = exp(-exp(-(a1+a2+a3+a4+a5)))*
(-exp(-(a1+a2+a3+a4+a5))*
(-(modelo2$coef$mean[2])))
#Fixando as demais covariáveis no 3º quartil
a1=modelo2$coef$mean[1]#intercepto
a2=modelo2$coef$mean[2]*inat.g
a3=modelo2$coef$mean[3]*quantile(urb, 0.75)
a4=modelo2$coef$mean[4]*quantile(alc, 0.75)
a5=modelo2$coef$mean[5]*quantile(vida, 0.75)
impactod3 = exp(-exp(-(a1+a2+a3+a4+a5)))*
(-exp(-(a1+a2+a3+a4+a5))*
(-(modelo2$coef$mean[2])))
####Curva de impacto
plot(inat.g, impactod1, type="l",  ylab="Impacto estimado",
xlab="taxas de inatividade física", lwd=2,lty=2,ylim=c(0,0.009))
lines(inat.g, impactod2,lty=1, type="l", lwd=2)
lines(inat.g, impactod3,lty=3, type="l", lwd=2)
```

```
legend("topleft", c("Quantil 0.25","Quantil 0.50","Quantil 0.75"),
lty = c(2,1,3), bty="n", lwd=2)
################################################################
####Variação percentual na estimativa dos parâmetros
################################################################
#Ajuste sem os pontos de alavanca:14 64 68
modelo2=  betareg(y ~inat+urb+alc+vida|vida+educ+alc, link="loglog",
link.phi="log")
coefAjusteTODOASobs=coefficients(modelo2)
x1 = c(14)
ajuste1 = update(modelo2, subset = -x1)
summary(ajuste1)
coefSEMobs14=coefficients(ajuste1)
TT1=cbind(round(((coefSEMobs14/coefAjusteTODOASobs) - 1)*100, digits=2))
x2 = c(64)
ajuste2 = update(modelo2, subset = -x2)
summary(ajuste2)
coefSEMobs64=coefficients(ajuste2)
TT2=cbind(round(((coefSEMobs64/coefAjusteTODOASobs) - 1)*100, digits=2))
x3 = c(68)
ajuste3 = update(modelo2, subset = -x3)
summary(ajuste3)
coefSEMobs68=coefficients(ajuste3)
TT3=cbind(round(((coefSEMobs68/coefAjusteTODOASobs) - 1)*100, digits=2))
#Retirando as três observações
x4 = c(14,64,68)
ajuste4 = update(modelo2, subset = -x4)
summary(ajuste4)
coefSEMobsT=coefficients(ajuste4)
TT4=cbind(round(((coefSEMobsT/coefAjusteTODOASobs) - 1)*100, digits=2))
#LAMBDA
phihat = predict(modelo2, type = "precision")
phihat
lambdacomp = max(phihat)/min(phihat)
lambdacomp
####
phihatSEMobs14 = predict(ajuste1, type = "precision")
lambdaSEMobs14 = max(phihatSEMobs14)/min(phihatSEMobs14)
lambdaSEMobs14
MudancaSEMobs14Lbd = round(((lambdaSEMobs14/lambdacomp)-1)*100,digits=2)
MudancaSEMobs14Lbd
```

```
####
phihatSEMobs64 = predict(ajuste2, type = "precision")
lambdaSEMobs64 = max(phihatSEMobs64)/min(phihatSEMobs64)
lambdaSEMobs64
MudancaSEMobs64Lbd = round(((lambdaSEMobs64/lambdacomp)-1)*100,digits=2)
MudancaSEMobs64Lbd
####
phihatSEMobs68 = predict(ajuste3, type = "precision")
lambdaSEMobs68 = max(phihatSEMobs68)/min(phihatSEMobs68)
lambdaSEMobs68
MudancaSEMobs68Lbd = round(((lambdaSEMobs68/lambdacomp)-1)*100,digits=2)
MudancaSEMobs68Lbd
####
phihatSEMobsT = predict(ajuste4, type = "precision")
lambdaSEMobsT = max(phihatSEMobsT)/min(phihatSEMobsT)
lambdaSEMobsT
MudancaSEMobsTLbd = round(((lambdaSEMobsT/lambdacomp)-1)*100,digits=2)
MudancaSEMobsTLbd
respA1=cbind(TT1,TT2,TT3,TT4)
respA2=as.vector(cbind(MudancaSEMobs14Lbd,MudancaSEMobs64Lbd,MudancaSEMobs68Lbd
,MudancaSEMobsTLbd))
rbind(respA1,respA2)

####Modelo de regressão Quantílica
#################################################################
####Envelope simulado via resíduos quantílicos
#################################################################
envel.rq<-function(model, data, ncolunas=1, scales="fixed")
{
tau<-model$tau
rho<-model$rho
if(length(tau)==1)
{
n<-length(residuals(model))
sigmahat<-model$rho/n
predicted<-fitted(model)
res.quant=qnorm(palap(as.numeric(model$y),predicted, model$rho/n, tau=tau))
e<-matrix(0,n,1000)
e1<-numeric(n)
e2<-numeric(n)
for(i in 1:1000)
```

