Universidade Estadual de Maringá
Departamento de Estatística
Disciplina: 8077 - Modelos Lineares Generalizados
Professor: Dra. Rosangela Getirana Santana
Acadêmico: André Felipe Berdusco Menezes

Modelos de regressão: aplicações em dados de contagem e proporções

Maringá
Julho de 2017

Sumário

1 Introdução

É de interesse comum, nos mais diversos âmbitos do conhecimento, compreender a relação entre uma variável resposta e possíveis variáveis explicativas (covariáveis), e na sequência realizar predições a partir da relação estabelecida. Seguramente, como aponta diversos autores, McCulagh e Nelder (1983), Neter et al. (1996), McCulloch e Searle (2001), Dobson (2002), Weisberg (2005) e Hosmer, Lemeshow e Sturdivant (2013) os modelos de regressão tem sido exaustivamente aplicados para estes propósitos. Além disso, o modelo mais conhecido e, sem dúvida, o mais utilizado por usuários da estatística aplicada é o modelo linear geral.

Um importante avanço na modelagem estatística, baseada modelos de regressão, ocorreu no início da década de 70 com a criação dos modelos lineares generalizados (MLG's), cujo o marco inicial é o artigo de Nelder e Wedderburn (1972). Os autores estendem de uma forma mais geral a classe de modelos de regressão, acomodando sob a mesma abordagem outras distribuições agrupadas na família exponencial.

Neste trabalho iremos analisar dois conjuntos de dados em que as variáveis respostas são, em sua natureza, discreta proveniente de uma determinada contagem e, por ultimo contínua restrita ao intervalo $(0, 1)$. O primeiro conjunto de dados foi retirado de Long (1990) e esta relacionado ao número de publicações dos Ph.D. em Bioquímica. Por outro lado, o segundo conjunto de dados refere-se a proporção de votos válidos da ex-presidenta Dilma no segundo turno da eleição de 2010 considerando os municípios do estado do Paraná. Nesta análise o objetivo é avaliar quais foram os principais fatores que influenciaram a proporção de votos obtidos por Dilma. Vale ressaltar, que um estudo semelhante foi realizado por Kieschnick e McCullough (2003) considerando a proporção de votos que o ex-presidente George Bush recebeu nas eleições presidenciais do Estados Unidos no ano de 2000.

Para melhor sistematização e organização este trabalho foi dividido em cinco seções, sendo a primeira destinada a uma revisão dos modelos utilizados para analisar os dois conjuntos de dados. Na Seção 3 são apresentados os conjuntos de dados. Os resultados e discussões das aplicações são expostos na Seção 4. Finalmente, na Seção 5 aponto algumas considerações sobre o trabalho e seus resultados.

2 Modelos Lineares Generalizados

No início da década de 70 Nelder e Wedderburn (1972) unificaram uma classe de modelos de regressão e denominaram de modelos lineares generalizados (MLG's). De forma geral, como discute Paula (2013) a ideia consistiu em abrir o leque de opções para a distribuição da variável resposta, no entanto restringindo com que a mesma pertença à família exponencial de distribuições, além disso, dar maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear.

Considere uma variável aleatória Y e associada a ela um conjunto de covariáveis $x_1, \ldots, x_p$. Assim, dada uma amostra aleatória de n observações $(y_i, \boldsymbol{x}_i)$, em que $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, definimos os MLG's por três componentes:

(i) **Componente aleatório:** representado por um conjunto de variáveis aleatórias independentes $Y_1, \ldots, Y_n$ provenientes de uma mesma distribuição da família exponencial, com médias $\mu_i, i = 1, \ldots, n$. Genericamente, sua função densidade de probabilidade é expressa por:

$$f(y_i \mid \theta_i, \phi) = \exp\left\{\frac{1}{a(\phi)}\,[y_i\,\phi + b(\theta_i)] + c(y_i, \phi)\right\} \tag{1}$$

sendo $a(\cdot), b(\cdot)$ e $c(\cdot)$ funções conhecidas, ϕ o parâmetro de dispersão e θ_i o parâmetro canônico. De (1) temos que:

$$E\,(Y_i) = b'(\theta_i) = \mu_i \quad \text{e} \quad V\,(Y_i) = \phi\, b''(\theta_i) = \phi\, V(\mu_i) \tag{2}$$

em que $V(\mu_i) = \mathrm{d}\mu_i/\mathrm{d}\theta_i$ é denominada de função de variância que depende unicamente da média μ_i.

(ii) **Componente sistemático:** as variáveis explicativas entram na forma de uma soma linear de seus efeitos, isto é,

$$\eta_i = \sum_{j=1}^{p} x_{ij}\,\beta_j = \boldsymbol{x}_i^\top\,\boldsymbol{\beta} \quad \text{ou} \quad \boldsymbol{\eta} = \boldsymbol{X}\,\boldsymbol{\beta} \tag{3}$$

sendo $\boldsymbol{X}$ a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ o vetor de parâmetros desconhecidos e $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ o preditor linear.

(iii) **Função de ligação:** uma função que relaciona o componente aleatório ao componente sistemático, em outras palavras vincula a sua média ao preditor linear, isto é,

$$\eta_i = g(\mu_i) \tag{4}$$

sendo $g(\cdot)$ uma função monótona diferenciável.

A partir destas especificações e utilizando a função de verossimilhança de (1) Nelder e Wedderburn (1972) desenvolveram um método genérico para inferência sobre os parâmetro $\boldsymbol{\beta}$, conforme a escolha da distribuição dos Y_i e função de ligação $g(\mu_i)$.

Nas próximas seções iremos apresentar alguns modelos pertencentes a teoria dos MLG's, como o caso do modelo Poisson e Binomial Negativa. Também será abordado alguns modelos cuja as distribuições de probabilidade não pertencem a família exponencial, entretanto a construção do modelo segue de forma a análoga as componentes dos MLG's.

2.1 Modelos para dados de contagem

Podemos considerar dados de contagens como variáveis aleatórias que assumem valores inteiros não negativos. Na análise deste tipo especifico de dados o uso do modelo de regressão Poisson tem ocorrência predominante. Entretanto, uma característica peculiar da distribuição Poisson é que sua média e variância são iguais, e na prática quando isso não é observado têm-se um problema de superdispersão ou subdispersão, inviabilizando o uso deste modelo

Nesse sentido, diversas alternativas foram e tem sido desenvolvidas na literatura estatística, talvez a mais comum seja o modelo Binomial Negativa. Na literatura brasileira, um trabalho recente que vale ressaltar foi o de Junior (2016), no qual apresenta-se uma excelente revisão de vários modelos alternativos a Poisson, com um enfoque especial a distribuição COM-Poisson. Vale destacar ainda, que o autor considerou o modelo COM-Poisson sobre diferentes perspectivas, isto é, excesso de zeros, super e subdispersão e ainda na presença de efeitos aleatórios.

2.1.1 Modelo Poisson

Dizemos que uma variável aleatória discreta de Poisson, se sua função massa de probabilidade é expressa por:

$$P(Y = y \mid \mu) = \frac{\mu^y \, e^{-\mu}}{y!}, \qquad y = 0, 1, 2, \ldots \tag{5}$$

em que $\mu > 0$ representa a taxa de ocorrência do evento. Uma característica importante desta distribuição é que $E(Y) = V(Y) = \mu$. Essa propriedade é denominada de equidispersão.

Afim de estabelecer um modelo de regressão por meio da distribuição de Poisson, (5), considere que $Y_1, \ldots, Y_n$ são variáveis aleatórias condicionalmente independentes, dado um vetor de covariáveis $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. A partir da teoria dos MLG's, define-se então o modelo de regressão log-linear Poisson por:

$$Y_i \mid \boldsymbol{x}_i \sim \text{Poisson}(\mu_i) \quad \text{com} \quad \log(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{6}$$

em que $\boldsymbol{x}_i$ e $\boldsymbol{\beta}$ são vetores desconhecidos de dimensões $(p \times 1)$, representando respectivamente, as covariáveis e seus parâmetros fixos e desconhecidos. Note que a função de ligação canônica da Poisson foi empregada, isto é, $g(\mu_i) = \log(\mu_i)$.

Realizando as substituições de (6) em (5), pode-se mostrar que a função log-verossimilhança para $\boldsymbol{\beta}$ é dada por:

$$\ell(\boldsymbol{\beta} \mid y_i, \boldsymbol{x}_i) = \sum_{i=1}^{n} y_i \, \boldsymbol{x}_i^\top \, \boldsymbol{\beta} - \exp\left(\boldsymbol{x}_i \, \boldsymbol{\beta}\right) - \log(y_i!) \tag{7}$$

Maximizando a função (7) em relação a $\boldsymbol{\beta}$ obtemos as estimativas de máxima verossimilhança, $\widehat{\boldsymbol{\beta}}$, de $\boldsymbol{\beta}$. Para as distribuições pertencentes a família exponencial, Nelder e Wedderburn (1972) propuseram o algoritmo de mínimos quadrados ponderados ou também conhecido como método de escore de Fisher.

2.1.2 Método da Quase-Verossimilhança

Uma opção para contornar a propriedade de equidispersão da Poisson foi proposta por Wedderburn (1974). O autor propõe uma nova forma de estimação com base em uma função biparamétrica, denominada de função quase-verossimilhança.

Sejam $Y_1, \ldots, Y_n$ variáveis aleatórias independentes com $E(Y_i) = \mu_i$ e $\text{Var}(Y_i) = \sigma^2 V(\mu_i)$, em que $V(\mu_i)$ é uma função conhecida. Então, sob certas condições de regularidades, a função de quase-verossimilhança é definida como:

$$Q(\mu_i \mid y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V(t)} \, \mathrm{d}t. \tag{8}$$

Note que a estimação por quase-verossimilhança necessita apenas das especificações do primeiro e segundo momentos, isto é, $E(Y_i)$ e $V(Y_i)$. A estimação dos parâmetros, é obtida maximizando a expressão (8) em relação a μ_i, lembrando que nos modelos de regressão estamos interessados em $g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

Uma importante propriedade mostrada por Wedderburn (1974) é a seguinte

$$-E\left(\frac{\partial^2}{\partial \mu_i^2} Q(\mu_i \mid y_i)\right) \leq -E\left(\frac{\partial^2}{\partial \, \mu_i^2} \ell(\mu_i \mid y_i)\right) \tag{9}$$

Essa relação mostra que a informação a respeito de μ_i quando se conhece apenas a relação entre a variância e a média é menor ou igual do que a informação a respeito de μ quando se conhece a distribuição da variável resposta, a qual é dada pelo log-verossimilhança. Na prática isso influencia na estimação dos erros padrão, podendo assim, ocasionar estimativas menos precisas do que a abordagem da verossimilhança.

Ressalta-se que o parâmetro de dispersão σ^2 é estimado a partir do método dos momentos (PAULA, 2013), o qual possui expressão analítica dada por:

$$\widehat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \tag{10}$$

Portanto verifica-se que o método da quase-verossimilhança permite estabelecer uma relação entre a média e a variância, corrigindo a dispersão da variável resposta. No entanto, devido a propriedade (9) este método pode produzir estimativas menos precisas do que a abordagem baseada na verossimilhança.

2.1.3 Modelo Binomial Negativo

A distribuição Binomial Negativa é uma escolha natural para análises de dados de contagem que apresentam problemas de superdispersão. Sua função massa de probabilidade pode surge de um processo hierárquico de efeitos aleatórios onde se assume

$$\begin{aligned} Y \mid \lambda &\sim \text{Poisson} \\ \lambda &\sim \text{Gama}(\mu, \theta) \end{aligned} \tag{11}$$

ou seja, a distribuição binomial negativa pode ser vista como uma distribuição de Poisson(λ), em que λ é uma variável aleatória com distribuição Gama(μ, θ). Formalmente, considere que $f(y \mid \lambda)$ e $g(\lambda \mid \mu, \theta)$ denotam a função massa de probabilidade da Poisson (5) e a função densidade de probabilidade da Gamma, respectivamente. Logo têm-se:

$$\begin{aligned} P(Y = y \mid \mu, \theta) &= \int_0^\infty f(y \mid \lambda) \, g(\lambda \mid \mu, \theta) \, \mathrm{d}\lambda \\ &= \frac{\Gamma(\theta + y)}{\Gamma(y+1)\,\Gamma(\theta)} \left(\frac{\mu}{\mu + \theta}\right)^y \left(\frac{\theta}{\mu + \theta}\right)^\theta, \qquad y = 0, 1, 2, \ldots \end{aligned} \tag{12}$$

em que $\mu, \theta > 0$. Sua média e variância são definidas respectivamente por, $E(Y) = \mu$ e $\text{Var}(Y) = \mu + \mu^2\,\theta^{-1}$.

