Artigo Original *DOI: 10.5902/2179460X21393*

Ciência e Natura, Santa Maria, v. 38 n. 3, 2016, p. 1146–1156
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA

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

Modeling the proportion of obese in the United States using beta regression model with variable dispersion

Saul de A. Souza*[1], André A. de Oliveira[2], Tatiene C. Souza[3] e Caliandra M. B. L. Lima[4]

[1,2,3]Dep. de Estatística, Universidade Federal da Paraíba, Cidade Universitária, João Pessoa/PB, 58089-900, Brasil
[4]Dep. de Fisiologia e Patologia, Universidade Federal da Paraíba, Cidade Universitária, João Pessoa/PB, 58089-900, Brasil

Resumo

Neste artigo tivemos como objetivo modelar a proporção de adultos obesos nos estados dos Estados Unidos considerando os indivíduos que apresentaram IMC (Índice de Massa Corporal) maior ou igual a 30.0 kg/m^2. Utilizamos o modelo de regressão beta com dispersão variável objetivando explicar a proporção de adultos obesos, uma vez que os dados apresentam assimetria e estão restritos ao intervalo (0,1)*. Os resultados mostraram que a falta de atividade física, o pouco consumo de vegetais por dia, o hábito de fumar e as taxas de insegurança alimentar nos estados, apresentam um efeito positivo no aumento da proporção média de adultos obesos. Por outro lado, as taxas de desemprego e o escore de bem-estar exibem uma relação negativa com o desfecho. Estimamos o impacto das taxas de inatividade física sobre a proporção média de adultos obesos e os resultados revelaram que o efeito desse impacto é positivo e apresenta uma forma acelerada para valores de inatividade física menores do que 0.85.*

Palavras-chave: *Estados Unidos, modelo de regressão beta, obesos.*

Abstract

In this article we aimed to model the proportion of obese adults in the U.S. states considering the subjects who had BMI (Body Mass Index) greater than or equal to 30.0 kg/m^2. We used the beta regression model with variable dispersion with purpose to explain the proportion of obese adults, since the data shows asymmetry and are restricted to the interval (0.1)*. The results showed that lack of physical activity, the low consumption of vegetables per day, smoking and food insecurity rates in the states, have a positive effect on the increase in the average proportion of obese adults, on the other hand, the unemployment rate and the score well-being exhibit a negative relationship with the outcome variable. We estimate the impact of physical inactivity rates on the average proportion of obese adults and the results revealed that the effect of that impact is positive and shows an accelerated way to values of physical inactivity smaller than 0.85.*

Keywords: *United States, beta regression model, obese.*

*Autor para correspondência: saul_ asouza@hotmail.com
Recebido: 11/03/2016 Revisado: 17/08/2016 Aceito: 01/09/2016

1 Introdução

A obesidade é uma doença de abrangência mundial podendo afetar tanto países desenvolvidos quanto subdesenvolvidos. Segundo a OMS (Organização Mundial da Saúde) a obesidade é definida como a excessiva concentração de gordura que pode prejudicar a saúde do indivíduo, sendo a falta de atividade física e o consumo exagerado de alimentos altamente energéticos dois dos principais fatores para o surgimento dessa morbidade. Além disso ela pode estar associada a outras doenças, a exemplo dos problemas respiratórios, problemas circulatórios, diabetes ou até mesmo o surgimento do câncer. Cabrera e Filho (2001) apresentam três medidas antropométricas distintas para avaliar a concentração e o volume de gordura no indivíduo, a saber, o IMC (Índice de Massa Corporal), o RCQ (Razão Cintura-Quadril) e o CA (Circunferência Abdominal). O IMC é definido como a razão entre o peso do indivíduo dado em quilogramas (kg) e sua altura ao quadrado (m^2). Dessa forma, Stol et al. (2011) apresentam três classificações para a obesidade: grau I com 30.0 $\leq$IMC$\leq$ 34.9 kg/m^2, grau II com 35.0 $\leq$IMC$\leq$ 39.9 kg/m^2 e grau III com IMC$\geq$ 40.0 kg/m^2. O RCQ é uma medida utilizada para verificar o risco do indivíduo apresentar doenças cardiovasculares, definida como a razão entre a circunferência da cintura e a circunferência do quadril. Assim, homens que apresentarem RCQ$\geq$0.9 e mulheres que apresentarem RCQ$\geq$0.85 estão mais sujeitas a tais riscos (Gabriele, 2011). O CA é também uma medida de risco para doenças cardiovasculares, de modo que homens e mulheres que apresentarem CA$>$ 0.94 cm e CA$>$ 0.80 cm, respectivamente, apresentam maiores riscos de adquirirem doenças cardíacas (Rezende et al., 2006).

Segundo estimativas da OMS, cerca de 13% da população mundial adulta, 11% dos homens e 15% das mulheres, eram obesos em 2014. Em 2008 verificou-se que a maior prevalência de pessoas com sobrepeso ou obesas encontravam-se na América, com cerca de 62% de pessoas com sobrepeso e 26% obesas, e a menor no Sudeste da Ásia, com cerca de 14% com excesso de peso e 3% obesas. Além disso é verificado que a obesidade é responsável por cerca de 2.8 milhões de mortes no mundo devido às suas complicações. O consumo de refrigerantes tem um papel fundamental para explicar o grande aumento da proporção de obesos ao redor do mundo. Basu et al. (2013), por exemplo, mostram através de um modelo de regressão que a variável consumo de refrigerantes está relacionada de forma significativa com o aumento do sobrepeso, obesidade e diabetes no mundo. Mostrando assim, que os cuidados com a alimentação são de fundamental importância para prevenir essas e outras doenças.

