UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

DILTOMAR SOUZA ALELUIA

MODELO DE REGRESSÃO BETA INFLACIONADO APLICADO A DADOS DE ABSENTEÍSMO POR LICENÇA MÉDICA

Salvador
2015

DILTOMAR SOUZA ALELUIA

MODELO DE REGRESSÃO BETA INFLACIONADO APLICADO A DADOS DE ABSENTEÍSMO POR LICENÇA MÉDICA

Projeto apresentado ao Curso de Graduação em Estatística do Departamento de Estatística e Instituto de Matemática da Universidade Federal da Bahia, como requisito para aprovação na disciplina de Trabalho de Conclusão de Curso II.

Orientador: Profa. Dra. Verônica Maria Cadena Lima

Salvador
2015

SUMÁRIO

Lista de ilustrações

Lista de tabelas

1 INTRODUÇÃO

Em muitas situações, existe a necessidade de modelar dados na forma de taxas ou proporções, como, por exemplo, a taxa de homicídios em certa cidade, observada num determinado período, ou a proporção de alunos aprovados em dada disciplina, ou ainda, a proporção de crianças que participam de certo projeto social em determinada comunidade.

Quando a variável resposta é medida de forma contínua no intervalo $(0, 1)$, o modelo de regressão linear não é o mais indicado a ser usado, pois, podem surgir vários problemas na modelagem deste dados. Por exemplo, existe a possibilidade de se observar valores no processo de predição que extrapolam o intervalo $(0, 1)$. Para contornar este problema, é muito comum a utilização de transformações na variável dependente. Entretanto, a interpretação dos resultados pode se tornar uma tarefa difícil.

Ferrari & Cribari-Neto (2004) propõem o modelo de regressão beta para modelar variáveis distribuídas de forma contínua no intervalo (0,1). Porém, dados na forma de taxa ou proporção podem conter valores iguais a zeros e/ou uns. Ospina & Ferrari (2010) propõem a distribuição beta inflacionada em zero e/ou um que permite modelar dados que assumem valores no intervalo [0,1), (0,1] ou [0,1]. Para dados observados no intervalo [0,1), (0,1], os autores utilizam uma mistura da distribuição Beta e uma distribuição degenerada que atribui probabilidade não-negativa a 0 ou 1, dependendo do caso. Para dados observados em [0,1], uma mistura das distribuições Beta e Bernoulli é usada. Além disso, Ospina & Ferrari (2012) propõem o modelo de regressão beta inflacionado para modelar dados no intervalo $[0, 1)$ ou no intervalo $(0, 1]$, que é uma extensão do modelo de regressão beta proposto por Ferrari & Cribari-Neto (2004). O modelo de regressão beta inflacionado inclui um submodelo para a probabilidade de que a variável dependente seja igual a um dos limites do intervalo zero ou um. Os parâmetros de média, de precisão e da probabilidade de um ponto de massa em zero ou em um são relacionados a preditores lineares ou não-lineares através de funções de ligação.

Na literatura, a ideia de misturar uma distribuição degenerada em zero e/ou um com uma distribuição contínua não é nova. Observa-se que modelos inflacionados foram recebendo nomenclaturas diferentes de acordo com qual extremo do intervalo $(0, 1)$ está presente nos dados e a distribuição a ser utilizada. Por exemplo, quando o extremo zero está presente nos dados e a distribuição Binomial é utilizada, a nomenclatura é *ZIB - zero-inflated binomial model*; para o extremo zero e a distribuição Poisson, a nomenclatura utilizada é *ZIP - zero-inflated Poisson model*; para zero e a distribuição Binomial Negativa, *ZINB - zero-inflated negative binomial model.* Uma revisão da literatura dos modelos ZIP e ZINB é encontrada no trabalho de Ridout, Demétrio & Hinde (1998).

O objetivo deste trabalho é estudar o modelo de regressão beta inflacionado e investigar se este modelo é adequado para analisar dados sobre absenteísmo com licença

médica dos trabalhadores de uma empresa de petróleo. Este trabalho está organizado da seguinte forma. Na Seção 2.1 é apresentada a distribuição Beta, sua parametrização convencional e a parametrização adotada por Ferrari & Cribari-Neto (2004). Em seguida, na Seção 2.2, apresentamos a distribuição Beta inflacionada em zero e/ou um e algumas de suas propriedades. O modelo de regressão Beta inflacionado e suas propriedades são apresentados na Seção 2.3. No Capítulo 3, o modelo de regressão Beta inflacionado é aplicado a dados de absenteísmo com licença médica de uma coorte de trabalhadores de uma empresa de petróleo.

A avaliação numéria realizada no presente trabalho foi realizada em computador utilizando o sistema operacional *Windows 7*, *software* estatístico *R* na versão 3.1.1 com o uso do pacote *GAMLSS*. *GAMLSS* (*Generalized Additive Model for Location, Scale and Shape*) é uma ampla família de modelos de regressão conhecidos como modelos aditivos generalizados para locação, escala e forma.

Queremos salientar que as funções utilizadas para produção dos gráficos apresentados no capítulo Aplicação foram cedidas pelo Professor Dr. Raydonal Ospina Martínez do Departamento de Estatística da Universidade Federal de Pernambuco.

2 METODOLOGIA

2.1 A DISTRIBUIÇÃO BETA

Dizemos que uma variável aleatória Y possui distribuição Beta com parâmetros $a, b > 0$, se sua função de densidade é dada por

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1},\ y \in (0,1), \tag{2.1}$$

em que $\Gamma(.)$ é a função gama. A distribuição Beta, apresentada em (2.1), possui esperança e variância dadas respectivamente por:

$$E(Y) = \frac{a}{a+b} \tag{2.2}$$

e

$$Var(Y) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.3}$$

A equação em (2.1) é a forma usual de apresentar a distribuição Beta, entretanto com o objetivo de definir um modelo de regressão para variáveis aleatórias com distribuição Beta, Ferrari e Cribari-Neto (2004) propõem a seguinte parametrização. Seja $\mu = a/(a+b)$ e $\phi = a+b$, i.e, $a = \mu\phi$ e $b = (1-\mu)\phi$. Com isso, $E(Y) = \mu$ e a $Var(Y) = \mu(1-\mu)/(1+\phi)$. Neste caso, observa-se que o parâmetro μ representa a média e ϕ pode ser interpretado como o parâmetro de precisão, i.e., quanto maior for o valor de ϕ menor será o valor da variância.

Com esta nova parametrização, a função de densidade de Y pode ser escrita como:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, y \in (0,1) \tag{2.4}$$

para $0 < \mu < 1$, $\phi > 0$.