Assumindo $\theta = \phi^{-1}$ obtemos uma nova parametrização da Binomial Negativa com média $E(Y) = \mu$ e variância $\text{Var}(Y) = \mu\,(1 + \phi\,\mu)$. Conforme aponta Lawless (1987) ϕ é um parâmetro de dispersão. Além disso, se ϕ for igual a zero, então a média e variância são iguais, resultando na distribuição de Poisson. Por outro lado, se $\phi > 0$, então a variância é maior que a média, caracterizando dados com superdispersão. Ressalta-se, que esta parametrização da Binomial Negativa é utilizada pelo `PROC GENMOD` (SAS Institute Inc., 2011).

De modo análogo a Poisson, o emprego da Binomial Negativa na regressão é realizado definindo por:

$$Y_i \mid \boldsymbol{x}_i \sim \text{BN}(\mu_i, \phi) \quad \text{com} \quad g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{13}$$

em que $\boldsymbol{x}_i$ e $\boldsymbol{\beta}$ são vetores desconhecidos de dimensões $(p \times 1)$, representando respectivamente, as covariáveis e seus parâmetros fixos e desconhecidos. Além disso, $g(\mu_i)$ denota uma função de ligação, usualmente $g(\mu_i) = \log(\mu_i)$.

2.2 Modelos para dados no intervalo unitário

A teoria dos MLG's unificou uma classe ampla de distribuições para aplicação em modelos de regressão. Contudo, os modelos lineares generalizados usuais apresentam fortes limitações para variáveis respostas cujo o suporte é limitado a um intervalo (a, b), sendo o intervalo unitário $(0, 1)$ o mais habitual. Neste contexto diversos autores vem propondo alternativas para este tipo especifico de dados. Destaca-se o trabalho de Kieschnick e McCullough (2003) no qual foi considerado o modelo linear Gaussiano com censura, o modelo de regressão Beta, simplex e um modelo semi-paramétrico estimado por quase-verossimilhança. Na sequência e independentemente, Ferrari e Cribari-Neto (2004) apresentaram uma definição mais formal para modelo de regressão Beta a partir da reparametrização da densidade Beta. Por fim, ressalta-se que uma comparação por meio de aplicações dos modelos linear Gaussino, Beta, Simplex e Kumaraswamy foi recentemente realizada por Bonat, JR e Zeviani (2013).

2.2.1 Modelo Beta

O modelo de regressão Beta introduzido por Ferrari e Cribari-Neto (2004) consite em uma nova parametrização da distribuição Beta indexada por sua média e um parâmetro de precisão. A função densidade de probabilidade da distribuição Beta em sua forma reparametrizada é definida por:

$$f(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi - 1}(1-y)^{(1-\mu)\phi - 1}, \qquad 0 < y < 1 \tag{14}$$

em que $0 < \mu < 1$ e $\phi > 0$. Assim, a média e variância de Y são dadas pelas seguintes expressões,

$$\mathbb{E}(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \frac{\text{V}(\mu)}{(1+\phi)} \tag{15}$$

sendo $V(\mu) = \mu\,(1-\mu)$ a função variância.

Para estabelecer um modelo de regressão utilização a distribuição Beta, (14), considere que $Y_1, \ldots, Y_n$ são variáveis aleatórias condicionalmente independentes, dado um vetor de covariáveis $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Assim, o modelo de regressão Beta é definido por:

$$Y_i \mid \boldsymbol{x}_i \sim \text{Beta}(\mu_i, \phi) \quad \text{com} \quad g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{16}$$

em que $\boldsymbol{x}_i$ e $\boldsymbol{\beta}$ são $p \times 1$ vetores de covariáveis conhecidas e desconhecidos parâmetros de regressão, respectivamente. Além disso, $g(\cdot)$ representa a função de ligação.

As estimativas de máxima verossimilhança são obtidas pela maximização da função log-verossimilhança, em relação aos parâmetros, a qual é definida pelo logaritmo natural do produto das função densidades definida em (14). É importante notar que o parâmetro de precisão, ϕ, será considerado constante.

2.2.2 Modelo Simplex

Proposta por Barndorff-Nielsen e Jørgensen (1991) a distribuição Simplex também é parametrizada por sua esperança (μ) e um parâmetro extra de dispersão (ϕ). Sua função densidade de probabilidade pode ser expressa por:

$$f(y \mid \mu, \phi) = \left[2\,\pi\,\phi^2\{y\,(1-y)\}^3\right]^{-1/2} \exp\left\{-\frac{1}{2\,\phi^2}\,\left[\frac{(y-\mu)^2}{y\,(1-y)\,\mu^2(1-\mu)^2}\right]\right\}, \tag{17}$$

em que $y, \mu \in (0,1)$ e $\phi > 0$. Na realidade, a distribuição Simplex faz parte dos modelos de dispersão propostos por Jørgensen (1997) e que estendem os MLG's.

Sejam $Y_1, \ldots, Y_n$ variáveis aleatórias condicionalmente independentes, dado um vetor de covariáveis $\boldsymbol{x}_i$, em que cada Y_i, $i = 1, \ldots, n$, segue uma distribuição Simplex com média μ_i e parâmetro de dispersão ϕ constante. Formalmente, o modelo de regressão Simplex pode ser definido por:

$$Y_i \mid \boldsymbol{x}_i \sim \text{Simplex}(\mu_i, \phi) \quad \text{com} \quad g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{18}$$

Assim como no modelo Beta a esperança da distribuição é ligada ao preditor linear por meio de uma função de ligação $g(\cdot)$. A estimação dos parâmetros, $\boldsymbol{\beta}$ e ϕ, do modelo de regressão Simplex é realizada sob o paradigma da verossimilhança, ou seja, maximizando a função log-verossimilhança, a qual dada pelo logaritmo do produto das densidades (17), vista como uma função de $\boldsymbol{\beta}$ e ϕ. Ressalta-se que os parâmetros $\boldsymbol{\beta}$ e ϕ são ortogonais, isto é, são assintoticamente independentes.

3 Materiais

3.1 Conjuntos de dados

No que se segue, estarei descrevendo os dois conjuntos de dados utilizados para modelagem. Em especifico as próximas subseções consistem em i) descrição do estudo, ii) definição das variáveis e iii) caracterização dos dados por meio de uma análise descritiva.

3.1.1 Número de publicações produzidas por Ph.D. em Bioquímica

Originalmente apresentado e analisado por Long (1990), este conjunto de dados é proveniente de um estudo observacional em que a população foi definida como todos os bioquímicos que receberam o título de Ph.D. durante o período de 1956–1958 e 1961–1963, além disso, todas as bioquímicas que obtiveram o título de Ph.D. no período entre 1950 à 1967.

Conforme discute Long (1990), o principal objetivo deste estudo foi identificar os efeitos de determinadas covariáveis sobre a produtividade dos bioquímicos durante seu Ph.D. Ressalta-se que a produtividade foi mensurada por meio do número de publicações em revistas cientificas nos últimos três anos do Ph.D. (`art`). O levantamento contou com 915 observações das quais têm-se informações sobre sexo (`fem`), estado civil (`mar`), número de filhos menores que 6 anos (`kid5`), prestigio do programa de Ph.D. (`phd`), número de artigos publicados pelo orientador nos últimos três anos (`ment`).

Na Figura 1 apresenta-se a distribuição empírica do número de publicações. Observa-se que durante os três anos do Ph.D. a maioria dos bioquímicos não tiveram publicações, de fato pouco mais da metade tiverem nenhum publicação. Além disso, é importante destacar, que o número médio de publicações foi de 1,69 com uma variância de 3,71. Portanto, embora simples essas informações fornecem indícios de um fenômeno com excesso de zeros e superdispersão.

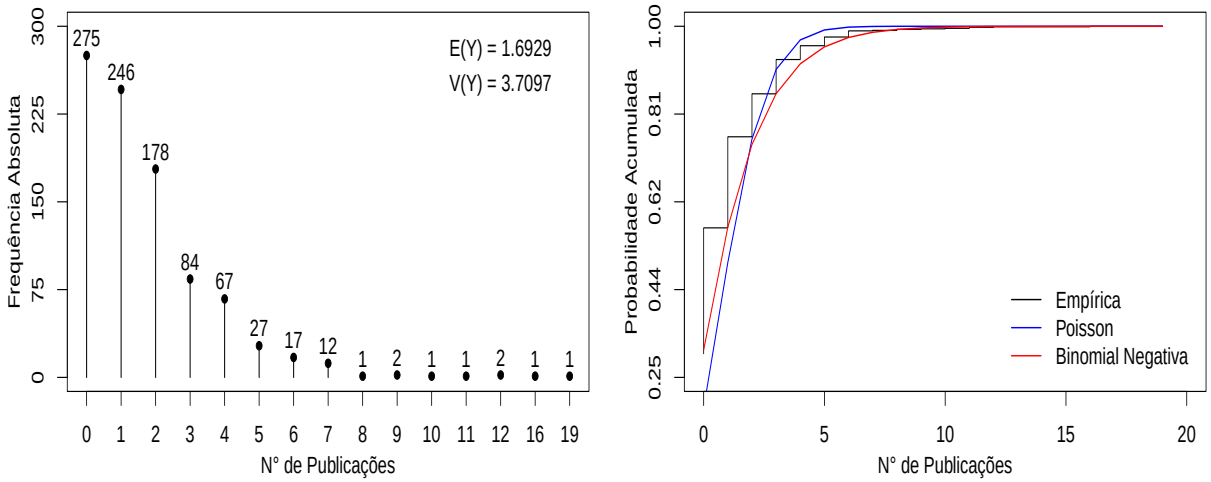

Figura 1: Distribuição empírica da quantidade de publicações dos(as) bioquímicos(as) Ph.D.

Nas Tabelas 1 e 2 são apresentadas algumas medidas amostrais do número de publicações conforme o gênero, estado civil e filhos. Em ambas as tabelas é possível identificar

a característica de superdispersão, uma vez que as variâncias amostrais são relativamente maiores que as respectivas médias.

Tabela 1: Medidas amostrais do número de publicações dos(as) bioquímicos(as) Ph.D. conforme gênero e estado civil.

Gênero	Estado Civil	N	Mínimo	Máximo	Média	Variância	CV (%)
Homem	Solteiro	113	0	7	1.9469	4.0507	103.3766
	Casado	381	0	19	1.8635	4.9655	119.5775
Mulher	Solteira	196	0	7	1.3878	2.2796	108.7979
	Casada	225	0	10	1.5422	2.5172	102.8751

Tabela 2: Medidas amostrais do número de publicações dos(as) bioquímicos(as) Ph.D. conforme número de filhos.

N° de Filhos	N	Mínimo	Máximo	Média	Variância	CV (%)
0	599	0	19	1.7212	3.7365	112.3057
1	195	0	12	1.7590	4.1942	116.4298
2	105	0	11	1.5429	3.0198	112.6320
3	16	0	3	0.8125	0.8292	112.0721

É importante pontuar que este conjunto de dados já foi analisado por Long e Freese (2001) e Mazucheli, Oliveira e Achcar (2017) utilizando os modelos de regressão Binomial Negativa e Lindley discreta, respectivamente.

3.1.2 Votação presidencial de 2010

Nesta segunda aplicação o interesse reside em identificar os fatores que influenciaram a proporção de votos por municípios do Paraná que a candidata, filiada ao Partido dos Trabalhadores (PT), Dilma Roussef recebeu nas eleições presidenciais de 2010. As fontes dos dados relacionados a proporção de votos do PT em 2006 e 2010 foram utilizadas por Furriel (2017) e gentilmente cedidas pelo autor. Enquanto que as variáveis demográficas de cada cidade foram obtidas do censo de 2010 realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e estão disponíveis pelo sitio eletrônico `<http://www.atlasbrasil.org.br>`. A Tabela 3 apresenta uma breve descrição das variáveis utilizadas nesta análise, vale mencionar que a variável resposta esta codificado como sendo pt2010. É importante destacar também, que os dados cobrem 395 municípios do Paraná, uma vez que em 4 cidades não foi possível encontrar a proporção de votos do segundo turno da eleição de 2006.

Tabela 3: Descrição das variáveis consideradas.