Os Estados Unidos são um dos países desenvolvidos que mais sofrem com os problemas relacionados a obesidade. De 2011 a 2012 cerca de um terço da população adulta era obesa, sendo maior entre negros não-hispânicos (47.8%), hispânicos (42.5%), brancos não-hispânicos (32.6%) do que entre os asiáticos não-hispânicos (10.8%), contudo não existia diferença entre a prevalência de homens e mulheres (Ogden et al., 2014). A obesidade é uma doença que pode ser prevenida ou tratada. O tratamento dessa doença e dos problemas relacionados a ela causam um grande impacto nos cofres públicos, que poderiam ser evitados se existissem políticas públicas no setor de saúde que conscientizassem as pessoas sobre os cuidados com a saúde, a importância de uma boa alimentação ou as vantagens de se praticar atividades físicas. Segundo dados de Arterburn et al. (2005) os gastos no setor de saúde pública relacionados a obesidade adulta nos Estados Unidos custaram cerca de 11 bilhões de dólares em 2000. Ou seja, o simples cuidado com a saúde poderia significar uma redução nos custos relacionados a medicamentos, internações hospitalares, atendimentos médicos, redução de doenças associadas e uma melhor qualidade de vida. Danaei et al. (2009) apontam que as mortes nos Estados Unidos no ano de 2005 relacionados a inatividade física, obesidade/sobrepeso e glicemia elevada mataram de 24% a 27% dos estadunidenses, já o tabagismo foi responsável por cerca de 16.7% a 20% das mortes adultas. Pietiläinen et al. (2008) apresentam em seu estudo que a inatividade física na adolescência se mostra como um fator de risco para a obesidade e a obesidade abdominal em adultos com 25 anos, ou seja, a falta de atividade física aumentava os riscos dessas condições em 3.9 e 4.8 vezes, respectivamente, constituindo um dos principais fatores para o aumento dessa morbidade ao longo dos anos.

Neste cenário, o nosso objetivo com o presente artigo é avaliar a proporção de adultos obesos nos Estados Unidos, uma vez que esse país está entre as nações desenvolvidas que mais sofrem com os problemas relacionados à obesidade. Para isso, utilizamos o modelo de regressão beta com dispersão variável (Simas et al., 2010). A classe de modelos de regressão beta tem como objetivo permitir a modelagem de respostas restritas ao intervalo $(0,1)$, por meio de uma estrutura de regressão que contém funções de ligação para modelar a média e a precisão, além de covariáveis e parâmetros de regressão desconhecidos. Os dados utilizados nesse artigo foram extraídos a partir de algumas fontes de informações públicas. O procedimento computacional foi desenvolvido utilizando o pacote `betareg` (Cribari-Neto e Zeileis, 2010) do *software* estatístico `R` (Kleiber e Zeleis, 2008; R Core Team, 2013).

O presente artigo encontra-se dividido em cinco seções. A Seção 2 apresenta o modelo de regressão beta com dispersão variável. Uma breve descrição dos dados encontra-se na Seção 3. Na Seção 4 foram apresentados

os resultados obtidos a partir do ajuste do modelo selecionado. Por último, na Seção 5 são apresentadas as conclusões e considerações finais.

2 Modelo de regressão beta

A classe de modelos de regressão beta é comumente utilizada em modelagens de variáveis que assumem valores no intervalo unitário $(0,1)$, a exemplo de taxas e proporções. Estes modelos são baseados na suposição de que a variável dependente tem distribuição beta e que a sua média é relacionada a um preditor linear por meio de uma função de ligação. O preditor linear envolve covariáveis e parâmetros de regressão desconhecidos. O modelo também inclui um parâmetro de dispersão, que em certas situações podem variar ao longo das observações (Smithson e Verkuilen, 2006; Espinheira et al., 2008a,b; Simas et al., 2010; Cribari-Neto e Souza, 2012, 2013; Silva e Souza, 2014; Souza e Cribari-Neto, 2015; Almeida Junior e Souza, 2015).

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

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

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

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

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

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

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

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

em que $\gamma = (\gamma_1,\ldots,\gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1},\ldots,z_{tq}$ são observações de q covariáveis ($k + q < n$), assumidas fixas e conhecidas, ϑ_t é o preditor linear, e $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável que mapeia os pontos positivos da reta, $h : (0,\infty) \to \mathbb{R}$. Portanto, $\phi_t = h^{-1}(\vartheta_t)$. Há várias possíveis escolhas para as funções de ligação $g(\cdot)$ e $h(\cdot)$. Para $g(\cdot)$ pode-se utilizar a função de ligação logit, $g(\mu) = \log\{\mu/(1-\mu)\}$, ou cloglog, $g(\mu) = \log\{-\log(1-\mu)\}$, entre outras. Já para $h(\cdot)$, pode-se utilizar a função *logarítmica*, $h(\phi) = \log(\phi)$, ou *raíz quadrada*, $h(\phi) = \sqrt{\phi}$, entre outras. Para maiores detalhes sobre as funções de ligação ver Mccullagh e Nelder (1989).