2.2 DISTRIBUIÇÃO BETA INFLACIONADA

Na prática, dados na forma de proporção podem apresentar valores zeros e/ou uns. Para modelar dados desta forma, Ferrari & Ospina (2010) propõem a distribuição beta inflacionada com a qual é possível modelar dados que estejam contidos nos intervalos [0,1), (0,1] ou ainda [0,1]. Para dados observados no intervalo [0,1) ou (0,1], os autores utilizam uma mistura de uma distribuição contínua em (0,1) e uma distribuição degenerada, que atribui probabilidade não-negativa a 0 ou 1. A função de densidade de probabilidade da distribuição beta inflacionada em c, denotada por bi_c, é dada por:

$$bi_c(y; \alpha, \mu, \phi) = \begin{cases} \alpha, & \text{se } y = c \\ (1-\alpha) f(y; \mu, \phi), & \text{se } y \in (0,1), \end{cases} \tag{2.5}$$

em que $0 < \alpha, \mu < 1$, $\phi > 0$, $f(y; \mu, \phi)$ é a função de densidade em (2.4) com seus respectivos parâmetros μ e ϕ; e α é a massa de probabilidade em c e representa a probabilidade de se observar $(c = 0)$ ou $(c = 1)$.

A função em (2.5) é conhecida como distribuição Beta inflacionada, porém, vale ressaltar que ela é inflacionada em $c = 0$ ou $c = 1$ e não nos dois ao mesmo tempo. Para os dados no intervalo [0,1], a mistura será entre a distribuição Beta para a região contínua dos dados e a distribuição de Bernoulli para os valores zero e um. Mais sobre esta distribuição pode ser visto em Ferrari & Ospina (2010).

Seja Y uma variável aleatória com distribuição Beta inflacionada com densidade definida em (2.5). Se $c = 0$, a densidade da distribuição em (2.5) é chamada distribuição Beta inflacionada em zero (em inglês, *zero-inflated beta distribution* - BEZI). A notação usual é, $Y \sim \text{BEZI}(\alpha, \mu, \phi)$. Se $c = 1$, a densidade da distribuição em (2.5) é chamada distribuição Beta inflacionada em um (em inglês, *one-inflated beta distribution* - BEOI), e é denotada por $Y \sim \text{BEOI}(\alpha, \mu, \phi)$.

A distribuição BEZI possui esperança e variância dadas por

$$E(Y) = (1 - \alpha)\mu$$

e

$$Var(Y) = (1 - \alpha)\frac{V(\mu)}{\phi + 1} + \alpha(1 - \alpha)\mu^2,$$

respectivamente. Para a distribuição BEOI, a média e a variância são

$$E(Y) = \alpha + (1 - \alpha)\mu$$

e

$$Var(Y) = (1 - \alpha)\frac{V(\mu)}{\phi + 1} + \alpha(1 - \alpha)(1 - \mu)^2,$$

respectivamente, em que $V(\mu) = \mu(1 - \mu)$.

Na Figura 1 são apresentados os gráficos da distribuição Beta inflacionada em zero para diferentes valores dos parâmetros. Observa-se que quando $\mu = 1/2$, a forma da densidade em (2.5) será simétrica, enquanto que quando $\mu \neq 1$ a densidade será assimétrica.

2.3 MODELO DE REGRESSÃO BETA INFLACIONADO

Ospina e Ferrari (2012) propõem o modelo de regressão Beta inflacionado para a modelagem de proporções e taxas quando zeros ou uns estão contidos nos dados. Apresentamos a seguir uma descrição deste modelo. Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes com distribuição Beta inflacionada dada em (2.5), com parâmetros $\alpha = \alpha_t$, $\mu = \mu_t$ e $\phi = \phi_t$, $t = 1, 2, \ldots, n$. Suponha que os parâmetros α_t, μ_t e ϕ_t satisfazem as

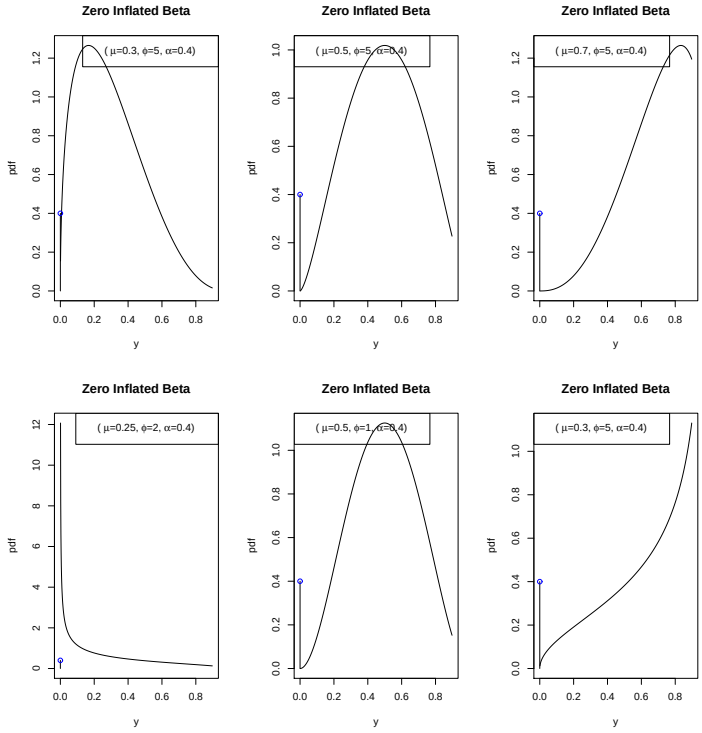

Figura 1: Densidade da distribuição Beta inflacionada em zero para diferentes valores de μ e ϕ.

seguintes relações funcionais:

$$\begin{aligned} h_1(\alpha_t) &= \eta_1 = f_1(v, \rho), \\ h_2(\mu_t) &= \eta_2 = f_2(x, \beta), \\ h_3(\phi_t) &= \eta_3 = f_3(z, \gamma), \end{aligned} \tag{2.6}$$

em que $\rho = (\rho_1, \ldots, \rho_p)^T$, $\beta = (\beta_1, ..., \beta_k)^T$ e $\gamma = (\gamma_1, ..., \gamma_m)^T$ são vetores de parâmetros de regressão desconhecidos, $(p + k + m < n)$; $v_t = (v_{t1}, ..., v_{tp'})$, $x_t = (x_{t1}, .., x_{tk'})$ e $z_t = (z_{t1}, ..., z_{tm'})$ são observações referentes as $p' + k' + m'$ variáveis explanatórias do modelo; $\eta_1 = (\eta_{11}, \ldots, \eta_{1n})^T$, $\eta_2 = (\eta_{21}, \ldots, \eta_{2n})^T$ e $\eta_3 = (\eta_{31}, \ldots, \eta_{3n})^T$ são vetores de predição. As funções $f_1(.,.)$, $f_2(.,.)$ e $f_3(.,.)$ são contínuas, lineares ou não-lineares e duas vezes diferenciáveis, de forma que, $\mathcal{V} = \partial\eta_1/\partial\rho$, $X = \partial\eta_2/\partial\beta$ e $Z = \partial\eta_3/\partial\gamma$ têm ranks p, k e m, respectivamente, para todo ρ, β e γ. Além disso, as funções de ligação $h_1 : (0, 1) \to R$, $h_2 : (0, 1) \to R$ e $h_3 : (0, \infty) \to R$ são funções estritamente monótonas e duas vezes diferenciáveis. Várias funções de ligação podem ser usadas. Para μ e α, as funções de ligação mais comuns são: logit, probit, log-log complementar e a log-log. Para o parâmetro de precisão ϕ, as funções de ligação mais utilizadas são a função log e a raiz quadrada.