Variável	Definição
pt2010	Proporção de votos válidos do PT no segundo turno da eleição de 2010
pt2006	Proporção de votos válidos do PT no segundo turno da eleição de 2006
gini	Índice de Gini
idhm_e	Índice de Desenvolvimento Humano Municipal – Dimensão Educação
idhm_l	Índice de Desenvolvimento Humano Municipal – Dimensão Longevidade
idhm_r	Índice de Desenvolvimento Humano Municipal – Dimensão Renda
urb	Proporção da população vivendo em áreas urbanas
des	Proporção da PEA com 18 ou mais de idade em 2010 que estava desocupada
pib	PIB per capita do município

Na Figura 2 é possível observar o comportamento da variável resposta, bem como a curva teórica das duas distribuições de probabilidade que serão utilizadas para o modelo de regressão. Destaca-se que a proporção de votos do PT no segundo turno de 2010 nos municípios do Paraná apresenta um comportamento aproximadamente simétrico. Além disso, é visível que ambas as distribuições possuem uma ajuste similar, diferenciando somente em seus pontos de máximo.

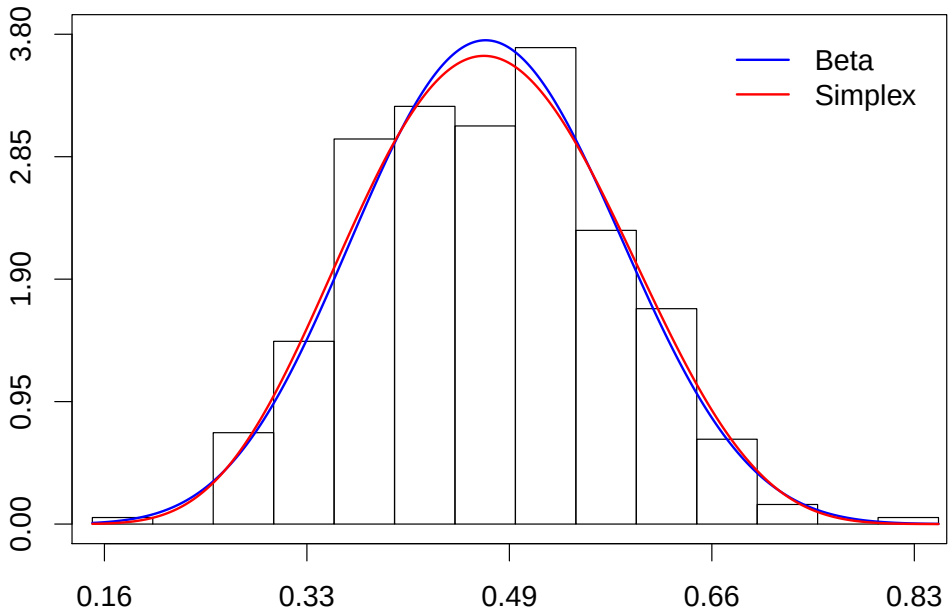

Figura 2: Comportamento empírico da proporção de votos válidos do PT em 2010 .

As seguintes medidas descritivas: mínimo, primeiro quartil ($Q_{0.25}$), mediana, média, terceiro quartil ($Q_{0.75}$), máximo e coeficiente de variação ($CV\%$) são apresentadas na Tabela 4. Tais estatísticas foram obtidas utilizando a `PROC MEANS` do software SAS.

Tabela 4: Estatísticas descritivas das variáveis consideradas.

Variável	Mínimo	$Q_{0.25}$	Mediana	Média	$Q_{0.75}$	Máximo	CV%
pt2010	0.1877	0.3995	0.4784	0.4782	0.5492	0.8098	21.5757
pt2006	0.1426	0.3611	0.4271	0.4291	0.5063	0.7539	22.8908
gini	0.3300	0.4300	0.4700	0.4657	0.5000	0.6600	12.2630
idhm_e	0.3620	0.5760	0.6210	0.6110	0.6550	0.7680	10.2870
idhm_l	0.7650	0.8050	0.8210	0.8205	0.8360	0.8700	2.5558
idhm_r	0.5700	0.6690	0.6920	0.6919	0.7150	0.8500	5.5289
urb	0.0935	0.5527	0.7190	0.6839	0.8423	1.0000	29.6121
des	0.0036	0.0289	0.0392	0.0410	0.0501	0.1013	40.6904
pib	5.8737	10.4045	12.9056	14.6271	16.2779	103.8509	57.5005

Algumas conclusões gerais podem ser extraídas da Tabela 4. Em 50% dos municípios do Paraná o percentual de votos válidos do PT no segundo turno da eleição de 2006 foi quase 43%, enquanto que em 2010 este percentual foi superior a 47%. O maior índice de Gini observado no paraná foi de 66%, enquanto que o menor foi de 33%. O terceiro quartil da variável des é 0.0501. Isto indica que em 75% dos municípios a proporção da população economicamente ativa que estava desocupada é menor ou igual a 5%. Por fim, verifica-se que na época 50% dos municípios o PIB per capita era menor ou igual a 12,90 mil/hab.

3.2 Recursos computacionais

Neste trabalho as análises foram conduzidas nos softwares SAS® versão 9.4, e R versão 3.3.2. Para os ajustes dos modelos optou-se pelo software SAS®, sendo que os seguintes procedimentos foram utilizados: `PROC GENMOD`, `PROC GLIMIX` e `PROC NLMIXED`. Já a apresentação gráfica dos resultados optou-se pelo software R e sua biblioteca `ggplot2`. Finalmente, para elaboração do relatório utilizou-se o sistema LaTeX.

4 Resultados e Discussões

Nesta seção serão apresentados os resultados relacionados aos ajustes dos modelos de regressão discutidos na seção 2 para os dados descritos na seção 3. A comparação dos modelos é realizada sob o ponto de vista da verossimilhança. Na avaliação do ajuste serão discutidas as principais abordagens utilizadas na literatura.

4.1 Análise do número de publicações produzidas por Ph.D. em Bioquímica

Conforme a descrição dos dados, realizada na seção 3, podemos notar indícios de superdispersão. O modelo considerado inicialmente contempla todas as covariáveis e para todas as distribuições a função de ligação $\log(\mu_i)$ foi considerada. Portanto, os modelos especificados modelam a média do número de publicações dos bioquímicos Ph.D., por meio da expressão:

$$\log(\mu_i) = \beta_0 + \beta_1 \,\text{fem} + \beta_2 \,\text{mar} + \beta_3 \,\text{kid5} + \beta_4 \,\text{phd} + \beta_5 \,\text{ment}. \tag{19}$$

em que a descrição das covariáveis foi exposta na seção 3.

A seguir é apresentada, na Tabela 5, as estimativas dos parâmetros juntamente com seus erros padrões para os três modelos ajustados.

Tabela 5: Estimativas dos parâmetros e erro padrão para os três modelos considerados.

	Poisson		Quase-Poisson		Binomial Negativo	
Parâmetro	Estimativa	EP	Estimativa	EP	Estimativa	EP
β_0	0.0800	0.0986	0.0800	0.1334	0.0397	0.1328
β_1	0.2246	0.0546	0.2246	0.0739	0.2164	0.0727
β_2	0.1552	0.0614	0.1552	0.0830	0.1505	0.0821
β_3	-0.1849	0.0401	-0.1849	0.0543	-0.1764	0.0531
β_4	0.0128	0.0264	0.0128	0.0357	0.0153	0.0360
β_5	0.0255	0.0020	0.0255	0.0027	0.0291	0.0035
σ^2, ϕ	1.0000	—	1.8290	—	0.4416	0.0530

Nota: estimativas estatisticamente significativas.

No que se refere aos modelos Poisson e Quase-Poisson as estimativas são idênticas por construção, divergindo seus erros padrões, que no caso Quase-Poisson é corrigido pelo parâmetro σ^2. Dessa forma, verifica-se que a covariável `mar` não é significativa considerando o modelo Quase-Poisson. As estimativas dos parâmetros bem como os erros padrões do modelo Binomial Negativo são semelhantes ao caso Quase-Poisson, no qual a covariável `mar` também não é estatisticamente significativa para o modelo. As estimativas dos parâmetros σ^2 e ϕ dos modelos Quase-Poisson e Binomial Negativo indicam problemas de superdispersão, uma vez que são menor que 1 e zero, respectivamente.

As medidas utilizadas para discriminação entre os modelos considerados estão expostas na Tabela 6. Podemos identificar de imediato que o modelo Binomial Negativo apresentou

os menores valores de AIC e BIC, bem como o maior valor da log-verossimilhança (ℓ) quando comparado com o modelo Poisson, caracterizando um melhor ajuste aos dados. A discriminação entre os modelos Binomial Negativo e Quase-Poisson pode ser realizada analisando os valores da deviance e estatística de Pearson (χ^2), uma vez que o modelo Quase-Poisson não faz suposição da distribuição dos dados. Fica claro que, outra vez, que o modelo Binomial Negativo apresenta menores valores da deviance e χ^2, indicando o melhor ajuste que os demais.

Tabela 6: Medidas de ajuste para avaliação e comparação entre os modelos.

Modelo	Deviance	Pearson χ^2	ℓ	AIC	BIC
Poisson	1634.3710	1662.5466	-642.0261	3314.1126	3343.0262
Binomial Negativo	1004.2815	944.5494	-551.9281	3135.9167	3169.6491
Quase-Poisson	1284.0522	1296.0522	—	—	—

Conforme foi discutido o modelo Binomial Negativo apresentou melhor ajuste aos dados comparado com as alternativas consideradas neste trabalho. Assim sendo, estarei considerando ele para modelar o número de publicações produzidas por Ph.D. em Bioquímica. Dos resultados apresentados na Tabela 5 verificou-se pelo teste de Wald que a covariável prestigio do programa de Ph.D. (`phd`) não é significativa para o modelo. Além disso, pode-se notar que seu erro padrão é aproximadamente três vezes maior que sua estimativa, comprovando sua insignificância para a análise. Portanto o modelo que será considerado é formalmente definido por:

$$\begin{aligned} Y_i \mid \boldsymbol{x}_i &\sim \text{BN}(\mu_i, \phi) \\ \log(\mu_i) = \eta_i &= \beta_0 + \beta_1 \,\text{fem} + \beta_2 \,\text{mar} + \beta_3 \,\text{kid5} + \beta_4 \,\text{ment}. \end{aligned} \tag{20}$$

Os resultados do ajuste do modelo definido em (20) estão apresentados na Tabela 7. É possível notar, conforme o intervalo baseado na estatística Wald, que a covariável `mar`, isto é, o estado civil do pesquisador, pode influenciar negativamente ou positivamente o número de artigos produzidos, indicando uma certa imprecisão em sua estimativa.

Analisando as outras estimativas é esperado que o número de publicações dos homens seja $\exp(0.2167) \simeq 1.2394$ maior do que produtividade das mulheres. Também podemos notar que existe uma relação negativa entre o número médio de publicações e o número de filhos, de fato para cada filho espera-se $\exp(-0.1768) \simeq 0.8379$ publicações. Já o número de artigos publicados pelo orientador nos últimos três anos tem efeito positivo na produtividade do pesquisador. Especificamente, para cada artigo publicado pelo orientador, espera-se que o Ph.D. também tenha mais uma publicação ($\exp(0.0294) \simeq 1.0298$).

Tabela 7: Resumo do ajuste para o modelo Binomial Negativo.

Parâmetro	Estimativa	EP	LI-Wald	LS-Wald	S_W	$P(S_W > \chi^2)$
β_0	0.0867	0.0731	-0.0566	0.2299	1.41	0.2359
β_1	0.2167	0.0727	0.0742	0.3591	8.89	0.0029
β_2	0.1469	0.0817	-0.0131	0.3070	3.24	0.0720
β_3	-0.1768	0.0531	-0.2808	-0.0728	11.10	0.0009
β_4	0.0294	0.0034	0.0228	0.0360	75.98	$<.0001$
ϕ	0.4417	0.0530	0.3491	0.5587	—	—

Utilizando a aproximação quadrática da verossimilhança foi possível construir intervalo de confiança para o parâmetro de dispersão, o resultado mostrou que ϕ é maior que zero (ver Tabela 7). Alternativamente, na Figura 3 temos os intervalos de confiança baseados na log-verossimilhança perfilada, o valor zero também não esta dentro dos limites de confiança de 99, 95 e até 90%. Portanto, mais uma vez comprova-se o fenômeno de superdispersão para os dados em estudo.