Segue de (1) que o logaritmo da função de verossimilhança é

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

em que

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

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

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

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

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 Descrição dos dados

A Tabela 1 apresenta uma breve descrição das variáveis utilizadas neste estudo. As fontes de dados consultadas foram as páginas da web: `http://stateofobesity.org`, `http://healthstats.azurewebsites.net/`, `http://www.gallup.com` e `http://map.feedingamerica.org`.

A Tabela 2 apresenta algumas estatísticas descritivas, como mínimo, primeiro quartil $(Q_{1/4})$, mediana, média, terceiro quartil $(Q_{3/4})$ e máximo das variáveis utilizadas. Essas estatísticas são baseadas em 50 observações referentes aos estados dos Estados Unidos. Algumas conclusões podem ser feitas através de sua análise. Para a variável proporção de adultos obesos, o valor máximo encontrado foi de 0.36, ou seja, para o estado correspondente a este valor, cerca de 36% dos adultos apresentaram obesidade no ano de 2014, enquanto que o valor mínimo encontrado para essa mesma variável foi de 0.21.

Para a variável porcentagem de adultos considerados inativos físicos, temos que 75% dos estados norte-americanos apresentaram uma porcentagem menor do que 25.23%. Considerando a porcentagem de indivíduos que consumiam vegetais menos de uma vez ao dia, 50% dos estados apresentaram um valor menor do que 22.90%. Para a variável FUM, que representa a porcentagem de adultos que fumavam cigarro no ano de 2012, o menor valor encontrado foi de 10.60%.

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

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

Destacamos que a maior proporção de adultos obesos no ano de 2014 foi registrada no estado do Arkansas, enquanto que a menor proporção foi encontrada no estado do Colorado. Para a variável inatividade física, os valores de mínimo e máximo foram encontrados em Colorado e Mississipi, respectivamente. Enquanto que Oregon apresentou a menor proporção de indivíduos que consumiam vegetais menos de uma vez por dia. Já o estado de Kentucky apresentou o maior percentual de adultos fumantes de cigarro no ano de 2012.

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

A Figura 1 apresenta o histograma e o *Box-plot*, respectivamente, da variável proporção de adultos obesos nos Estados Unidos. Dessa forma, é possível visualizar a distribuição dos dados e uma certa assimetria a esquerda, uma vez que a mediana está mais próxima do terceiro quartil. No *Box-plot* é possível destacar uma observação discrepante que excede seus limites, referente ao estado do Colorado. Portanto a utilização do modelo de regressão beta se faz necessário, visto que a variável resposta é uma proporção e foi verificada uma certa assimetria em sua distribuição.

4 Especificação do modelo

Inicialmente, ao ajustarmos o modelo de regressão beta, estamos interessados em testar a hipótese nula que a dispersão dos dados é fixa versus a hipótese alternativa de que a dispersão dos dados é variável. Para tanto, utilizamos o teste da razão de verossimilhanças (Almeida Junior e Souza, 2015; Neyman e Pearson, 1928; Silva e Souza, 2014) e obtivemos um p-valor<0.001 (valor obtido a partir dos dados amostrais e reflete a probabilidade de rejeitar a hipótese nula dado que ela é verdadeira), ou seja, ao nível de significância de 5% rejeitamos a hipótese nula de que a dispersão é fixa, portanto se faz necessário o ajuste de um modelo para a dispersão.

O modelo de regressão beta com dispersão variável selecionado foi:

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

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

A análise de diagnóstico é uma etapa da regressão que permite verificar algumas suposições do modelo, tais como: aleatoriedade dos resíduos, adequação da

Tabela 1: Descrição das variáveis utilizadas.

Variáveis	Definição	Fonte consultada
$OB2014$	Proporção de adultos obesos (2014)	http://stateofobesity.org
$INAT$	Porcentagem de inatividade física entre adultos (2014)	http://stateofobesity.org
$VEGET$	Porcentagem de adultos que consomem vegetais menos de uma vez por dia (2011)	http://stateofobesity.org
FUM	Porcentagem de fumantes de cigarro (2012)	http://healthstats.azurewebsites.net/
$DESEMP$	Porcentagem de residentes desempregados ou empregados em tempo parcial (2014)	http://www.gallup.com
$INSEG$	Taxa de insegurança alimentar (2013)	http://map.feedingamerica.org
BST	Escore de bem-estar (2014)	http://www.gallup.com
$DESCOB$	Porcentagem de residentes que não tem cobertura de seguro de saúde (2014)	http://www.gallup.com

Tabela 2: Estatística descritiva das variáveis utilizadas.