2.3.1 INFERÊNCIA

A função de verossimilhança para $\theta = (\rho^T, \beta^T, \gamma^T)^T$ é obtida fazendo

$$L(\theta; y_1, y_2..., y_n) = \prod_{i=1}^{n} bi_c(y_t; \alpha_t, \mu_t, \phi_t) = L_1(\rho)L_2(\beta, \gamma),$$

em que,

$$L_1(\rho) = \prod_{i=1}^{n} \alpha_t^{\mathbb{1}_c(y_t)}(1-\alpha_t)^{1-\mathbb{1}_c(y_t)}$$

e

$$L_2(\beta, \gamma) = \prod_{t:y_t \in (0,1)} f(y_t; \mu_t, \phi_t),$$

em que $\mathbb{1}_A(y_t)$ é uma função indicadora que é igual a 1 se $y_t \in A$ e 0 se $y_t \notin A$; $\alpha_t = h_1^{-1}(\eta_{1t})$, $\mu_t = h_2^{-1}(\eta_{2t})$ e $\phi_t = h_3^{-1}(\eta_{3t})$ são funções de ρ, β e γ, respectivamente, como definido em (2.6). O log da função de verossimilhança é dado por

$$\ell(\theta) = \ell_1(\rho) + \ell_2(\beta, \gamma) = \sum_{t=1}^{n} \ell_1(\alpha_t) + \sum_{t:y_t \in (0,1)} \ell_t(\mu_t, \phi_t), \tag{2.7}$$

em que

$$\ell(\alpha_t) = \mathbb{1}_c(y_t)\log\alpha_t + (1 - \mathbb{1}_c(y_t))\log(1-\alpha_t),$$

$$\ell(\mu_t, \phi_t) = \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) +$$

$$+(\mu_t\phi_t - 1)\log(y_t) + \{(1-\mu_t)\phi - 1\}\log(1-y_t),$$

Como a função de verossimilhança $L(\theta, y_1, \ldots, y_n)$ pode ser fatorada em dois termos, os parâmetros são separáveis e a inferência sobre $(\beta^T, \gamma^T)^T$ pode ser realizada separadamente da inferência sobre ρ e vice-versa (ver Ospina e Ferrari (2012)). Da separabilidade dos vetores de parâmetros ρ e $(\beta^T, \gamma^T)^T$ é possível obter independentemente as funções escores para ρ e $(\beta^T, \gamma^T)^T$. A função escore, obtida pela diferenciação do log da verossimilhança com respeito aos parâmetros desconhecidos, é dada por

$$U(\theta) = (U_\rho(\rho)^T, U_\beta(\beta, \gamma)^T, U_\gamma(\beta, \gamma)^T)^T,$$

em que

$$U_\rho(\rho) = \mathcal{V}^T\mathcal{A}\mathcal{D}\mathcal{A}^*(y^c - \alpha),$$

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

$$U_\gamma(\beta, \gamma) = Z^T(I_n - Y^c)H[M(y^* - \mu^*) + (y^\dagger - \mu^\dagger)].$$

Aqui $y^* = (y_1^*, \ldots, y_n^*)^T$, $y^\dagger = (y_1^\dagger, \ldots, y_n^\dagger)^T$, $y^c = (\mathbb{1}_{\{c\}}(y_1), ..., \mathbb{1}_{\{c\}}(y_n))^T$, $\mu^* = (\mu_1^*, ..., \mu_n^*)^T$, $\mu^\dagger = (\mu_1^\dagger, ..., \mu_n^\dagger)^T$ e $\alpha^* = (\alpha_1^*, ..., \alpha_n^*)^T$ são vetores de dimensão $(n \times 1)$ e $M = \text{diag}(\mu_1, \ldots, \mu_n)$, $\mathcal{A} = \text{diag}(1/\alpha_1, \ldots, 1/\alpha_n)$, $\mathcal{A}^* = \text{diag}(1/(1-\alpha_1), \ldots, 1/(1-\alpha_n))$, $D = \text{diag}(1/h_1'(\alpha_1), \ldots, 1/h_1'(\alpha_n))$, $T = \text{diag}(1/h_2'(\mu_1), \ldots, 1/h_2'(\alpha_n))$, $H = \text{diag}(1/h_3'(\phi_1), \ldots,$

$1/h_3'(\phi_n))$, $\Phi = \text{diag}(\phi_1, \ldots, \phi_n)$ e $Y^c = (\mathbb{1}_{\{c\}}(y_1), ..., \mathbb{1}_{\{c\}}(y_n))^T$ são matrizes diagonais de dimensão $(n \times n)$. Além disso, I_n representa a matrix identidade $(n \times n)$. Os estimadores de máxima verossimilhança de ρ e $(\beta^T, \gamma^T)^T$ são obtidos como solução dos sistemas não-lineares $U_\rho(\rho) = 0$ e $(U_\beta(\beta, \gamma)^T, U_\gamma(\beta, \gamma)^T) = 0$. Observa-se que, como não existem expressões em fórmulas fechadas para estes estimadores, as estimativas dos parâmetros devem ser obtidas pela maximização do log da função de verossimilhança através de um algoritmo de otimização não-linear.

Para obter os erros-padrão assintóticos dos estimadores de máxima verossimilhança é necessário calcular a inversa da matriz de informação de Fisher. Ospina & Ferrari (2012) mostram que a inversa da matriz de informação é dada por:

$$\begin{pmatrix} K^{\rho\rho} & 0 & 0 \\ 0 & K^{\beta\beta} & K^{\beta\gamma} \\ 0 & K^{\gamma\beta} & K^{\gamma\gamma} \end{pmatrix}.$$