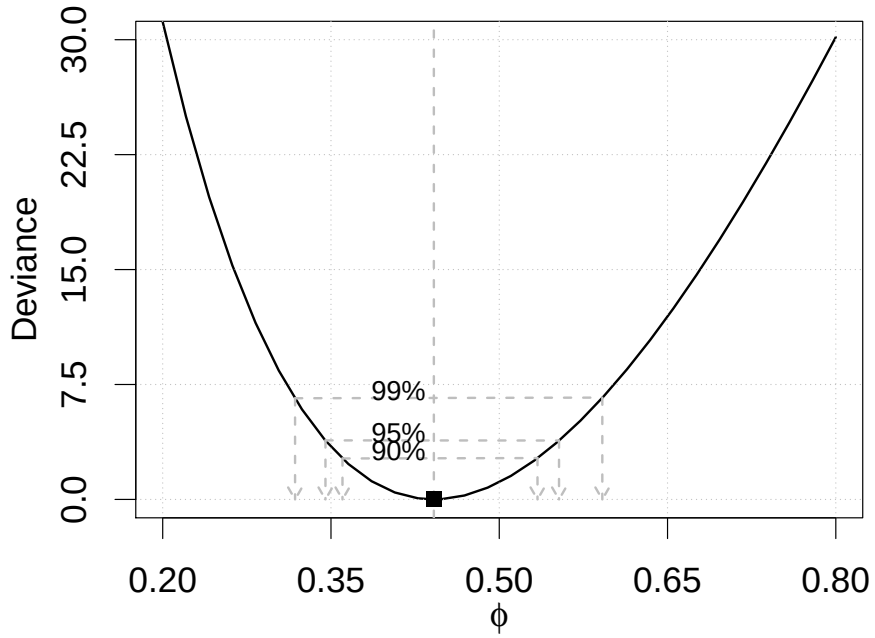

Figura 3: Intervalos de confiança determinados com base na função de log-verossimilhança perfilada.

Outro aspecto relevante que devemos observar é a correlação empírica entre os parâmetros estimados, este fato é importante pois toda inferência sobre os parâmetros é realizada de forma independente. Na Figura 4 destaca-se que correlação entre o parâmetro de dispersão $\widehat{\phi}$ e os parâmetros de regressão $\widehat{\beta}$'s são pequenas, indicando uma certa ortogonalidade empírica dos parâmetros para este conjunto de dados.

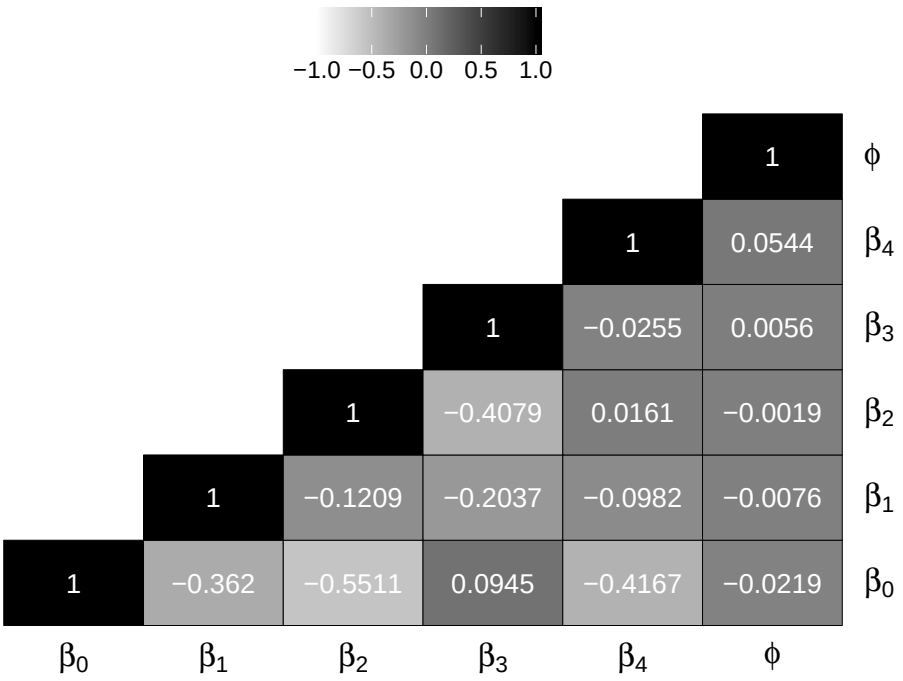

Figura 4: Correlação entre os parâmetros do modelo Binomial Negativo.

No que tange a crítica do modelo é apresentado a seguir uma série de métodos e gráficos que avaliam se o modelo especificado (20) possui um ajuste adequado aos dados. A Figura 5 ilustra o gráfico dos resíduos de pearson e da deviance versus o preditor linear. Examinando os resultados apresentados na Figura 5 é possível constatar uma certa padrão de decaimento conforme aumenta o preditor linear, uma explicação para este comportamento seria a presença das variáveis categóricas (`fem` e `mar`).

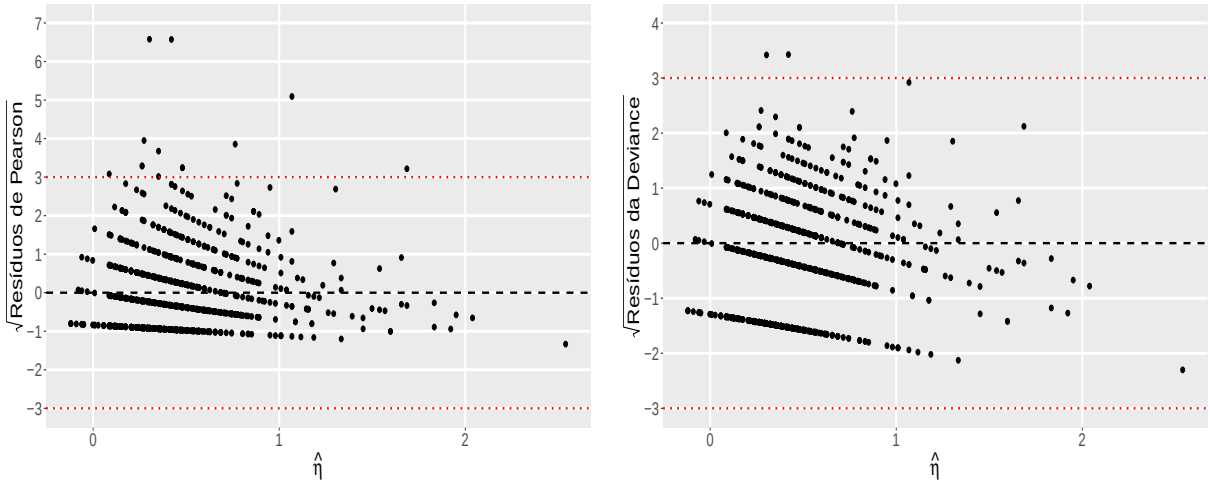

Figura 5: Gráficos de resíduos versus o preditor linear η.

Uma vez que a distribuição dos resíduos não é conhecida um excelente dispositivo para seu diagnóstico é o gráfico meio-normal de probabilidade com o envelope simulado exposto na Figura 6. Nota-se que somente 1.31% das observações estiveram foram do envelope, entretanto a forma entre os percentis da Normal e os resíduos não representa

uma reta como era de se esperar. Portanto existe certas evidências contra a adequacidade do modelo.

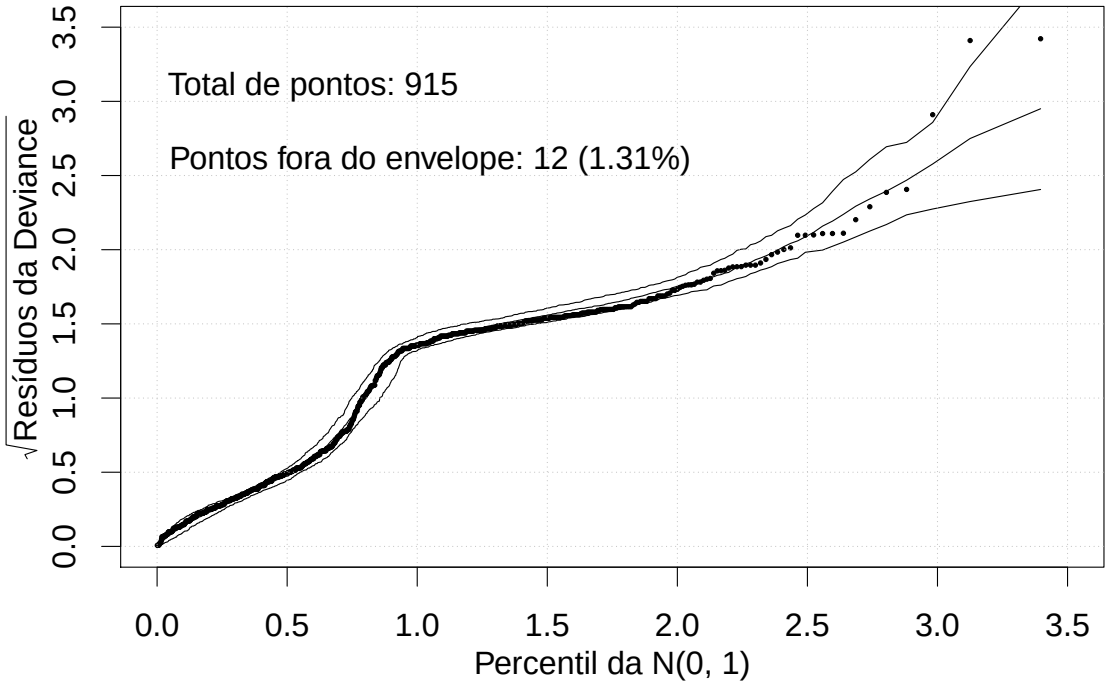

Figura 6: Gráfico meio-normal de probabilidades com envelope simulado.

Este comportamento dos resíduos pode estar associado grande número de zeros que existem no conjunto de dados. Dessa forma, uma alternativa seria modelar separadamente os zeros. Assim optou-se por comparar o comportamento dos resíduos entre os modelos Binomial Negativo, Poisson, Poisson Inflacionada de Zero e Binomial Negativa Inflacionada de zero. Para isso, foi construído o gráfico meio-normal de probabilidades com envelope simulado utilizando o resíduo de Pearson com intuito de comparar na mesma escala, os gráficos estão apresentados na Figura 7. Percebe-se que nos mesmo modelos inflacionados de zero o resíduo apresenta um comportamento próximo a reta. No entanto, o modelo Binomial Negativo apresentou o menor percentual de pontos fora do envelope, indicando melhor comportamento de forma geral.

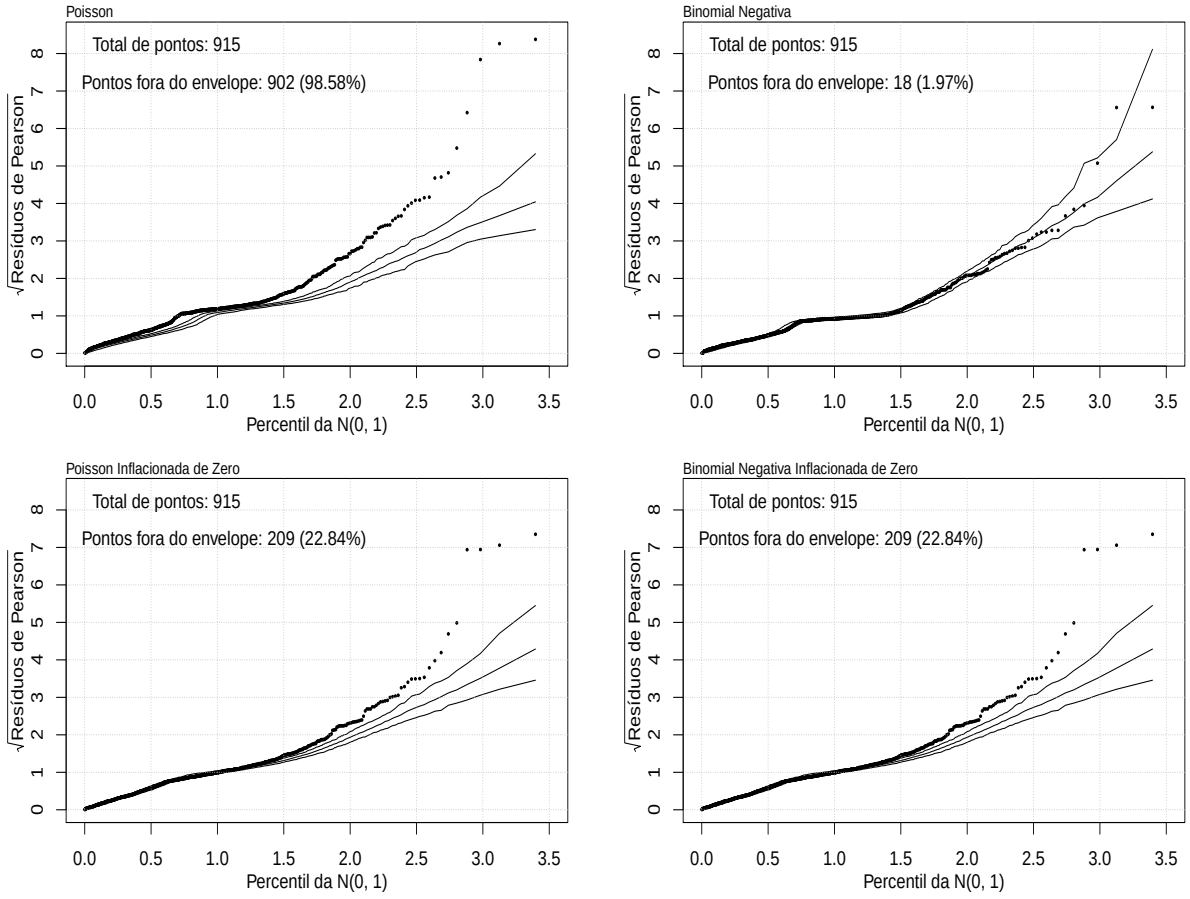

Figura 7: Gráficos meio-normal de probabilidades com envelope simulado.