Variáveis	Mínimo	$Q_{1/4}$	Mediana	Média	$Q_{3/4}$	Máximo
$OB2014$	0.21	0.27	0.29	0.29	0.31	0.36
$INAT$	16.40	20.30	22.90	23.02	25.23	31.60
$VEGET$	15.30	20.70	22.90	23.21	25.85	32.50
FUM	10.60	17.32	19.55	19.84	22.38	28.30
$DESEMP$	9.00	13.35	15.20	15.07	16.88	20.70
$INSEG$	7.80	13.30	14.60	14.97	16.98	22.70
BST	59.00	61.10	61.90	61.86	62.62	64.70
$DESCOB$	4.60	10.15	12.75	12.73	15.30	24.40

distribuição de probabilidade suposta para a variável resposta e por fim, identificar possíveis pontos de influência e de alavanca por meio da análise dos resíduos. Neste artigo utilizamos os resíduos ponderados padronizados (para maiores detalhes ver Espinheira et al. (2008b)). Para verificarmos a qualidade do ajuste do modelo foi utilizado o coeficiente de determinação ajustado (pseudo-R^2), o teste *RESET* (Ramsey, 1969) e o gráfico de probabilidade normal com envelope simulado.

O pseudo-R^2 é uma medida global da variação explicada e análoga ao coeficiente de determinação, utilizada em modelos lineares de regressão. Ferrari e Cribari-Neto (2004) propuseram um pseudo-R^2 para os modelos de regressão beta, definido como o quadrado do coeficiente de correlação entre $\hat{\eta}$ e $g(y)$. Dessa forma com um pseudo-R^2=0.75 constatamos que as variáveis independentes são capazes de explicar cerca de 75% da variabilidade total da proporção de adultos obesos. Para testar a correta especificação do modelo, utilizamos o teste *RESET* para modelos de regressão beta (Lima, 2007; Souza e Cribari-Neto, 2015). A hipótese nula deste teste sugere que o modelo proposto está bem especificado contra a hipótese alternativa de que o modelo está mal especificado. O teste consiste em adicionar como variável de teste o preditor linear estimado elevado a segunda potência ($\hat{\eta}^2$) ao submodelo da média. Desta forma, obtivemos um p-valor=0.208, ou seja, como a variável de teste não mostrou-se significativa, podemos concluir que o modelo proposto não apresenta nenhum erro de especificação ao nível de significância de 5%.

O gráfico de probabilidade normal com envelope simulado é uma técnica gráfica que permite identificar possíveis observações discrepantes, bem como a adequação da distribuição de probabilidade que foi suposta para o modelo. Na Figura 2 notamos que as observações encontram-se distribuídas de forma aleatória dentro do envelope e próximo a linha central, não sendo possível detectar observações discrepantes. Portanto, não temos evidências de que o modelo especificado não está adequado. Já a Figura 3 mostra o gráfico dos resíduos ponderados padronizados versus a ordem das observações, sendo possível visualizar que os resíduos estão distribuídos de forma aleatória entre os limites $[-2,2]$ com apenas as observações 2, 6 e 11 ultrapassando esse intervalo, correspondendo aos estados do Alaska, Colorado e Hawaii. Portanto, temos que a suposição de que os resíduos são uma sequência aleatória não foi violada.

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 et al. (2008a) propuseram uma medida similar a distância de Cook e medidas de influência

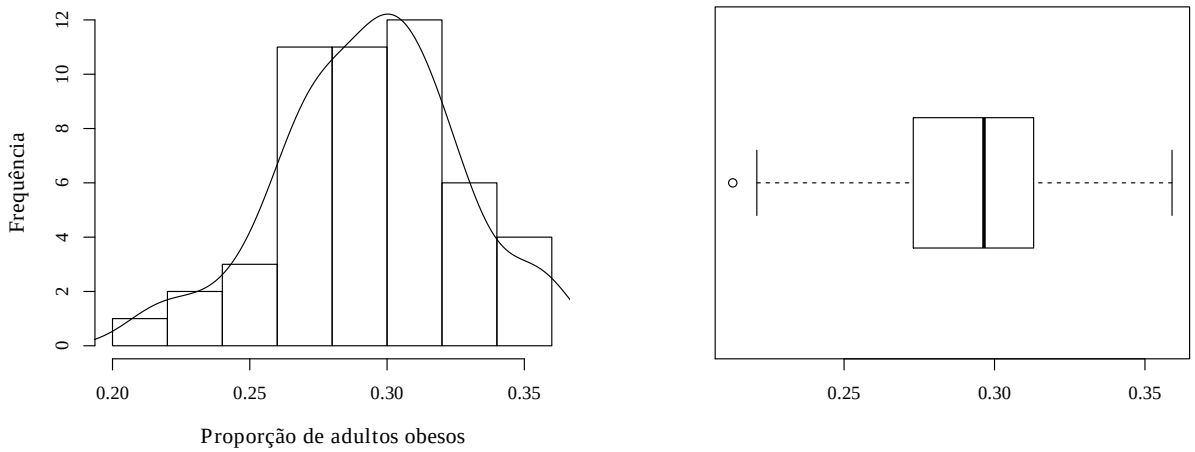

Figura 1: Histograma e *Box-plot* da variável proporção de adultos obesos nos estados dos Estados Unidos em 2014.