em que
$K^{\rho\rho} = (\mathcal{V}^T\mathcal{W}_1\mathcal{V})^{-1}$, $K^{\beta\beta} = \{\mathcal{X}^T(\mathcal{W}_2 - \mathcal{W}_3\mathcal{Z}(\mathcal{Z}^T\mathcal{W}_4\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{W}_3)\mathcal{X}\}^{-1}$, $K^{\beta\gamma} = (K^{\beta\phi})^T = -(\mathcal{Z}^T\mathcal{W}_4\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{W}_3\mathcal{X}(\mathcal{X}^T(\mathcal{W}_2 - \mathcal{W}_3\mathcal{Z}(\mathcal{Z}^T\mathcal{W}_4\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{W}_3)\mathcal{X})^{-1}$, e $K^{\gamma\gamma} = (\mathcal{Z}^T\mathcal{W}_4\mathcal{Z})^{-1} + (\mathcal{Z}^T\mathcal{W}_4\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{W}_3\mathcal{X}(\mathcal{X}^T(\mathcal{W}_2 - \mathcal{W}_3\mathcal{Z}(\mathcal{Z}^T\mathcal{W}_4\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{W}_3)\mathcal{X})^{-1}l\mathcal{X}^T\mathcal{W}_3\mathcal{Z}X(\mathcal{Z}^T\mathcal{W}_4\mathcal{Z})^{-1}$
com $\mathcal{W}_1 = (\mathcal{A}^* + \mathcal{A})\mathcal{D}$, $\mathcal{W}_2 = \Phi\mathcal{T}\{\mathcal{V}^*\mathcal{A}^{*-1}\mathcal{T}\Phi\}$ $\mathcal{W}_3 = \mathcal{T}\{\Phi(\mathcal{M}\mathcal{V}^* + C)\mathcal{A}^{*-1}\}\mathcal{H}$, e $\mathcal{W}_4 = \mathcal{H}\{(\mathcal{M}^2\mathcal{V}^* + 2\mathcal{M}C + \mathcal{V}^\dagger)\mathcal{A}^{*-1}\}$.

Para verificar a significância das variáveis explicativas podem ser realizados testes sobre os parâmetros do modelo. Suponha que seja de interesse do pesquisador testar um subconjunto dos vetores de parâmetros ρ, β e γ. Particionando os vetores de parâmetros $\rho = (\rho_1^T, \rho_2^T)^T$, $\beta = (\beta_1^T, \beta_2^T)^T$ e $\gamma = (\gamma_1^T, \gamma_2^T)^T$, em que $\rho_1 = (\rho_1, \ldots, \rho_{m_1})^T$, $\rho_2 = (\rho_{m_1+1}, \ldots, \rho_m)^T$, $\beta_1 = (\beta_1, \ldots, \beta_{k_1})^T$, $\beta_2 = (\beta_{k_1+1}, \ldots, \beta_k)^T$, $\gamma_1 = (\gamma_1, \ldots, \gamma_{p_1})^T$, $\gamma_2 = (\gamma_{p_1+1}, \ldots, \gamma_p)^T$, pode-se testar a hipótese $H_0 : \rho_1 = \rho_1^{(0)}; \beta_1 = \beta_1^{(0)}; \gamma_1 = \gamma_1^{(0)}$ versus a hipótese H_1 : violação de pelo menos uma igualdade. A estatística de teste da razão de verossimilhanças é

$$\Lambda = 2\ell(\hat{\rho}, \hat{\beta}, \hat{\gamma}) - \ell(\tilde{\rho}, \tilde{\beta}, \tilde{\gamma}), \tag{2.8}$$

em que $\ell(\rho, \beta, \gamma)$ é a função de log-verossimilhança em (2.7) e $(\tilde{\rho}^T, \tilde{\beta}^T, \tilde{\gamma}^T)^T$ é o estimador de máxima verossimilhança restrito de $(\rho^T, \beta^T, \gamma^T)^T$, obtido pela imposição da hipótese nula. Sob condições usuais de regularidade, (2.8) converge em distribuição para uma variável aleatória com distribuição qui-quadrado com $(m_1 + k_1 + p_1)$ graus de liberdade.

2.4 CRITÉRIO DE SELEÇÃO DE MODELOS

Para o modelo de regressão Beta inflacionado os três critérios de seleção de modelos que são mais usados são: o critério de informação de Akaike (AIC), o critério de informação

bayesiano de Schwarz (BIC) e o critério consistente de informação de Akaike (CAIC).

$$AIC = -2\hat{\ell} + 2d,$$

$$BIC = -2\hat{\ell} + dlog(n),$$

$$CAIC = -2\hat{\ell} + d(log(n) + 1),$$

em que $\hat{\ell}$ representa o máximo da função de log-verossimilhança, d é o número de parâmetros do modelo e n é o número de observações. Observe que o critério AIC é o único que não utiliza o valor de n. O modelo escolhido como mais adequado será o que apresentar o menor valor do critério empregado.

2.5 ANÁLISE DE RESÍDUOS

A análise de resíduos é de muita utilidade para analisar um possível afastamento das suposições assumidas de um modelo ajustado, detectando pontos extremos no conjunto de dados e analisando seu impacto nos resultados inferenciais que possam prejudicar a bondade do ajuste.

2.5.1 Resíduos padronizados

Para o estudo dos resíduos no modelo de regressão Beta inflacionado, Ospina & Ferrari (2012) sugerem, primeiramente, que o estudo do componente discreto e do componente contínuo sejam realizados separadamente utilizando o resíduo padronizado de Pearson. Posteriormente, usando informações do componente discreto e do componente contínuo simultaneamente é definido um resíduo quantil aleatorizado como um resíduo global para o modelo. Apesar dos resíduos padronizados serem distribuídos de forma assimétrica, o que torna a tarefa do uso dos métodos de diagnóstico muito mais difícil, pode-se fazer uso de gráficos que confrontam os resíduos padronizados do componente discreto *versus* os valores ajustados de $\hat{\alpha}_t$ e os resíduos padronizados do componente contínuo *versus* os valores ajustados de $\hat{\mu}_t$, que têm como ponto de corte o intervalo (-4, 4), (Ospina & Ferrari, 2012). Estes gráficos são extremamente úteis para identificar valores discrepantes em cada componente. Baseado no algoritmo interativo score de Fisher, define-se os seguintes resíduos padronizados de Pearson para o modelo de regressão Beta inflacionado em zero ou um:

$$r_{pt}^{(1)} = \frac{\mathbb{1}_{\{c\}}(y_t) - \hat{\alpha}_t}{\sqrt{\hat{\alpha}_t(1-\hat{\alpha}_t)(1-\hat{h}_{1tt}^*)}}$$

para o componente discreto, e para o componente contínuo temos

$$r_{pt}^{(2)} = \frac{y_t^* - \hat{\mu_t^*}}{\sqrt{\hat{v}_t(1-\hat{\alpha}_t)(1-\hat{h}_{2tt}^*)}},$$

em que $\hat{h}^*_{1tt}$ e $\hat{h}^*_{2tt}$ são os t-ésimos componentes da diagonal principal da matriz de projeção,

$$\hat{H}^*_1 = \hat{Q}^{1/2} Z (Z^T \hat{Q} Z)^{-1} Z^T \hat{Q}^{1/2}$$

e

$$\hat{H}^*_2 = \hat{W}^{1/2}_{\beta\beta} X (X^T \hat{W}_{\beta\beta} X)^{-1} X^T \hat{W}^{1/2}_{\beta\beta},$$

respectivamente;

$$y^*_t = \begin{cases} log(\frac{y_t}{(1-y_t)}), & \text{se } y_t \,\epsilon\, (0,1) \\ 0, & \text{cc,} \end{cases}$$

e

$$\mu^*_t = E(Y^*_t | \mathbb{1}_{\{c\}}(y_t) = 0) = \psi(\mu_t \phi) - \psi((1-\mu_t)\phi).$$

As matrizes Z $(n \times m)$ e X $(n \times k)$ são matrizes de valores fixos e conhecidos cujas t-ésimas linhas são $z_t = (z_{t1}, ..., z_{tm})$ e $x_t = (x_{t1}, .., x_{tk})$ respectivamente, $Q = diag\{q_1, ..., q_n\}$ e $W = diag\{w_1, ..., w_n\}$.