Afim de formalizar a comparação entre os três modelo citados acima com o Binomial Negativo exibimos na Tabela 8 as estatísticas usuais para discriminação entre modelos (AIC e BIC), bem como os resultados do teste para modelos não encaixados proposto por Vuong (1989). É importante mencionar que o teste de Voung testa a hipótese de igualdade entre modelos não encaixados. Valores positivos da estatística indicam que o modelo Binomial Negativo (BN) é mais próximo do verdadeiro modelo.

Por exemplo, verifica-se que modelo BN está mais próximo do verdadeiro modelo do que os modelos Poisson e ZIP. Além disso, a hipótese de igualdade entre modelos também é rejeitada. Por outro lado, o modelo ZINB apresentou uma pequena vantagem sobre o modelo BN, entretanto é importante notar que esta pequena é compensada pelo número de parâmetro e a complexidade prática do modelo ZINB.

Tabela 8: Resultados da discriminação entre o modelo Binomial Negativo e os demais.

Modelo	np	AIC	BIC	Vuong (valor-p)
BN	6	3134	3163	—
Poisson	5	3312	3336	4.3423 (< 0.000)
ZIP	10	3230	3278	2.4882 (0.0128)
ZINB	12	3122	3175	-6.500 (0.6916)

Neste sentido, iremos escolher o modelo Binomial Negativo para modelar o número de artigos em função das covariáveis do estudo. Assim, iremos estudar sob um maior enfoque a função de ligação e variância utilizadas bem como verificar a existência de pontos de alavanca e influência.

A Figura 8 exibe o gráfico do preditor linear versus a variável z. Este gráfico é útil para identificar se a função de ligação esta adequada. Conforme podemos notar existe uma certa tendência linear crescente entre $\widehat{\eta}$ e z.

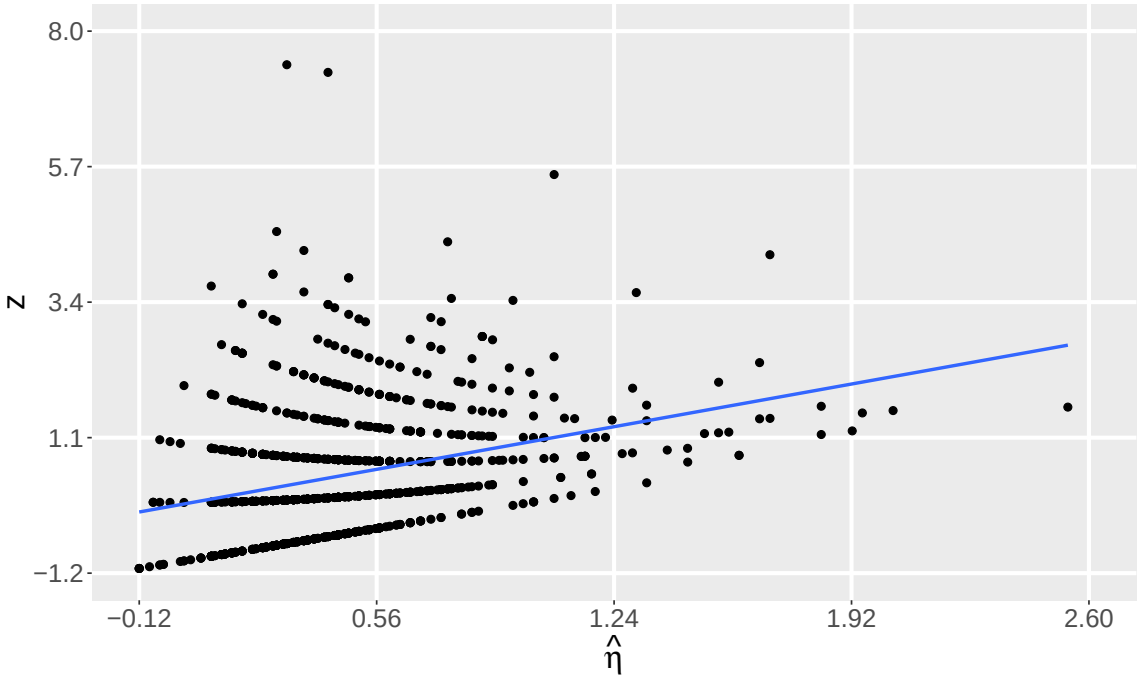

Figura 8: Método gráfico para avaliar a função de ligação.

Um método mais formal para avaliar a função de ligação é o teste RESET (Regression Specification Error Test) proposto por Ramsey (1969). Este teste consiste em incluir $\widehat{\eta}^2$, isto é os valores estimados do preditor ao quadrado, no modelo e testar se essa nova "covariável" deve permanecer no modelo. Caso a função de ligação esteja corretamente especificada então a hipótese nula ($\mathcal{H}_0 : \beta_{k+1} = 0$) não deve ser rejeitada. O resultado do teste RESET esta apresentado na Tabela 9 e indica que a função de ligação log não é adequada para este conjunto de dados.

Tabela 9: Estimativas e erro padrão com a adição de η^2

Parâmetro	Estimativa	EP	S_W	$P(S_W > \chi^2)$
β_0	-0.0710	0.0885	0.64	0.4225
β_1	0.3447	0.0827	17.38	<.0001
β_2	0.2423	0.0865	7.85	0.0051
β_3	-0.2711	0.0606	20.04	<.0001
β_4	0.0576	0.0095	36.93	<.0001
β_5	-0.5670	0.1761	10.36	0.0013

Nota: β_5 é o parâmetro associado a $\widehat{\eta}^2$

Na Figura 9 temos o gráfico os resíduos absolutos versus os valores ajustados. Nota-se que os pontos ficaram dispersos sem uma tendência definida implicando assim numa função de variância adequada.

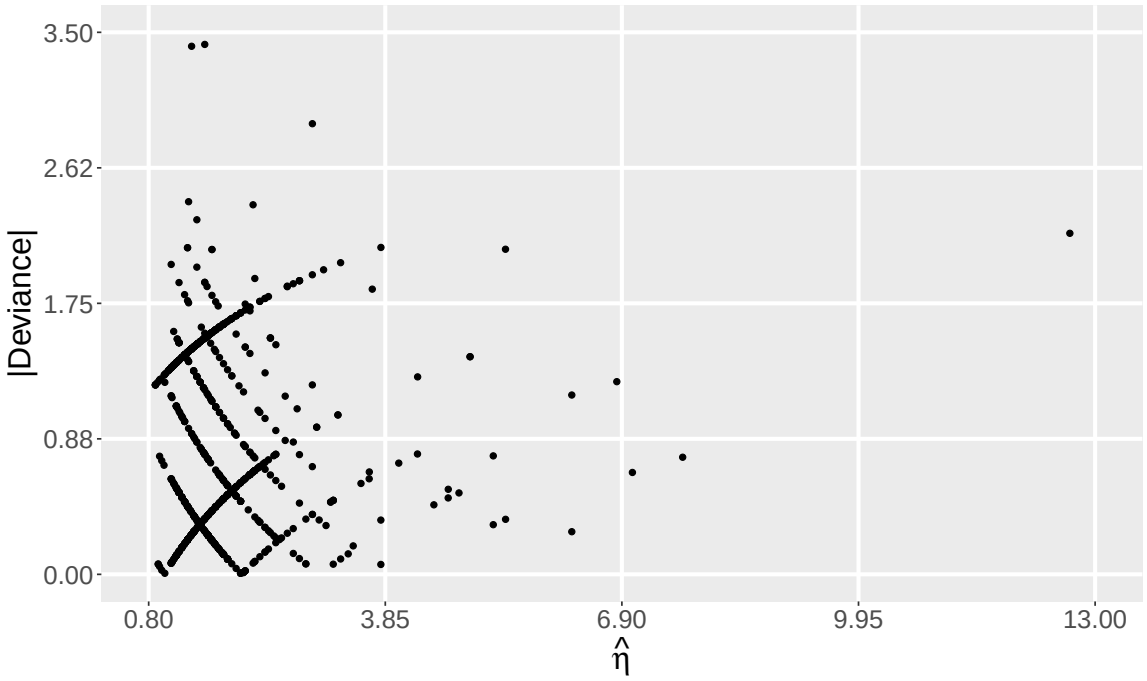

Figura 9: Método gráfico para avaliar a função de variância.

Para análise de alavancagem e influência foi utilizado os métodos gráficos exibidos na Figura 10. Observa-se que o maior valor do leverage foi aproximadamente 9 para a observação 328. Além disso, a mesma observação foi destaque no gráfico da distância de Cook. A observação 328 refere-se a um pesquisador casado que publicou um artigo durante seu Ph.D., ressalta-se que seu orientador teve 77 artigos publicados nos últimos três anos, um número 10 vezes maior que a média dessa covariável. Portanto, conclui-se que a observação 328 é fortemente influenciada pelo valor observado da covariável `ment`.

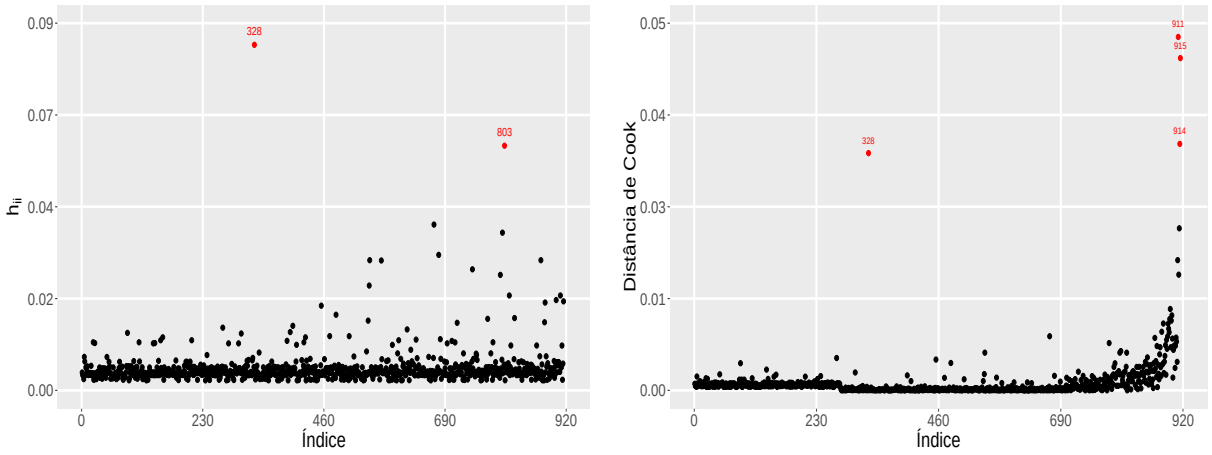

Figura 10: Gráficos para identificação de pontos de alavanca e influência.