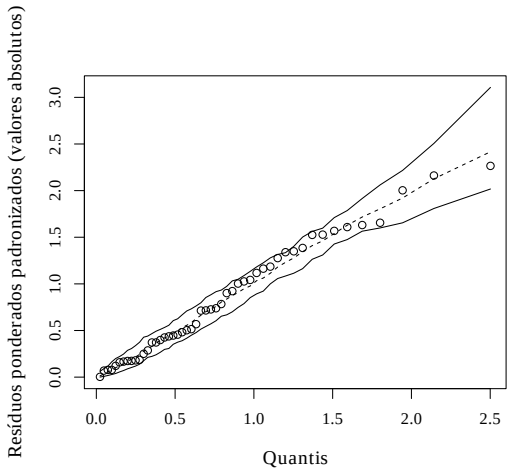

Figura 2: Gráfico da probabilidade normal com envelopes simulados.

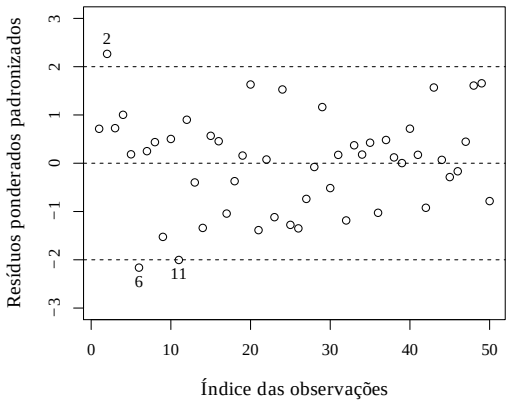

Figura 3: Resíduos ponderados padronizados versus os índices das observações.

local para modelos de regressão beta. A alavancagem generalizada proposta por Wei et al. (1998) é definida em modelos de regressão como uma medida da importância individual das observações. Ferrari et al. (2011) propuseram a alavancagem generalizada para modelos de regressão beta com dispersão variável.

A Figura 4 apresenta o gráfico das distâncias de Cook versus os valores preditos e da alavancagem generalizada versus os valores preditos. Essas técnicas gráficas permitem identificar observações que interferem nas estimativas dos parâmetros produzindo resultados distorcidos. No gráfico da distância de Cook não foi possível identificar nenhum ponto de influência. Entretanto algumas observações apresentam valores de Cook diferenciados, a saber, as observações 2, 24 e 43 referentes, respectivamente, aos estados do Alaska, Mississippi e Texas. No gráfico da alavancagem generalizada é possível identificar três pontos de alavanca referentes aos estados da Louisiana, North Dakota e South Dakota.

A Tabela 3 apresenta os resultados obtidos a partir do ajuste do modelo de regressão beta com dispersão variável. Esse modelo utiliza as funções de ligação cloglog e log para modelar a média e a precisão, respectivamente, uma vez que estas forneceram um melhor ajuste. Através do teste de Wald (Wald, 1943; Cribari-Neto e Zeileis, 2010) verificamos que as variáveis relevantes para explicar a proporção de adultos obesos na modelagem da média foram: $INAT$, $VEGET$, FUM, $DESEMP$, $INSEG$, BST e a interação entre as variáveis taxas de fumantes e de insegurança alimentar, denotada por $INTER$, pois

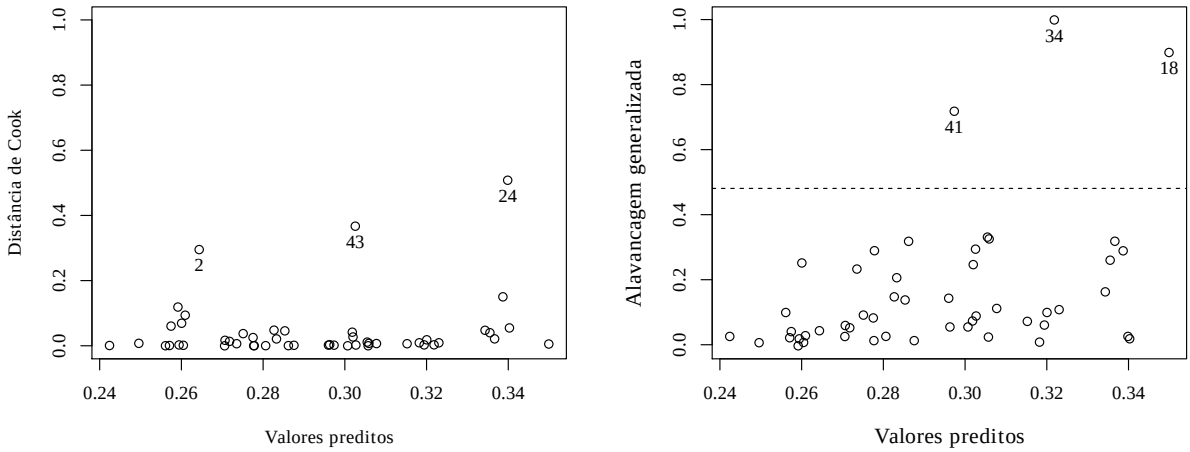

Figura 4: Gráfico da distância de Cook e da alavancagem generalizada.

apresentaram p-valores menores que o nível de significância de 5%, rejeitando assim a hipótese nula de que $\beta_j = 0$. Em relação a modelagem da precisão verificamos que as variáveis $INSEG$, $DESCOB$ e $VEGET$ foram significativas, pois apresentaram p-valores menores que o nível de significância de 5%. Através da análise dos coeficientes do modelo proposto, é possível verificar que as variáveis $INAT$, $VEGET$, FUM e $INSEG$ apresentaram efeito positivo na proporção de adultos obesos, por outro lado as variáveis $DESEMP$ e BST influenciaram negativamente esta mesma variável. Considerando a estrutura de regressão para o parâmetro de precisão, temos que à medida que a covariável $INSEG$ aumenta a precisão diminui, ou seja, os estados que apresentam maiores valores de insegurança alimentar tendem a apresentar respostas menos precisas. Por outro lado, à medida que as covariáveis $DESCOB$ e $VEGET$ aumentam, a precisão aumenta, ou seja, os estados com maiores valores de descobertos e consumo de vegetais tendem a apresentar respostas mais precisas.