Assim, um resíduo padronizado global para o modelo é dado da seguinte forma:

$$r_{pt} = \begin{cases} r^{(1)}_{pt}, & \text{se } y_t = c \\ r^{(2)}_{pt}, & \text{se } y_t \in (0,1), \end{cases}$$

2.5.2 Resíduos quantis aleatorizados

Para avaliar a adequação/ajuste do modelo de regressão Beta inflacionado, Ospina & Ferrari (2012) propõem o resíduo quantil aleatorizado dado por

$$r^q_t = \Phi^{-1}(u_t), t = 1, ..., n, \tag{2.9}$$

em que $\Phi(.)$ denota a função de distribuição acumulada da distribuição normal padrão e u_t é uma variável aleatória que se encontra no intervalo $(a, b]$ com $a = lim_{y \to y_t} BI_c(y; \hat{\alpha}, \hat{\mu}, \hat{\phi})$ e $b = BI_c(y_t; \hat{\alpha}, \hat{\mu}, \hat{\phi})$.

2.6 PONTOS INFLUENTES

No processo de diagnóstico nos modelos de regressão, a verificação de pontos influentes é de suma importância, pois tais pontos podem reduzir significativamente a qualidade do modelo ajustado.

Baseando-se na curva normal, uma das análises de pontos influentes mais utilizada é a distância de Cook. Esta distância pode ser assim descrita: Seja θ um vetor com os parâmetros a serem estudados, ω a influência que uma perturbação pode provocar no vetor de parâmetros θ e $\ell(\theta/\omega)$ a função log-verossimilhança do modelo perturbado. O grau da pertubação exercido por uma observação pode ser medido por $F(\omega) = 2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\omega)\}$, em que $\ell(\hat{\theta}_\omega)$ é a função de log-verossimilhança do estimador restrito e $\ell(\hat{\theta})$ a função de log-verossimilhança do modelo não perturbado. Para $F(\omega_0)$, em que ω_0 é o ponto de

pertubação nula, podemos obter a direção da maior mudança com o cálculo da curvatura normal na direção d da superfície do deslocamento pela verossimilhança com

$$C_d(\theta) = 2|d^T \Delta^T \ell^{-1} \Delta d|, \tag{2.10}$$

em que $\Delta = \partial^2 \ell(\theta/\omega)/\partial\theta\partial\omega^T|_{\theta=\hat{\theta},\omega=\omega_\theta}$, $\ell = \partial^2\ell(\theta)/\partial\theta\partial\theta^T|_{\theta=\hat{\theta}}$, e d é um vetor unitário.

Para avaliar a influência de apenas uma parte do vetor de parâmetros θ, tal que $\theta = (\theta_1^T, \theta_2^T)^T$, a curva normal na direção d será dada da seguinte forma

$$C_d(\theta_1) = |d^T \Delta^T (\ell^{-1} - \ell_{22}) \Delta d|,$$

em que

$$\ell_{22} = \begin{pmatrix} 0 & 0 \\ 0 & \ell^{-1}_{\theta_2\theta_2} \end{pmatrix}.$$

e $\ell_{\theta_2\theta_2} = \partial^2\ell(\theta)/\partial\theta_2\partial\theta_2^T$.

Para avaliar a curvatura normal na direção do t-ésimo indivíduo, temos

$$C_t = 2|\Delta_t^T \ell^{-1} \Delta_t|$$

que é a influência local total do t-ésimo indivíduo, em que Δ_t é a t-ésima coluna da matriz Δ. Para (2.10) podemos encontrar a direção da máxima curvatura normal usando as direções que produzem as maiores mudanças locais no deslocamento na verossimilhança em $F(\omega_0)$. Assim, utilizaremos d_{max} e obteremos

$$C_{max} = \max d\{C_d(\theta)\}.$$

É de grande utilidade observarmos os gráficos de C_t contra os índices das observações para encontrarmos pontos que possam influenciar na qualidade do ajuste do modelo.

3 APLICAÇÃO

Neste capítulo, vamos considerar uma aplicação do modelo de regressão Beta inflacionado a um conjunto de dados reais. Os dados foram coletados através de um estudo de coorte retrospectivo com todos os trabalhadores de uma empresa de petróleo no período de 1 de janeiro de 2007 a 31 de dezembro de 2009. No final, informações sobre 776 trabalhadores foram obtidas dos seus prontuários eletrônicos.

O interesse é identificar os principais fatores associados ao absenteísmo com licença médica, o qual pode ser entendido como a ausência do indivíduo ao trabalho, seja por motivo de doença, acidente ou por atendimento médico e justificado por licença médica (Oenning, Carvalho & Lima 2014). A variável resposta é o índice médio de faltas ao trabalho com licença médica no período (IMFalta), a qual é obtida pela razão entre o total de dias com licença médica e os dias potencialmente trabalháveis no período. Estes dados foram cedidos pela pesquisadora Nágila Soares Xavier Oenning, e foram usados em sua dissertação de mestrado intitulada "Absenteísmo com licença médica em uma coorte de trabalhadores da área de serviços de uma indústria de petróleo" (Oenning, 2011).

As covariáveis da pesquisa encontram-se descritas na Tabela 1. Algumas delas foram consideradas quanto à sua definição no início da coorte, tais como: sexo, idade, tempo de atuação, etc. Outras variáveis foram classificadas de forma a apresentar sua evolução no período. Mais informações sobre a coleta e definição das variáveis podem ser encontradas em Oenning (2011).

Na Tabela 2 encontram-se algumas medidas resumo das variáveis contínuas envolvidas no estudo. Através desta tabela, podemos notar que a variável resposta apresenta-se no intervalo que vai de zero até 0,89750 e 50% dos trabalhadores apresentaram índice médio de faltas no período entre 0% e 3,35%. A idade média dos trabalhadores é de 43,57 anos (desvio-padrão = 8,51 anos), idade mínima de 21,6 anos e máxima 71,1 anos. O tempo médio de atuação na empresa é de 17,81 anos (dp = 9,79 anos) e 25% dos indivíduos apresentam tempo de atuação superior a 24,45 anos. Pelo menos, 75% dos trabalhadores apresentam níveis de glicose dentro do padrão (Valor de Referência: 70 a 99 mg/dL) e mais de 50% deles apresentam excesso de peso (IMC$>$25 Kg/m^2).