Dessa forma, constatou-se que dentre os quatro modelos alternativos (Poisson, Quase-Poissonm, ZIP e ZINB) o modelo Binomial Negativo apresentou melhor ajuste, conforme análise de resíduo apresentada pela Figura 7 e os critérios de discriminação baseados na função log-verossimilhança (AIC, BIC). Todavia, foi verificado uma má especificação na função de ligação de acordo com o teste RESET. Além disso, é possível notar que os resíduos não possuem o comportamento desejado. Em relação a pontos de alavanca e influência a observação 328 surgiu como potencial ponto de alavancagem. Portanto, recomenda-se o estudo de outras distribuições alternativas, tais como a Lindley Discreta, Weibull Disceta, Poisson Generalizada, etc, afim captar os dados analisados de forma mais adequada do que o modelo Binomial Negativo.

4.2 Análise da votação presidencial de 2010

Nesta seção apresento uma modelagem empírica relacionada à eleição presidencial no Brasil no ano de 2010. O objetivo é identificar quais foram os principais fatores que influenciaram a proporção de votos que a ex-presidenta Dilma recebeu nas eleições presidenciais de 2010. Uma vez que a natureza da variável resposta, proporção de votos, é limitada ao intervalo $(0, 1)$ devemos considerar uma distribuição de probabilidade apropriada a natureza da variável resposta.

Seguramente a distribuição Beta é a mais utilizada e conhecida para modelar dados desta natureza, por isso ela é nossa primeira opção nesta modelagem. Alternativamente, também, consideramos a distribuição Simplex proposta por Barndorff-Nielsen e Jørgensen (1991). Dessa forma, para ambos as distribuições o modelo de regressão para modelar a média pode ser expresso por:

$$\begin{aligned}\text{logit}(\mu_i) = \beta_0 + \beta_1\,\texttt{pt2006} + \beta_2\,\texttt{gini} + \beta_3\,\texttt{idhm\_e} + \beta_4\,\texttt{idhm\_l} + \beta_5\,\texttt{idhm\_r} + \beta_6\,\texttt{urb} \\ + \beta_7\,\texttt{des} + \beta_8\,\texttt{pib}\end{aligned} \tag{21}$$

em que a descrição das covariáveis foi exposta na seção 3. Note que para ambas as distribuições será utilizado a função de ligação $\eta_i = \text{logit}(\mu_i) = \log\left(\dfrac{\mu_i}{1-\mu_i}\right)$. Ressalta-se também que o parâmetro de dispersão é considerado constante, isto é, não há estrutura de regressão para ele.

Na Tabela 10 apresenta-se as estimativas e os erros padrões para os modelos de regressão Beta e Simplex. Estes resultados foram obtidos utilizando a `PROC NLMIXED` (SAS, 2010). Os resultados apresentados na Tabela 10 mostram que ambas os modelos de regressão, Beta e Simplex, concordam com o efeito das covariáveis, cem como sua significância.

Tabela 10: Estimativas dos parâmetros e erro padrão para os modelos considerados.

	Beta		Simplex	
Parâmetro	Estimativa	EP	Estimativa	EP
β_0	-1.1352	0.5211	-1.1305	0.5161
β_1 (`pt2006`)	3.0309	0.1255	3.0343	0.1205
β_2 (`gini`)	-0.7275	0.2634	-0.7397	0.2632
β_3 (`idhm_e`)	-0.9978	0.3041	-0.9828	0.2998
β_4 (`idhm_l`)	0.0316	0.6582	0.0597	0.6556
β_5 (`idhm_r`)	1.8597	0.6001	1.8277	0.5966
β_6 (`urb`)	-0.7461	0.0933	-0.7621	0.0922
β_7 (`des`)	-2.4220	0.8331	-2.3907	0.8244
β_8 (`pib`)	-0.0008	0.0015	-0.0009	0.0015
ϕ	78.5775	5.5570	0.4719	0.0168

Nota: estimativas estatisticamente significativas.

Uma vez que a log-verossimilhança é uma medida de compatibilidade do modelo com o particular conjunto de dados, os valores maximizados das log-verossimilhanças permitem comparar o ajuste de modelos que tenham a mesma estrutura de covariáveis e o mesmo número de parâmetros, mas que também tenham diferentes especificações na distribuição da variável resposta (BONAT; JR; ZEVIANI, 2013). Neste sentido, para escolha e comparação entre os modelos Beta e Simplex consideramos o valor -2 $\hat{\ell}$ e os critérios de informação AIC e BIC. Tais critérios penalizam os valores maximizados das log-verossimilhanças pelo número de parâmetros no modelo.

Na Tabela 11 apresenta-se os valores das medidas discutidas, bem como o resultado do teste de Vuong. Embora a diferença seja miníma observa-se que o modelo Simplex possui os menores valores dos critérios baseados na log-verossimilhança. Além disso, conforme o teste de Vuong não rejeita-se a hipótese nula de que ambos os modelos são iguais sob o ponto de vista da estatística.

Tabela 11: Medidas utilizados para discriminação entre os modelos.

Modelo	-2 $\hat{\ell}$	AIC	BIC	Vuong
Beta	-1169.4060	-1149.4060	-1109.6172	—
Simplex	-1171.3461	-1151.3461	-1111.5573	-0.8206 (0.7940)

Além das estatísticas baseadas na log-verossimilhança é importante identificar a qualidade do ajuste por meio da análise de resíduos. Dessa forma, na Figura 11 exibimos o gráfico normal de probabilidade com o envelope simulado. Verifica-se que ambos os modelos não apresentam afastamentos das especificações impostas. De fato, os resíduos parecem se comportar de maneira muito parecida.

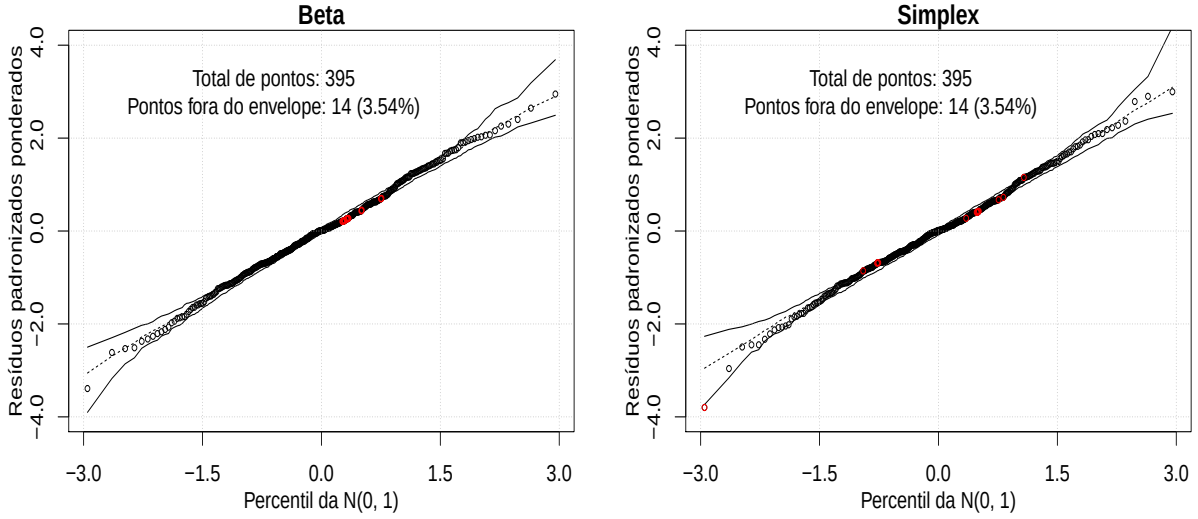

Figura 11: Gráficos normais de probabilidades com envelope simulado.

Com base nesta análise inicial pode-se verificar que o modelo Simplex apresentou melhor ajuste do que o modelo Beta, assim ele será considerado para modelar o proporção de votos em função das covariáveis especificadas. Foi exposto na Tabela 10 que somente as

covariáveis `idhm_l`, `idhm_r` e `pib` não foram significantes para o modelo. Verifica-se também que o erro padrão relacionado ao coeficiente da covariável `idhm_l` é substancialmente maior que sua estimativa, já a covariável `pib` apresentou um efeito muito próximo de zero assim como seu erro padrão. Por outro lado, a variável independente `idhm_r` embora, tenha sido insignificante do ponto de vista estatística, possui suas interpretações práticas para o problema em estudo.

Afim, de formalizar a insignificância das covariáveis `idhm_l` e `pib` testamos a hipótese nula $\mathcal{H}_0 : \boldsymbol{\theta} = 0$ versus a alternativa $\mathcal{H}_1 : \boldsymbol{\theta} \neq 0$, em que $\boldsymbol{\theta} = (\beta_4, \beta_8)$ representa os coeficientes associados as covariáveis. A estatística do teste de Wald apresentou valor igual a 0.4090 e valor-p correspondente de 0.8151. Assim, não rejeita-se a hipótese nula de que o vetor $\boldsymbol{\theta}$ é igual a zero. O mesmo resultado pode ser encontrado por meio do teste da razão de verossimilhanças.

Logo, o modelo selecionado assume distribuição Simplex para a variável resposta com parâmetro de precisão (ϕ) constante, e a sua média é modelada pela expressão:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1\,\texttt{pt2006} + \beta_2\,\texttt{gini} + \beta_3\,\texttt{idhm\_e} + \beta_4\,\texttt{idhm\_r} + \beta_5\,\texttt{urb} + \beta_6\,\texttt{des} \quad (22)$$

Através da análise dos resultados apresentados na Tabela 12 podemos retirar algumas importantes conclusões. Primeiramente, ressalta-se que as covariáveis `gini`, `idhm_e`, `urb` e `des` exercem efeito negativo sobre a proporção de votos recebidas pela ex-presidenta Dilma Rousseff nas eleições de 2010 no estado do Paraná. Isto significa que quanto maior for o índice de Gini (maior desigualdade), a dimensão educação do IDH, a proporção de pessoas residentes em áreas urbanas e a proporção de pessoas desocupadas, estes municípios tendem a apresentar, em média, um menor percentual de votos a favor de Dilma. Em contrapartida, tiveram uma influencia positiva na proporção de votos da Dilma as covariáveis `pt2006` e `idhm_r`, indicando que quanto maior o percentual de votos de Lula em 2006 e a dimensão renda do IDH, espera-se um maior percentual de votos da Dilma nestes municípios.

Tabela 12: Resumo do ajuste considerando o modelo (22).

Parâmetro	Estimativa	EP	LI	LS	t	P($T > t$)
β_0	-1.0463	0.2685	-1.5725	-0.5201	-3.8974	0.0001
β_1 (`pt2006`)	3.0302	0.1202	2.7947	3.2658	25.2130	<0.000
β_2 (`gini`)	-0.7283	0.2628	-1.2434	-0.2131	-2.7708	0.0056
β_3 (`idhm_e`)	-0.9832	0.2984	-1.5681	-0.3984	-3.2951	0.0010
β_4 (`idhm_r`)	1.7495	0.5684	0.6355	2.8635	3.0780	0.0021
β_5 (`urb`)	-0.7558	0.0915	-0.9352	-0.5765	-8.2599	<0.000
β_6 (`des`)	-2.4523	0.8190	-4.0575	-0.8471	-2.9943	0.0028
ϕ	0.4722	0.0168	0.4392	0.5051	—	—

Com o intuito de realizar uma critica sobre o modelo ajustado, isto uma análise de resíduo e diagnóstico, algumas ferramentas gráficas foram utilizadas para detectar possíveis afastamentos das suposições feitas pelo modelo de regressão definido em (22). Entre elas, apresenta-se na Figura 12 o gráfico de probabilidades normal com envelopes simulados. O resíduo proposto por Miyashiro (2008) foi utilizado, pode-se ver na Figura 12

que a grande maioria dos resíduos permanecem dentro do envelope, então designa-se que o modelo escolhido aparenta fornecer boa representação dos dados. Adicionalmente, o teste RESET (RAMSEY, 1969) foi considerado. A hipótese nula de que o modelo sob julgamento esta bem especificado não foi rejeitada, em que o valor-p obtido foi de 0.2977, muito acima dos usuais níveis de significância.