```
{
e[,i]<-ralap(n, predicted, sigmahat, tau=tau)
sim.model<-rq(as.formula(paste("e[,i]~", model$formula[3])), data, tau=tau)
e[,i]<-qnorm(palap(as.numeric(sim.model$y), fitted(sim.model), sim.model$rho/n
, tau=tau))
e[,i]<-sort(e[,i])
}
for(i in 1:n)
{
eo<-sort(e[i,])
e1[i]<-(eo[2]+eo[3])/2
e2[i]<-(eo[997]+eo[998])/2
}
theoretical.quant<-qnorm(1:n/(n+1))
sample.quant<-sort(res.quant)
db<-data.frame(theoretical.quant, sample.quant, Tau=paste("Tau=",
tau,sep=""))
db1<-data.frame(e1=sort(e1),theoretical.quant)
db2<-data.frame(e2=sort(e2),theoretical.quant)
g<-ggplot(db, aes(x=theoretical.quant,y=sample.quant))+geom_point()+
xlab("Quantis teóricos")+ylab("Quantis amostrais")
graph<-g+geom_line(aes(y=e1),db1)+geom_line(aes(y=e2),db2)+
facet_wrap(~Tau)
}
else
{
n<-nrow(residuals(model))
residuos<-list()
for(k in 1:length(tau))
{
residuos[[k]]<-qnorm(palap(as.numeric(model$y),fitted(model)[,k],
model$rho[k]/n,tau=tau[k]))
}
residuos<-lapply(residuos, sort)
residuos<-unlist(residuos)
predicted<-as.vector(fitted(model))
tau.total<-rep(tau, each=n)
e<-list()
e1<-list(n)
e2<-list(n)
for(k in 1:length(tau))
```

```
{
e[[k]]<-matrix(0,n,1000)
e1[[k]]<-numeric(n)
e2[[k]]<-numeric(n)
for(i in 1:1000)
{
e[[k]][,i]<-ralap(n, fitted(model)[,k], model$rho[k]/n,
tau=tau[k])
sim.model<-rq(as.formula(paste("e[[k]][,i]~",model$formula[3])),data,
tau=tau[k])
e[[k]][,i]<-qnorm(palap(as.numeric(sim.model$y),fitted(sim.model),
sim.model$rho/n, tau=tau[k]))
e[[k]][,i]<-sort(e[[k]][,i])
}
for(i in 1:n)
{
eo<-sort(e[[k]][i,])
e1[[k]][i]<-(eo[1]+eo[2])/2
e2[[k]][i]<-(eo[997]+eo[998])/2
}
}
e1<-as.numeric(unlist(lapply(e1,sort)))
e2<-as.numeric(unlist(lapply(e2,sort)))
quantis.teoricos<-vector(length=n*length(tau))
quantis.teoricos<-qnorm(1:n/(n+1))
for(j in 2:length(tau))
{
quantis.teoricos<-c(quantis.teoricos,qnorm(1:n/(n+1)))
}
dados<-data.frame(residuos, quantis.teoricos, tau=tau.total,e1,e2)
g<-ggplot(dados, aes(x=quantis.teoricos, y=residuos, group=tau))+geom_point()+
facet_wrap(~tau, scales="fixed", ncol=ncolunas)
graph<-g+geom_line(aes(y=e1,group=tau),
dados)+geom_line(aes(y=e2,group=tau),dados)+xlab("Quantis teoricos")+
ylab("Quantis amostrais")
}
return(graph)
}
#################################################################
####Medida da bondade de ajuste
#################################################################
```

```
grafR1<-function(model.rqs, trueScale=T)
{
if(class(model.rqs)!="rqs")
{
stop("Você deve usar essa função com objetos do tipo rqs")
}
taus=model.rqs$tau
rho.c=model.rqs$rho
methods=model.rqs$method
y<-model.rqs$y
rho.r<-rq(y~1, tau=taus, method=methods)$rho
R1<-1-rho.c/rho.r
data.graph<-data.frame(taus,R1)
saida=list(values=data.graph, variable=paste(model.rqs$call$formula[3],"",
sep=""))
if(trueScale)graph<-ggplot(data.graph,aes(x=taus, y=R1))+ylim(c(0,1))
else graph<-ggplot(data.graph,aes(x=taus, y=R1))
graph<-graph+geom_line()+ylab(expression(R^{1}*(tau)))+xlab(expression(tau))
saida.final=list(data=saida, graph=graph)
return(saida.final)
}
#####################################################################
####Aplicando a transformação logit: log(y/(1-y))
#####################################################################
logit.y=log(y/(1-y))
####OLS
ols <- lm(logit.y~inat+urb+alc)
summary(ols)
#####################################################################
####Ajustando modelo para os quantis: 0.15, 0.25, 0.50, 0.75 e 0.90
#####################################################################
#Quantil 0.15
mod_15 <- rq(logit.y~inat+urb+alc,tau=0.15)
summary(mod_15,se="nid")
GOFTest(mod_15, alpha = 0.05, B = 1000, seed = 416)
envel.rq(mod_15, data=dados,ncolunas=1)
#Quantil 0.25
mod_25 <- rq(logit.y~inat+urb+alc, tau=0.25)
summary(mod_25,se="nid")
GOFTest(mod_25, alpha = 0.05, B = 1000, seed = 416)
envel.rq(mod_25, data=dados,ncolunas=1)
```