Na Tabela 3 encontram-se a frequência e a porcentagem das variáveis qualitativas do estudo. Através desta tabela, pode-se observar que a maioria dos indivíduos pesquisados é do sexo masculino (76%), trabalha em regime administrativo e de sobreaviso (56%), trabalha como inspetor de segurança interna (44%), está satisfeito no trabalho (99%), trabalha com atenção concentrada (71%) e apresenta um bom relacionamento com a chefia (99%). Além disso, 75% dos indivíduos relataram não-fumar, 82% não desenvolvem atividade física regular e a maioria apresenta sono normal (85%). Observa-se ainda que apenas 7% dos indivíduos foram classificados como levemente hipertensos e 13% hiper-

Tabela 1: Descrição das variáveis

Descrição da variável	**V**ariável	**C**ategorias/Valores
Incidência média de faltas com licenças médicas no período	IMFalta	
Atividade física	AF	0-Muito Ativo, Ativo e Regu. Ativo; 1-Irreg. Ativo, Fisicamente Inativo, Sedentário
Atenção concentrada no trabalho	ATCONT	0-Não; 1-Sim
Denominação do cargo correspondente	CARGO	0-Outros; 1-Téc. de adm. e controle 2-Inspetor de segurança
Doenças do sistema cardiovascular	DCARDVAS	0-Não; 1-Sim
Diabetes não controlada	DIANCONTROL	0-Não; 1-Sim
Doenças do sistema digestivo	DSISDIG	0-Não; 1-Sim
Doenças do sistema locomotor	DSISLOCO	0-Não; 1-Sim
Doenças do sistema neurológico	DSNEURO	0-Não; 1-Sim
Glicemia em jejum	GLI	em mg/dL
Hipertensão arterial	HAS	0-Normal e Normal Limítrofe; 1-Hipertensão Leve; 2-Hipertenso
Idade	IDADE	em anos
Índice de massa corpórea	IMC	em kg/m^2
Neoplasias	NEOPLASIA	0-Não; 1-Sim
Posturas forçadas no trabalho	POSFOR	0-Não; 1-Sim
Risco coronariano	RC	0-Baixo; 1-Moderado e alto
Bom relacionamento com a chefia?	RELCHEF	0-Sim; 1-Não;
Satisfação no trabalho	SATISF	0-Sim; 1-Não;
Referência de gênero	SEXO	0-Masculino; 1-Feminino
Qualidade do sono	SONO	0-Normal; 1-Anormal
Regime de Trabalho	RT	0-Administrativo e Sobreaviso 1-Turno de 8 horas ou de 12 horas
Tempo de atuação	TATUA	em anos
Tabagismo	TAB	0-Não-fumante=0; 1-ex-fumante, 2-fumante

Tabela 2: Medidas resumo das variáveis contínuas

Estatísticas	**IMFalta**	**IDADE**	**TATUA**	**GLI**	**IMC**
MÍNIMO	0,0000	21,60	1,30	55,50	17,26
Q1	0,0000	36,90	5,23	82,30	24,23
MEDIANA	0,0081	44,60	20,64	87,30	26,72
MÉDIA	0,0326	43,57	17,81	93,84	27,21
Q3	0,0325	49,30	24,45	95,43	29,79
MÁXIMO	0,8975	71,10	38,76	343,70	44,20
DESVIO PADRÃO	0,0721	8,51	9,79	28,00	4,29

tensos, 13% apresentaram risco coronariano moderado a alto e apenas 1% apresentaram neoplasias. Por fim, observa-se também que 1% dos indivíduos relataram trabalhar com posturas forçadas, 1% apresentaram diabetes não controlada, 5% doenças do sistema digestivo, 9% doenças do sistema cardiovascular, 5% doenças do sistema locomotor e apenas 1% deles relataram doenças do sistema neurológico.

As variáveis satisfação no trabalho (SATISF), relacionamento com a chefia (RELACHEF), diabetes não controladas (DIANCONTROL), neoplasias (NEOPLASIA) e doenças do sistema neurológico (DSNEURO) foram excluídas da análise subsequente devido ao baixo número de observações em alguma das suas categorias.

A seguir apresentamos os box-plots para algumas variáveis qualitativas *versus* a variável resposta. Na Figura 2, observamos que a variável IMFalta apresenta uma maior

Tabela 3: Medidas resumo das variáveis qualitativas

Variável	**Categoria**	**Frequência**	**(%)**
SEXO	Feminino	189	24
	Masculino	587	76
RC	Baixo	675	87
	Moderado e alto	101	13
RT	Administrativo e Sobreaviso	437	56
	Turno de 8 horas e de 12 horas	339	44
HAS	Normal e Normal Limítrofe	623	80
	Leve hipertensão	51	7
	Hipertenso	102	13
CARGO	Tec. Administração e Controle	161	21
	Inspetor de segurança interno	338	44
	Outros	277	35
TABAGISMO	Ex-fumante	139	18
	Fumante	55	7
	Não-fumante	582	75
AF	Muito Ativo, Ativo e Regu. Ativo	141	18
	Irreg. Ativo, Fisicamente Inativo, Sedentário	635	82
SONO	Normal	661	85
	Anormal	115	15
SATISF	Sim	772	99
	Não	4	1
RELCHEF	Sim	774	99
	Não	2	1
ATCON	Não	228	29
	Sim	548	71
POSFOR	Não	766	99
	Sim	10	1
DIANCONTROL	Não	772	99
	Sim	4	1
DCARDVAS	Não	709	91
	Sim	67	9
DSISDIG	Não	736	95
	Sim	40	5
NEOPLASIA	Não	769	99
	Sim	7	1
DSISLOCO	Não	741	95
	Sim	35	5
DSNEURO	Não	774	99
	Sim	2	1

dispersão na categoria feminina. Além disso, também nesta categoria observamos o maior índice de faltas ao trabalho de 89,75%.

Na Figura 3, observamos que, nas três categorias da variável Cargo, as caixas dos gráficos são similares e todas mostram observações discrepantes.

Na Figura 4, observamos que na categoria "Irregularmente ativo, Fisicamente inativo, Sedentário"parece haver mais valores discrepantes, porém a distribuição da proporção de faltas é semelhante para as duas categorias.

Na Figura 5, observamos que a categoria sono Normal apresenta um percentual maior de trabalhadores com valores de IMFalta próximos de zero, indicando que estes trabalhadores tendem a faltar menos ao trabalho.

Na Figura 6, observamos que trabalhadores que relataram não fumar parecem faltar menos ao trabalho quando comparado com as outras categorias.