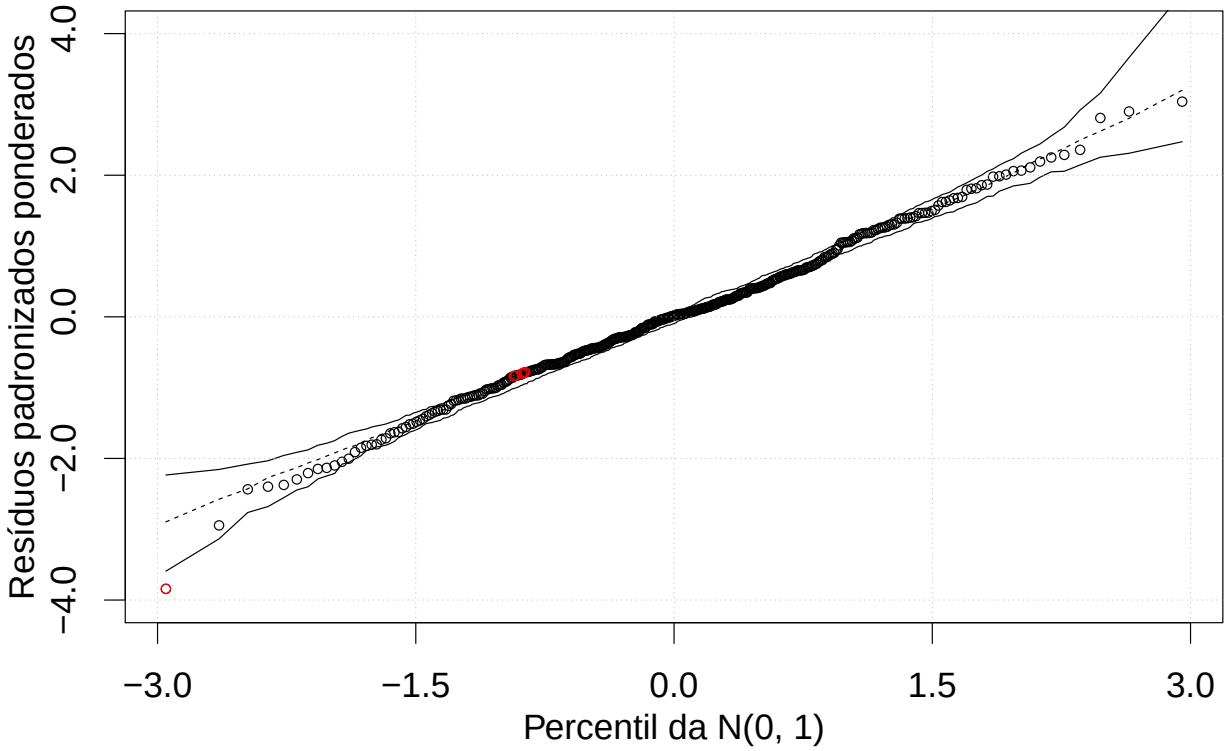

Figura 12: Gráfico normal de probabilidades com envelope simulado.

Ainda neste contexto, temos na Figura 13 o gráfico dos resíduos versus os índices das observações. Verifica-se que o modelo Simplex selecionado exibe bom ajuste, uma vez que somente uma observação se encontra fora do intervalo (-3,3).

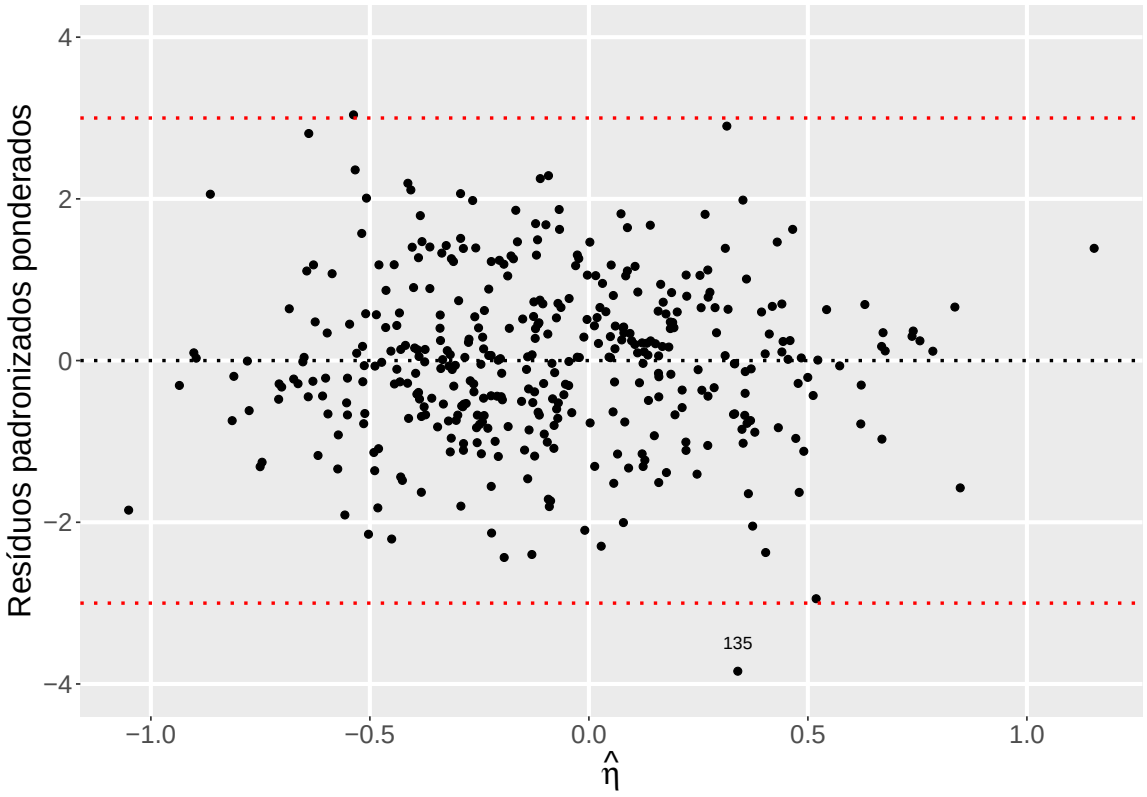

Figura 13: Gráfico dos resíduos versus os índices das observações

Na detecção de pontos influentes foi considerada a distância de Cook definida por Miyashiro (2008). A Figura 14 exibe o gráfico da distância de Cook versus os índices das observações. Nota-se que três municípios destacam-se dos demais, sendo eles Fênix (109), Guaraqueçaba (135) e Rio Azul (302). Destaca-se que no município de Fênix em 2010 a proporção de pessoas vivendo em áreas urbanas é superior a 83%, além disso 10% da população estava desocupada. Por outro lado, segundo os dados do IBGE nos municípios de Guaraqueçaba e Rio Azul haviam um pouco menos de 35% da população vivendo em áreas urbanas e além disso, em Guaraqueçaba a proporção de votos do PT teve uma queda de 48% em 2006 para 39% em 2010.

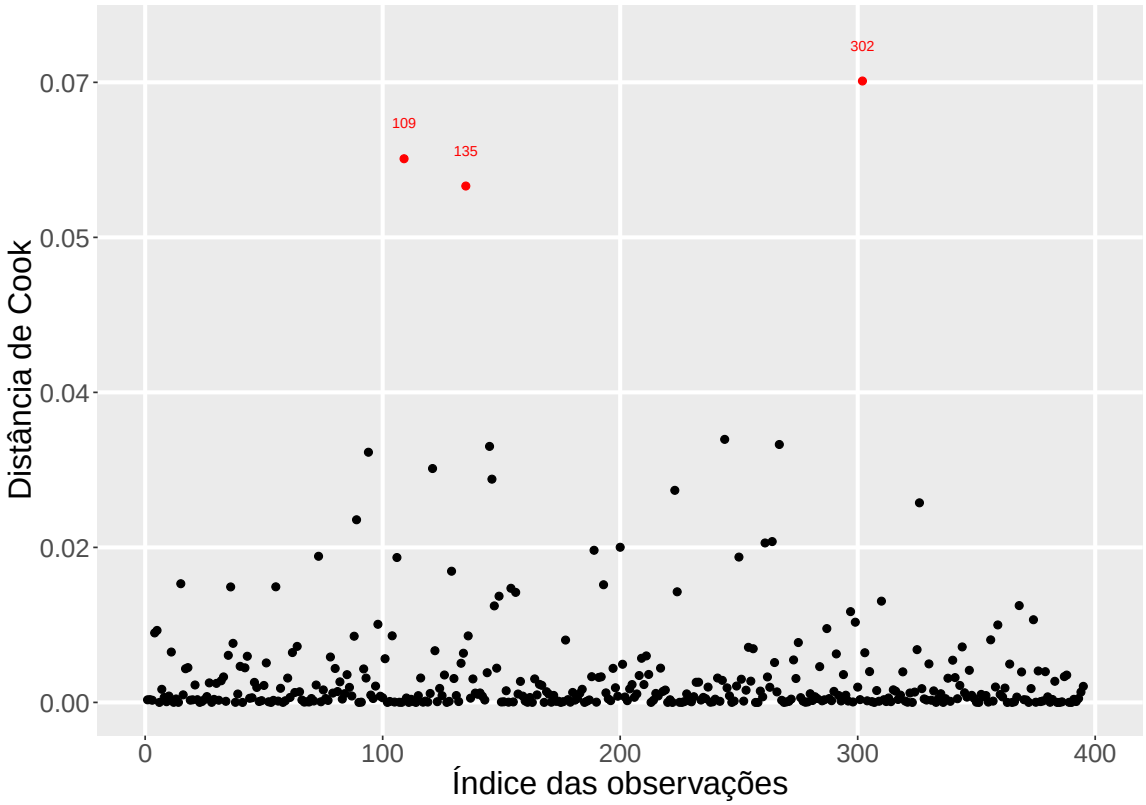

Figura 14: Gráfico da distância de Cook

Assim, conclui-se que o modelo Simplex obteve um ajuste satisfatório para os dados, sendo melhor que o usual modelo de regressão Beta. No que tange os resultados, verificou-se que no segundo turno das eleições de 2010 o percentual de votos de Lula em 2006 e municípios com maiores rendimentos per capita contribuíram positivamente para a proporção de votos recebidas por Dilma em 2010 no estado do Paraná. Por outro lado, observamos que a desigualdade de distribuição de renda, medida pelo índice de Gini, foi um efeito negativo sobre a proporção de votos da Dilma em 2010 no estado do Paraná.

5 Considerações Finais

Neste trabalho foi analisado sob o ponto de vista de regressão dois conjunto de dados, nos quais a variável resposta são de natureza discreta e continua. Uma breve contextualização sobre os modelos lineares generalizados foi exposta inicialmente. Em seguida, foi discutido três abordagens usuais para modelagem do primeiro conjunto de dados. No caso dos dados de natureza contínua considerei as distribuições Beta e Simplex, uma vez que a variável resposta é restrita ao intervalo unitário $(0, 1)$.

Em ambas as análises realizamos a comparação dos modelos considerados, análise de resíduos e diagnóstico, interpretação dos parâmetros estimados do modelo selecionado e por fim alguns considerações finais da modelagem. Ressalta-se, que os ajustes dos modelos foram realizados sob o paradigma da verossimilhança e utilizando o software SAS, além disso, leitores interessados podem encontrar os códigos em SAS no apêndice deste trabalho.

Referências

BARNDORFF-NIELSEN, O.; JØRGENSEN, B. Some parametric models on the Simplex. *Journal of Multivariate Analysis*, v. 39, n. 1, p. 106–116, 1991. ISSN 0047-259X.

BONAT, W. H.; JR, P. J. R.; ZEVIANI, W. M. Regression models for responses in the unit interval: specication, estimation and comparison. *Rev. Bras. Biom. (São Paulo)*, v. 20, n. 1, p. 1–10, 2013.

DOBSON, A. J. *An Introduction to Generalized Linear Models*. Second. Chapman & Hall/CRC, 2002.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, v. 31, n. 7, p. 799–815, 2004.

FURRIEL, O. W. *Determinantes do voto à presidência: análise espacial das eleições gerais no Brasil no período de 1994 a 2014*. 2017. PIC – Universidade Estadual de Maringá.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. Third. John Wiley & Sons, Inc, 2013.

JØRGENSEN, B. *The Theory of Dispersion Models*. Chapman & Hall/CRC, 1997.

JUNIOR, E. E. R. *Extensões e Aplicações do Modelo de Regressão Conway-Maxwell-Poisson para Modelagem de Dados de Contagem*. 2016. TCC, Universidade Federal do Paraná.

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

LAWLESS, J. F. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, Wiley-Blackwell, v. 15, n. 3, p. 209–225, 1987. ISSN 1708-945X. Disponível em: <http://dx.doi.org/10.2307/3314912>.

LONG, J. S. The origins of sex differences in science. *Social Forces*, Oxford University Press, v. 68, n. 4, p. 1297–1316, 1990. ISSN 00377732, 15347605.

LONG, J. S.; FREESE, J. Predicted probabilities for count models. *Stata Journal*, v. 1, n. 1, p. 7–51, 2001.