Os resultados obtidos na análise inicial da regressão beta são favoráveis com alguns resultados encontrados na literatura. Primeiro, no estudo de Cavalcanti et al. (2010) por meio da regressão logística mostrou-se que a prática de atividade física é um fator de proteção para a obesidade abdominal em adolescentes de 14 a 19 anos de idade, independente da presença do excesso de peso. Segundo, no estudo de Castanho et al. (2013) por meio da regressão logística verificou-se também que o consumo de frutas de maneira adequada reduz as chances de se adquirir obesidade abdominal. Ainda a ingestão de frutas, verduras e legumes apresentou um efeito significativo reduzindo o risco de se adquirir doenças cardiovasculares. Terceiro, no estudo de Flegal (2007) concluiu-se que grandes mudanças na prevalência de tabagismo causavam um pequeno efeito na prevalência da obesidade, ocasionando um aumento muitas vezes menor do que um ponto percentual na prevalência de obesos. Quarto, no estudo de Zhang et al. (2014) verificou-se uma associação entre as taxas de desemprego e o peso dos indivíduos nos estados e cidades. Concluindo a partir da regressão logística que as taxas de desemprego nos estados estão associadas negativamente com o IMC individual ao longo dos anos. Quinto, Dharod et al. (2013) por meio de um estudo transversal analisaram a associação entre insegurança alimentar, consumo alimentar e índice de massa corporal entre mulheres refugiadas Somalis que viviam nos Estados Unidos no período de outubro de 2006 até dezembro de 2007. Verificou-se que a insegurança alimentar estava associada positivamente ao sobrepeso e a obesidade, ou seja, ela apresentava-se como um fator de risco aumentando as chances do indivíduo vir a ser obeso. Sexto, Jagielski et al. (2014) exploraram a associação entre adiposidade, qualidade de vida e bem-estar mental entre indivíduos obesos que entraram em um serviço de gestão de peso. Verificando-se que a adiposidade está relacionada negativamente com o bem-estar e a qualidade de vida, sendo identificada uma alta prevalência de comorbidades psicológicas e uma redução da qualidade de vida dos indivíduos obesos. Segundo Holben (2010) a insegurança alimentar é uma ameaça a saúde que pode ser evitada. Além disso, ela está relacionada a diabetes, a incidência e o risco de doenças crônicas, ao excesso de peso e a obesidade. Entretanto é necessário maiores estudos para direcionar a relação entre a insegurança alimentar e o excesso de peso ou obesidade nos adultos.

Dando continuidade a análise de regressão do modelo ajustado, com o objetivo de verificarmos o impacto que as observações de alta alavancagem podem causar

Tabela 3: Estimativa dos coeficientes, erro padrão e p-valor do modelo de regressão beta com dispersão variável, por meio das funções de ligação cloglog e log para modelar a média e a precisão, respectivamente.

Função de Ligação	Variáveis	Parâmetros	Estimativa	Erro padrão	p-valor
cloglog(μ)	$INTERCEPTO$	β_0	-1.4737	0.4859	0.0024
	$INAT$	β_1	0.0127	0.0042	0.0027
	$VEGET$	β_2	0.0117	0.0024	<0.0001
	FUM	β_3	0.0556	0.0164	0.0007
	$DESEMP$	β_4	-0.0144	0.0028	<0.0001
	$INSEG$	β_5	0.0697	0.0251	0.0054
	BST	β_6	-0.0180	0.0061	0.0034
	$INTER$	β_7	-0.0032	0.0011	0.0047
log(ϕ)	$INTERCEPTO$	γ_0	3.6794	1.5563	0.0181
	$INSEG$	γ_1	-0.5097	0.0942	<0.0000
	$DESCOB$	γ_2	0.1966	0.0616	0.0014
	$VEGET$	γ_3	0.3704	0.0585	<0.0000

na estimativa dos parâmetros, decidimos por excluí-las individualmente e conjuntamente, sendo assim as variações percentuais das estimativas devido as observações 18, 34 e 41 podem ser visualizadas na Tabela 4. A partir da análise descritiva das variáveis relevantes ao modelo foi possível verificar que o estado da Louisiana, observação 18, se destaca por apresentar a maior taxa de $VEGET$ e uma das menores taxas de BST. Ao excluirmos a observação 18, temos que a variação percentual de $\hat{\beta}_2$ é 25.10%. Em relação ao submodelo da precisão, temos que o maior impacto ocorreu no intercepto, $\hat{\gamma}_0$, com uma redução de -37.49%, diminuindo a precisão das respostas.