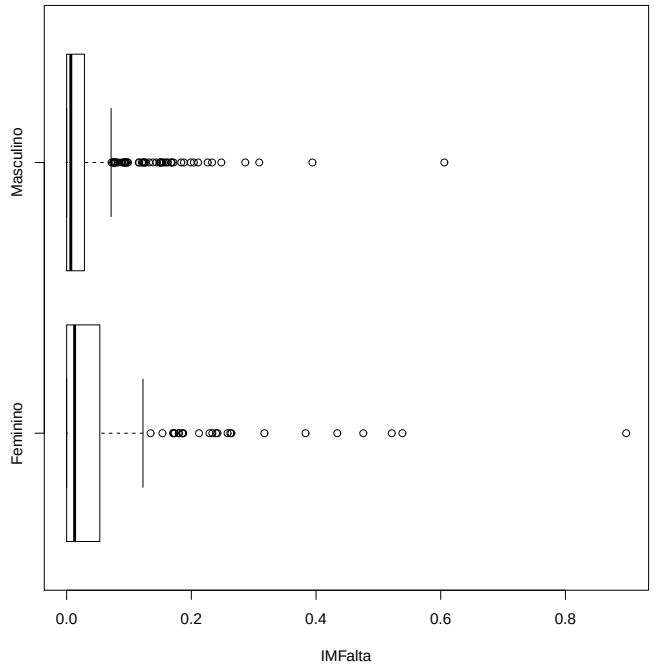

Figura 2: Box-plot da variável sexo por IMFalta.

Na Figura 7 apresentamos o histograma e o box-plot da variável resposta IMFalta. A barra vertical no histograma representa a quantidade de zeros nesta variável, que corresponde a 237, i.e., 30,54% das observações. No box-plot podemos observar a presença de vários valores discrepantes. A distribuição dos dados no intervalo (0,1) é assimétrica em forma de J invertido. Estes gráficos sugerem que o modelo de regressão Beta inflacionado em zero talvez seja um modelo conveniente para analisar estes dados.

Para a modelagem dos dados, optou-se por codificar as variáveis IMC (em 0: Baixo peso e peso normal (IMC $<$ 25), 1: Sobrepeso (25 $\leq$ IMC $<$ 30) e 2: Obeso (IMC $\geq$ 30)) e glicemia (0: GLI$<$100; 1:GLI$>$=100mg/dl). Para as variáveis que possuíam três categorias, foram utilizadas duas variáveis indicadoras para representá-las. Este foi o caso das variáveis CARGO, HAS e TAB. Por exemplo, para representar a variável CARGO foram usadas as variáveis indicadoras CARGO2 e CARGO3. A codificação utilizada para a variável CARGO foi: se CARGO = 0, então CARGO2 = 0 e CARGO3 = 0; se CARGO = 1, então CARGO2 = 1 e CARGO3 = 0; e se CARGO = 3, então CARGO2 = 0 e CARGO3 = 1. Para a variável HAS e TAB, foram criadas as variáveis indicadoras HAS2 e HAS3 e TAB2 e TAB3, respectivamente, com codificação similar a utilizada para a variável CARGO.

Inicialmente, foi verificada a necessidade de modelar o parâmetro de precisão ϕ. Para testar a hipótese $H_0 : \phi_1 = \phi_2 = \ldots = \phi_n = \phi$, foi realizado o teste da razão de verossimilhanças considerando para a estrutura de regressão dos parâmetros todas as covariáveis das Tabelas 2 e 3, exceto as variáveis já excluídas da análise. O valor da

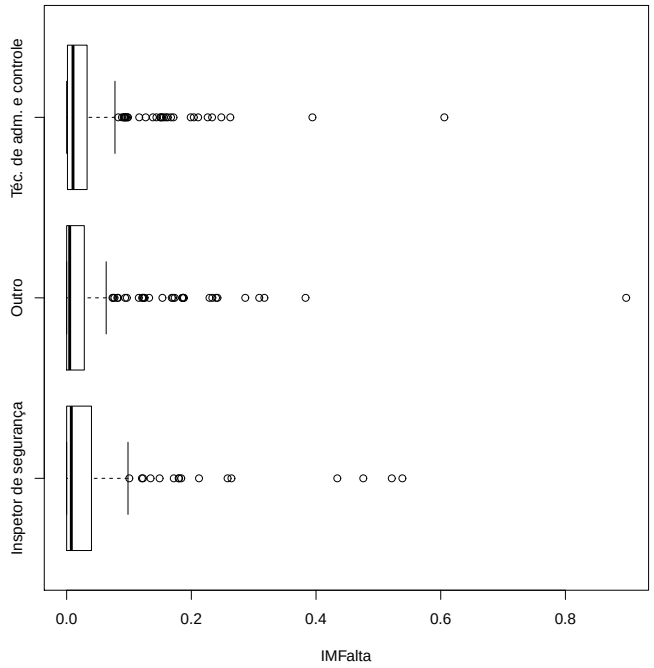

Figura 3: Box-plot da variável Cargo por IMFalta.

estatística foi $\Lambda = 129,7858$ (p-valor $= 0,0000$). Logo, rejeitamos a hipótese H_0 de que o parâmetro de dispersão seja constante.

Para o ajuste do modelo de regressão Beta inflacionado foi utilizada a função de ligação *logit* para o parâmetro μ e para o parâmetro α, e a função *log* para o parâmetro de precisão ϕ. Vale ressaltar que diferentes funções de ligação também foram usadas para modelagem destes parâmetros e foi escolhido o que apresentou o menor AIC. O modelo final ajustado foi:

$$\begin{aligned}
logit(\mu) &= \beta_0 + \beta_1 IDADE + \beta_2 GLI + \beta_3 SEXO + \beta_4 RT + \beta_5 HAS2 + \beta_6 HAS3 + \\
&\quad \beta_7 CARGO2 + \beta_8 CARGO3 + \beta_9 TAB2 + \beta_{10} TAB3 + \beta_{11} AF + \\
&\quad \beta_{12} SONO + \beta_{13} DSISDIG \\
logit(\alpha) &= \rho_0 + \rho_1 AF + \rho_2 SEXO + \rho_3 CARGO2 + \rho_4 CARGO3 \\
log(\phi) &= \gamma_0 + \gamma_1 IDADE + \gamma_2 GLI + \gamma_3 SEXO + \gamma_4 RT + \gamma_5 HAS2 + \gamma_6 HAS3 + \\
&\quad \gamma_7 CARGO2 + \gamma_8 CARGO3 + \gamma_9 TAB2 + \gamma_{10} TAB3 + \gamma_{11} AF + \\
&\quad \gamma_{12} DSISDIG.
\end{aligned}$$

Na Tabela 4 apresentamos as estimativas dos parâmetros do modelo escolhido com seus respectivos erros-padrão. Para interpretação dos parâmetros do modelo observa-se o sinal positivo ou negativo das estimativas dos parâmetros. Por exemplo, através da Tabela 4 observa-se que a variável IDADE tem um efeito positivo na probabilidade de