```
#Quantil 0.50
mod_50 <- rq(logit.y~inat+urb+alc, tau=0.50)
summary(mod_50,se="nid")
GOFTest(mod_50, alpha = 0.05, B = 1000, seed = 416)
envel.rq(mod_50, data=dados,ncolunas=1)
#Quantil 0.75
mod_75 <- rq(logit.y~inat+urb+alc, tau=0.75)
summary(mod_75,se="nid")
GOFTest(mod_75, alpha = 0.05, B = 1000, seed = 416)
envel.rq(mod_75, data=dados,ncolunas=1)
#Quantil 0.90
mod_90 <- rq(logit.y~inat+urb+alc, tau=0.90)
summary(mod_90,se="nid")
GOFTest(mod_90, alpha = 0.05, B = 1000, seed = 416)
envel.rq(mod_90, data=dados,ncolunas=1)
#Coeficientes estimados
cbind(coef(mod_15),coef(mod_25),coef(mod_50),coef(mod_75),coef(mod_90))
#################################################################
####Teste de Wald para igualdade dos parâmetros de inclinação
#################################################################
mod15 <- rq(logit.y~inat+urb+alc,tau=0.15)
mod25 <- rq(logit.y~inat+urb+alc, tau=0.25)
mod50 <- rq(logit.y~inat+urb+alc, tau=0.50)
mod75 <- rq(logit.y~inat+urb+alc, tau=0.75)
mod90 <- rq(logit.y~inat+urb+alc, tau=0.90)
#Teste de Wald de forma conjunta
anova(mod15,mod25,mod50,mod75,mod90)
#Teste de Wald para cada variável
anova(mod15,mod25,mod50,mod75,mod90, joint=F)
#################################################################
####Análise gráfica da variação da estimativa dos coeficientes
#################################################################
ajuste1 <- rq(logit.y~inat+urb+alc,tau=seq(0.10,0.90,by=0.05))
plot(summary(ajuste1),level=0.95)
plot(summary(ajuste1),level=0.95, lcol="black",parm=2,main="INAT")
plot(summary(ajuste1),level=0.95, lcol="black",parm=3,main="URB")
plot(summary(ajuste1),level=0.95, lcol="black",parm=4,main="ALC")
#################################################################
####Medida da bondade de ajuste
#################################################################
grafR1(ajuste1)
```

```
################################################################
####Efeito marginal da variável atividade física insuficiente
################################################################
#Efeito Marginal nos quantis 0.25, 0.50 e 0.75
modelo1 <- rq(logit.y~inat+urb+alc, tau=0.25)
modelo2 <- rq(logit.y~inat+urb+alc, tau=0.5)
modelo3 <- rq(logit.y~inat+urb+alc, tau=0.75)
#Considerando grade de valores para a atividade física insuficiente
summary(inat)
inat.g = seq(4,65, length=100)
#Efeito marginal no quantil 0.25: demais variáveis fixadas na mediana
a1=modelo1$coef[1]
a2=modelo1$coef[2]*inat.g
a3=modelo1$coef[3]*quantile(urb, 0.5)
a4=modelo1$coef[4]*quantile(alc, 0.5)
impactod1 = modelo1$coef[2]*(exp(a1+a2+a3+a4))/
((1+(exp(a1+a2+a3+a4)))^2)
#Efeito marginal no quantil 0.50: demais variáveis fixadas na mediana
a1=modelo2$coef[1]
a2=modelo2$coef[2]*inat.g
a3=modelo2$coef[3]*quantile(urb, 0.5)
a4=modelo2$coef[4]*quantile(alc, 0.5)
impactod2 = modelo2$coef[2]*(exp(a1+a2+a3+a4))/
((1+(exp(a1+a2+a3+a4)))^2)
#Efeito marginal no quantil 0.75: demais variáveis fixadas na mediana
a1=modelo3$coef[1]#
a2=modelo3$coef[2]*inat.g
a3=modelo3$coef[3]*quantile(urb, 0.5)
a4=modelo3$coef[4]*quantile(alc, 0.5)
impactod3 = modelo3$coef[2]*(exp(a1+a2+a3+a4))/
((1+(exp(a1+a2+a3+a4)))^2)
#Gráfico com o efeito marginal da inatividade física insuficiente
plot(inat.g, impactod1, type="l",  ylab="Impacto estimado",
xlab="Atividade física insuficiente", lwd=2,lty=2,ylim=c(0,0.009))
lines(inat.g, impactod2,lty=1, type="l", lwd=2)
lines(inat.g, impactod3,lty=3, type="l", lwd=2)
legend("topleft", c("Quantil 0.25","Quantil 0.50","Quantil 0.75"),
lty = c(2,1,3), bty="n", lwd=2)
```