O estado North Dakota, observação 34, apresenta a menor taxa de $DESEMP$ e $INSEG$, sendo assim a exclusão dessa observação diminui de forma considerável as estimativas de $\hat{\beta}_6$ e $\hat{\beta}_7$ respectivamente em -51.17% e -39.64%. Em relação ao submodelo da precisão, a exclusão da observação 34, diminui as estimativas de $\hat{\gamma}_0$ e $\hat{\gamma}_1$ respectivamente em -76.35% e -52.55%, influenciando negativamente a precisão.

O estado South Dakota, observação 41, se destaca por apresentar uma das menores taxas de $INAT$, além das maiores taxas de $DESEMP$ e $DESCOB$. A exclusão da observação 41 não causa grandes variações nas estimativas dos parâmetros. O maior impacto ocorreu na estimativa de $\hat{\beta}_6$, aumentando seu valor em 12.71%.

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

Tabela 4: Variações percentuais nas estimativas dos parâmetros ao se retirar observações influentes. Proporção de adultos obesos nos Estados Unidos.

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

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

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

em que

$$\mu_t = g^{-1}(\beta_0 + \beta_1 INAT_t + \beta_2 VEGETt + \beta_3 FUM_t + \beta_4 DESEMP_t + \beta_5 INSEG_t + \beta_6 BST_t + \beta_7 INTER_t).$$

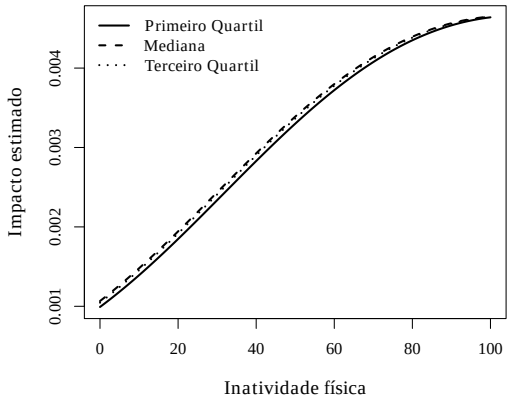

Figura 5: Impacto da taxa de adultos fisicamente inativos sobre a proporção de adultos obesos fixando-se a demais variáveis no primeiro, segundo e terceiro quartis.

Considerando que a função de ligação para a média é cloglog, o impacto pode ser expresso como:

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

Dessa forma, considerou-se o cenário em que as variáveis $VEGET$, FUM, $DESEMP$, $INSEG$, BST e $INTER$ estão fixadas no primeiro, segundo e terceiro quartis. A Figura 5 apresenta o impacto exercido pela inatividade física sobre a proporção de adultos obesos nos diferentes estados. Neste cenário é possível notar que o impacto da variável $INAT$ é positivo e assume forma acelerada para valores menores que 0.85. Para valores de $INAT > 0.85$ o efeito continua sendo positivo, contudo apresentando um crescimento mais lento. Notamos também que as curvas de impacto para o segundo e terceiro quartis não apresentam grandes diferenças, contudo quando as covariáveis estão fixadas no primeiro quartil a proporção de obesos é relativamente menor que nos demais quartis.

5 Conclusões

Neste artigo, através do modelo de regressão beta com dispersão variável, modelamos a proporção de adultos obesos e verificamos através do teste de Wald que as variáveis inatividade física, consumo de vegetais menos de uma vez por dia, o hábito de fumar e insegurança alimentar estavam relacionadas de forma positiva com a proporção média de adultos obesos. Já as taxas de desempregados e escore de bem-estar mostraram-se relacionados negativamente com a resposta média. A modelagem do submodelo da precisão permitiu constatar que os estados que apresentavam maiores taxas de descobertos quanto ao seguro de saúde e maiores taxas de consumo de vegetais, apresentavam respostas mais precisas, ou seja, menos dispersas.

Estimamos o impacto exercido pela taxa de inatividade física sobre a proporção média de adultos obesos nos Estados Unidos. Os resultados revelaram que o efeito desse impacto é positivo e apresenta uma forma acelerada para valores de inatividade física menores do que 0.85, considerando que as demais variáveis foram fixadas em um determinado quartil.

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

Agradecimentos

Agradecemos ao CNPq e à Capes pelo apoio financeiro.

Referências

Almeida Junior, P., Souza, T. (2015). Estimativas de votos da presidente Dilma Roussef nas eleições presidenciais de 2010 sob o âmbito do bolsa família. *Ciência e Natura*, *37*(1), 12–22.

Arterburn, D., Maciejewski, M., Tsevat, J. (2005). Impact of morbid obesity on medical expenditures in adults. *International Journal of Obesity*, *29*(3), 334–339.

Basu, S., Mckee, M., Galea, G., Stuckler, D. (2013). Relationship of soft drink consumption to global overweight, obesity, and diabetes: a cross-national analysis of 75 countries. *American Journal of Public Health*, *103*(11), 2017–2077.

Cabrera, M., Filho, W. (2001). Obesidade em idosos: prevalência, distribuição e associação com hábitos e

co-morbidades. *Arquivos brasileiros de Endocrinologia e Metabologia*, *45*(5), 494–501.