MAZUCHELI, J.; OLIVEIRA, R. P.; ACHCAR, J. A. A comparative study 1 between two discrete lindley distributions. *Ciencia & Natura*, 2017.

MCCULAGH, P.; NELDER, J. A. *Generalized Linear Models*. Second. Chapman & Hall, 1983.

MCCULLOCH, C. E.; SEARLE, S. R. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Inc, 2001.

MIYASHIRO, E. S. *Modelos de regressão Beta e Simplex para a análise de proporções*. Dissertação (Mestrado) — Universidade de São Paulo - USP, 2008.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. *Applied Linear Statistical Models*. Irwin Chicago, 1996. v. 4.

PAULA, G. A. *Modelos de Regressão com apoio computacional*. 2013.

RAMSEY, J. B. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, [Royal Statistical Society, Wiley], v. 31, n. 2, p. 350–371, 1969. ISSN 00359246.

SAS. *The NLMIXED Procedure, SAS/STAT® User's Guide, Version 9.22*. Cary, NC: SAS Institute Inc., 2010. 4967–5062 p.

SAS Institute Inc. *The GENMOD Procedure, SAS®/STAT User's Guide, Version 9.3*. Cary, NC: SAS Institute Inc., 2011. 4322–4412 p. 2605–2804.

VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, [Wiley, Econometric Society], v. 57, n. 2, p. 307–333, 1989.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 61, n. 3, p. 439–447, 1974. ISSN 00063444.

WEISBERG, S. *Applied Linear Regression*. John Wiley & Sons, 2005. v. 528.

Apêndice

Neste apêndice disponibilizo os códigos em SAS utilizados para a análise dos dois conjuntos de dados.

Listing 1: Códigos utilizados na modelagem dos dados discretos.

```
proc delete data = _all_ ; run;
proc import out = dados datafile = "~\couart2.dta";
run;
%include "~\envelope_macro.sas";
%include "~\voung-test.sas";

****** Comparacao de modelos;

/*Modelo Poisson*/
proc genmod data = dados;
        class fem mar;
        model art = fem mar kid5 phd ment / dist = poisson link = log;
        ods output ParameterEstimates=Estpoi(keep=parameter estimate StdErr where=(
            estimate <> 0));
        ods output ModelFit=gofpoi(drop = df ValueDF);
quit;

/*Modelo Binomial Negativa*/
proc genmod data = dados;
        class fem mar;
        model art = fem mar kid5 phd ment / dist = negbin link = log;
        ods output ParameterEstimates=EstBn(keep= estimate StdErr where=(estimate <> 0));
        ods output ModelFit=gofbn(drop = df ValueDF);
quit;

/*Modelo Quase-Poisson*/
proc glimmix data = dados ;
        class fem mar;
        _variance_ = _mu_;
        model art = fem mar kid5 phd ment / link = log solution;
        random _residual_;
        ods output ParameterEstimates=estQP(keep= estimate StdErr where=(estimate <> 0));
        ods output FitStatistics=gofqp;
quit;

****** Critica do modelo BN;
proc genmod data = dados;
        class fem mar;
        model art = fem mar kid5 ment / dist = negbin link = log;
        output out=bn predicted=predito xbeta=xb stdxbeta=stderror lower=lo upper=up
        reschi=rpearson resdev=rdeviance reslik=rlik stdreschi=sdpearson stdresdev=
            sdeviance
        hesswgt=weightmat leverage=leverage dfbetas=dfbeta cooksd=cook;
quit;

%envelope(data=bn, predict=predito, resid=rdeviance, class_v= fem mar, plinear=fem mar
    kid5 ment, dispersion=0.4417, family=nb, link=log, type=HN);
%envelope(data=bn, predict=predito, resid=rdeviance, class_v= fem mar, plinear=fem mar
    kid5 ment, dispersion= 0.4417, family=nb, link=log, type=N);

****** Discriminacao entre os modelos P, NB, ZIP e ZINB via teste de Voung;
proc countreg data=dados;
        model art=fem mar kid5 ment / d=zip;
        zeromodel art~fem mar kid5 ment;
        output out=outzip pred=predzip probzero=p0zip;
run;

proc countreg data=outzip method = qn;
        model art=fem mar kid5 ment / d=zinb;
        zeromodel art~fem mar kid5 ment;
        output out=outzinb pred=predzinb probzero=p0zinb;
run;
```

```
62
63 proc countreg data=outzinb;
64         model art=fem mar kid5 ment / d=Poisson;
65         output out=outp pred=predp;
66 run;
67
68 proc countreg data=outp method=qn;
69         model art=fem mar kid5 ment / d=negbin(p = 2);
70         output out=out pred=pnb;
71 run;
72
73 * BN vs Poisson;
74 %vuong(data=out, response=art, model1=neg. bin, p1=pnb, dist1=nb, nparm1=6, scale1
       =0.441673, model2=Poisson,  p2=predp, dist2=poi, nparm2=5)
75
76 * BN vs ZIP;
77 %vuong(data=out, response=art, model1=neg. bin, p1=pnb, dist1=nb, nparm1=6, scale1
       =0.441673, model2=zip, p2=predzip, dist2=zip, nparm2=10, pzero2=p0zip)
78
79 * BN vs ZINB;
80 %vuong(data=out, response=art, model1=neg. bin, p1=pnb, dist1=nb, nparm1=6, scale1
       =0.441673, model2=zip, p2=predzinb, dist2=zinb, nparm2=11, scale2=0.376253, pzero2=
       p0zinb)
```

Listing 2: Códigos utilizados na modelagem dos dados contínuos.

```
1  proc delete data = _all_ ; run;
2  ****** Importacao e ajuste;
3  proc import datafile  = '~idh_pt2010.txt' replace
4          out             = dados
5          dbms            = tab;
6          getnames        = yes;
7  run;
8
9  data parana(where = (ufn = 'Parana'));
10         set dados;
11         if (pt2006 = 0) then delete;
12 run;
13
14 ****** Descritiva das covariaveis;
15 proc means data = parana stackodsoutput min q1 median mean q3 max cv maxdec = 4;
16         var pt2010 pt2006 gini idhm_e idhm_l idhm_r urbprop pdes18m pibperc;
17 run;
18
19 ****** Ajuste e comparacao dos modelos;
20 * Regressao Beta;
21 proc nlmixed data = parana tech = tr update = bfgs df=999999;
22         parms b0 = 0, b1 = 0, b2 = 0, b3 = 0, b4 = 0, b5 = 0, b6 = 0, b7 = 0, b8 = 0,
               disp = 1;
23         y = pt2010;
24         eta = b0 + b1 * pt2006 + b2 * gini + b3 * idhm_e + b4 * idhm_l + b5 * idhm_r + b6
                * urbprop + b7 * pdes18m + b8 * pibperc;
25         mu  = exp(eta) / (1 + exp(eta));
26         phi = disp;
27         t = mu * phi;
28         w = (1 - mu) * phi;
29         ll = lgamma(phi) - lgamma(t) - lgamma(w) + (t - 1) * log(y) + (w - 1) * log(1 - y
               );
30         model y ~ general(ll);
31         predict mu out = betapred(rename = (pred = betapred Lower = betalower upper =
               betaupper) drop = DF StdErrPred tValue Probt alpha);
32         ods output ParameterEstimates = betaest(drop = df alpha lower upper gradient
               tValue Probt
33         rename = (estimate = estbeta StandardError = stdbeta));
34         ods output FitStatistics = betafit;
35 run;
36
37 proc glimmix data = parana;
38         model pt2010 = pt2006 gini idhm_e idhm_l idhm_r urbprop pdes18m pibperc / dist =
               beta solution;
39         output out=outbeta / allstats;
40 run;
41
```

```
42
43 * Regressao Simplex;
44 proc nlmixed data = parana tech = tr update = bfgs df=999999;
45         parms b0 = 0, b1 = 0, b2 = 0, b3 = 0, b4 = 0, b5 = 0, b6 = 0, b7 = 0, b8 = 0,
               disp = 1;
46         y = pt2010;
47         eta = b0 + b1 * pt2006 + b2 * gini + b3 * idhm_e + b4 * idhm_l + b5 * idhm_r + b6
                * urbprop + b7 * pdes18m + b8 * pibperc;
48         mu  = exp(eta) / (1 + exp(eta));
49         phi = disp;
50         pi = constant("pi");
51         ll = log(1/sqrt(2 * pi * phi**2 * (y * (1 - y))**3)) - 0.5 / phi**2 * ((y - mu)**
               2 / (y * (1 - y) * mu**2 * (1 - mu)**2));
52         model y ~ general(ll);
53         predict mu out = simppred(rename = (pred = simppred Lower = simplower upper =
               simpupper) drop = DF StdErrPred tValue Probt alpha);
54         ods output ParameterEstimates = simpest(drop = df alpha lower upper gradient
               tValue Probt
55         rename = (estimate = estsimp StandardError = stdsimp));
56         ods output FitStatistics = simpfit(rename = (value = svalue) drop = descr);
57 run;
58
59 ****** Ajuste do modelo Simplex;
60
61 * Verificando se as covariaveis pib e idhm_l sao conjuntamente significativas ;
62 proc nlmixed data = parana tech = tr update = bfgs df=999999;
63         parms b0 = 0, b1 = 0, b2 = 0, b3 = 0, b4 = 0, b5 = 0, b6 = 0, b7 = 0, b8 = 0,
               disp = 1;
64         y = pt2010;
65         eta = b0 + b1 * pt2006 + b2 * gini + b3 * idhm_e + b4 * idhm_l + b5 * idhm_r + b6
                * urbprop + b7 * pdes18m + b8 * pibperc;
66         mu = exp(eta) / (1 + exp(eta));
67         phi = disp;
68         pi = constant("pi");
69         ll = log(1/sqrt(2 * pi * phi**2 * (y * (1 - y))**3)) - 0.5 / phi**2 * ((y - mu)**
               2 / (y * (1 - y) * mu**2 * (1 - mu)**2));
70         model y ~ general(ll);
71         *       contrast 'b4 = b5 = b8 = 0' b4, b5, b8;
72         contrast 'b4 = b8 = 0'          b4, b8;
73         ods output Contrasts = Wald(drop = label DenDF rename = (NumDF = df FValue = est
               ProbF = pvalue));
74         ods output FitStatistics = H1(rename = (value = llh1 ) where = (descr = '-2 Log
               Likelihood'));
75 run;
76
77 * Ajuste sem as covariaveis pib e idhm_l;
78 proc nlmixed data = parana tech = tr update = bfgs df=999999;
79         parms b0 = 0, b1 = 0, b2 = 0, b3 = 0, b5 = 0, b6 = 0, b7 = 0, disp = 1;
80         y = pt2010;
81         eta = b0 + b1 * pt2006 + b2 * gini + b3 * idhm_e + b5 * idhm_r + b6 * urbprop +
               b7 * pdes18m;
82         mu = exp(eta) / (1 + exp(eta));
83         phi = disp;
84         pi = constant("pi");
85         ll = log(1/sqrt(2 * pi * phi**2 * (y * (1 - y))**3)) - 0.5 / phi**2 * ((y - mu)**
               2 / (y * (1 - y) * mu**2 * (1 - mu)**2));
86         model y ~ general(ll);
87         predict mu out = simppred(rename = (pred = simppred Lower = simplower upper =
               simpupper) drop = DF StdErrPred tValue Probt alpha);
88         ods output ParameterEstimates = simpest2(drop = df alpha gradient);
89         ods output FitStatistics = simpfit2;
90         ods output FitStatistics = H0(rename = (value = llh0 ) where = (descr = '-2 Log
               Likelihood'));
91 run;
92
93 ****** Teste de especificacao RESET;
94 data simppred;
95         set simppred;
96         simppred2 = simppred**2;
97 run;
98
99
100
```

```
proc nlmixed data = simppred tech = tr update = bfgs df=999999;
        parms b0 = 0, b1 = 0, b2 = 0, b3 = 0, b5 = 0, b6 = 0, b7 = 0, disp = 1, b8 = 0;
        y = pt2010;
        eta = b0 + b1 * pt2006 + b2 * gini + b3 * idhm_e + b5 * idhm_r + b6 * urbprop +
            b7 * pdes18m + b8 * simppred2;
        mu = exp(eta) / (1 + exp(eta));
        phi = disp;
        t = mu * phi;
        w = (1 - mu) * phi;
        ll = lgamma(phi) - lgamma(t) - lgamma(w) + (t - 1) * log(y) + (w - 1) * log(1 - y
            );
        model y ~ general(ll);
        ods output ParameterEstimates = TRESET(drop = df alpha gradient lower upper);
run;
```