Castanho, G. K. F., Marsola, F. C., Mclellan, K. C. P., Nicola, M., Moreto, F., Burini, R. C. (2013). Consumo de frutas, verduras e legumes associado à síndrome metabólica e seus componentes em amostra populacional adulta. *Ciência & Saúde Coletiva*, *18*(2), 385–392.

Cavalcanti, C. B. S., Barros, M. V. G., Menêses, A. L., Santos, C. M., Azevedo, A. M. P., Guimarães, F. J. d. S. P. (2010). Obesidade abdominal em adolescentes: prevalência e associação com atividade física e hábitos alimentares. *Arquivos Brasileiros de Cardiologia*, *94*(3), 371–377.

Cook, R. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Cribari-Neto, F., Souza, T. (2012). Testing inference in variable dispersion beta regressions. *Journal os Statistical Computation and Simulation*, *82*(12), 1827–1843.

Cribari-Neto, F., Souza, T. (2013). Religious belief and intelligence: Worldwide evidence. *Intelligence*, *41*(5), 482–489.

Cribari-Neto, F., Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, *34*(2), 1–24.

Danaei, G., Ding, E., Mozaffarian, D., Taylor, B., Rehm, J., Murray, C., Ezzati, M. (2009). The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLOS Medice*, *6*(4), 1–23.

Dharod, J. M., Croom, J. E., Sady, C. G. (2013). Food insecurity: its relationship to dietary intake and body weight among somali refugee women in the United States. *Journal of Nutrition Education and Behavior*, *45*(1), 47–53.

Espinheira, P., Ferrari, S., Cribari-Neto, F. (2008a). Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, *52*(9), 4417–4431.

Espinheira, P., Ferrari, S., Cribari-Neto, F. (2008b). On beta regression residuals. *Journal of Applied Statistics*, *35*(4), 407–419.

Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal os Applied Statistics*, *31*(7), 799–815.

Ferrari, S., Espinheira, P., Cribari-Neto, F. (2011). Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, *65*(3), 337–351.

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

Gabriele, R. (2011). Índice de massa corporal no diagnóstico de transtornos nutricionais em idosos institucionalizados no município de Fortaleza, Ceará. Dissertação de Mestrado, Universidade de Fortaleza.

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

Jagielski, A., Brown, A., Hosseini-Araghi, M., Thomas, N., Taheri, S. (2014). The association between adiposity, mental well-being, and quality of life in extreme obesity. *PLOS One*, *9*(3), 1–8.

Kleiber, C., Zeleis, A. (2008). *Applied econometrics with R*. New York: Springer.

Lima, L. (2007). Um teste de especificação correta para modelos de regressão beta. Dissertação de Mestrado, Universidade Federal de Pernambuco.

Mccullagh, P., Nelder, J. (1989). *Generalized linear models*. London: Chapman and Hall.

Neyman, J., Pearson, E. (1928). On the use and interpretation of certain teste criteria for purposes of statistical inference. *Biometrika*, *20*, 175–240.

Ogden, C., Carroll, M., Kit, B., Flegal, K. (2014). Prevalence of obesity among adults: United States, 2011-2012. *Medical Benefits*, *31*(1), 9.

Pietiläinen, K., Kaprio, J., Borg, P., Plasqui, G., Ykijärvinen, H., Kujala, U., Rose, R., Westerterp, K., Rissanen, A. (2008). Physical inactivity and obesity: A vicious circle. *Obesity (Silver Spring)*, *16*(2), 409–414.

Press, W., Teukolsky, S., Vetterling, W., Flannery, B. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge University Press.

R Core Team (2013). *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`.

Ramsey, J. B. (1969). Tests for specification erros in classical linear least squares regression analysis. *Journal of the Royal Statistical Society*, *31*(2), 350–371.

Rezende, F. A. C., Rosado, L. E. F. P., Ribeiro, R. C. L., Vidigal, F. C., Vasques, A. C. J., Bonard, I. S., Carvalho, C. R. (2006). Índice de massa corporal e circunferência abdominal: associação com fatores de risco cardiovascular. *Arquivos Brasileiros de Cardiologia*, *87*(6), 728–734.

Silva, C., Souza, T. (2014). Modelagem da taxa de analfabetismo no estado da Paraíba via modelo de regressão beta. *Revista Brasileira de Biometria*, *32*(3), 345–359.

Simas, A., Barreto-Souza, W., Rocha, A. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, *54*(2), 348–366.

Smithson, M., Verkuilen, J. (2006). A better lemon-squeezer? Maximum likelihood regression with beta-distribuited dependent variables. *Psychological Methods*, *11*(1), 54–71.

Souza, T., Cribari-Neto, F. (2015). Intelligence, religiosity and homosexuality non-acceptance: Empirical evidence. *Intelligence (Norwood)*, *52*, 63–70.

Stol, A., Gugelmin, G., Lampa-Junior, V. M., Frigulha, C., Selbach, R. A. (2011). Complicações e óbitos nas operações para tratar a obesidade mórbida. *Arquivos Brasileiros de Cirurgia Digestiva*, *24*(4), 282–284.

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

Wei, B., Hu, Y., Fung, W. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, *25*(1), 25–37.

Zhang, Q., Lamichhane, R., Wang, Y. (2014). Associations between U.S. adult obesity and state and county economic conditions in the recession. *Journal of Clinical Medicine*, *3*(1), 153–166.