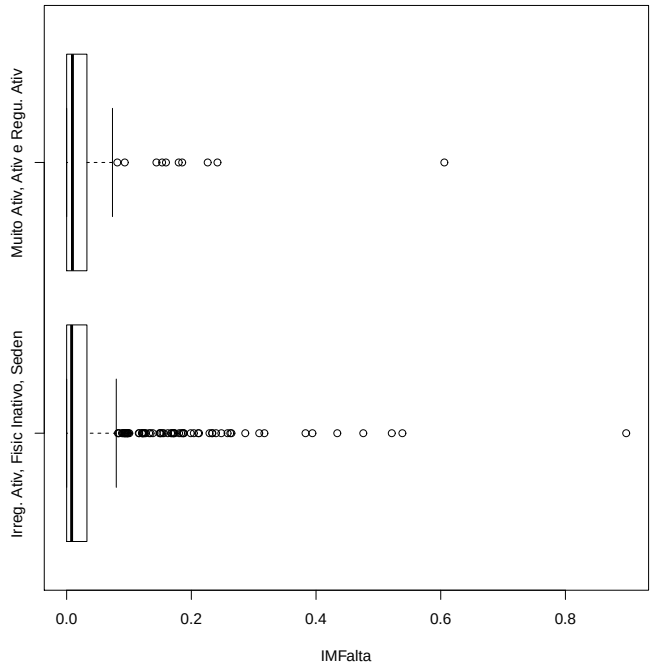

Figura 4: Box-plot da variável AF por IMFalta.

se observar no mínimo uma ausência ao trabalho por licença médica. Para a covariável SEXO observa-se que os trabalhadores do sexo feminino têm uma proporção média de faltas ao trabalho com licença médica maior quando comparado aos trabalhadores do sexo masculino.

Na Figura 8a apresentamos o gráfico dos resíduos quantis aleatorizados *versus* o índice das observações e, na Figura 8b, o gráfico do envelope simulado dos quantis aleatorizados *versus* os quantis de uma distribuição normal padrão. Ambos os gráficos não sugerem valores que possam ser considerados atípicos.

Na Figura 9a-c apresentamos os gráficos de diagnóstico para o componente discreto. O gráfico do índice das observações *versus* os resíduos padronizados (Figura 9a) sugere que os resíduos estão distribuídos aleatoriamente em torno do zero, não indicando má qualidade no ajuste do modelo, porém, apresenta a observação 742 como uma possível observação atípica. O gráfico dos valores ajustados de α *versus* os resíduos padronizados (Figura 9b) não sugere falta de ajuste. Por fim, o gráfico do índice das observações *versus* a estatística de Cook (Figura 9c), identificamos, novamente, a observação 742 como um possível valor discrepante.

Na Figura 10a-c apresentamos os gráficos para diagnóstico do componente contínuo. O gráfico do índice das observações *versus* os resíduos padronizados (Figura 10a) e o gráfico dos valores ajustados de μ *versus* os resíduos padronizados (Figura 10b) não sugerem falta de ajuste, porém a observação 619 é encontrada nos dois gráficos como um possível valor atípico. No gráfico do índice das observações *versus* a estatística de Cook (Figura

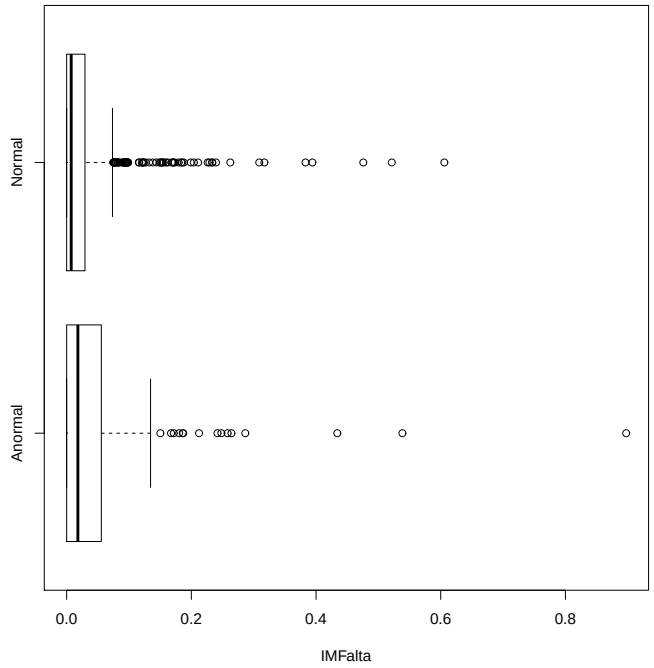

Figura 5: Box-plot da variável sono por IMFalta.

10c), observamos os valores 154 e 431 que podem ser considerados discrepantes.

Para avaliar o impacto das observações consideradas discrepantes encontradas nos gráficos 9 e 10, o modelo foi novamente ajustado após a retirada dos seguintes conjuntos de observações: { 742} e {154, 431, 619}. Nas Tabelas 5 e 6 encontram-se as colunas parâmetro estimado (Est.), erro padrão (S.e), p-valor (p), mudança relativa percentual da estimativa (Rel. est.) e mudança relativa percentual no erro padrão (Rel. s.e.) devido a exclusão da(s) observação(ões). Notamos que com a exclusão da observação 742, observa-se na Tabela 5 pequenas mudanças nos valores das estimativas do componente discreto, as quais não foram consideradas significativas. Porém, com a exclusão do conjunto de observações {154, 431, 619}, observa-se na Tabela 6, mudanças significativas nas estimativas dos parâmetros de todo o modelo. Portanto, pode-se concluir que o modelo escolhido não é adequado para modelar os dados.

Ferrari e Ospina (2012) sugerem a abordagem da quasiverossimilhança como um método alternativo para avaliar os dados. Nessa abordagem, nenhuma função de probabilidade é assumida para os dados, e sim uma relação funcional entre a média e a variância, que é especificada na forma de uma função da variância.

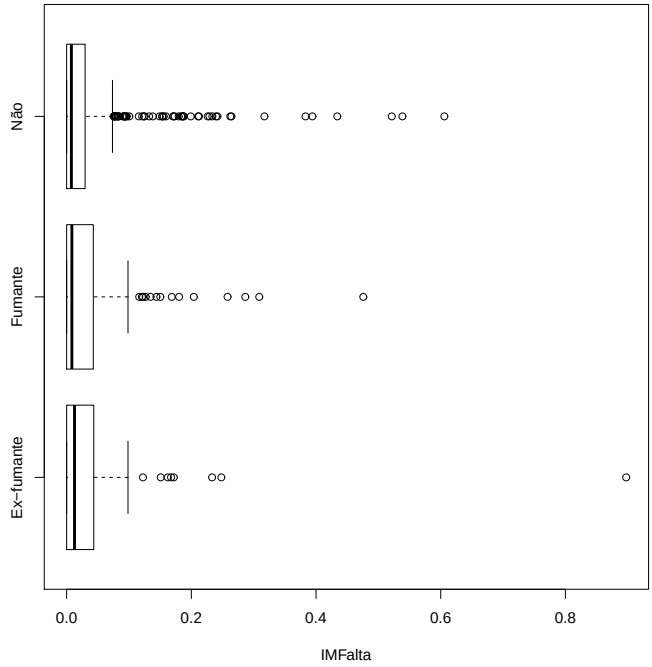

Figura 6: Box-plot da variável tabagismo por IMFalta.

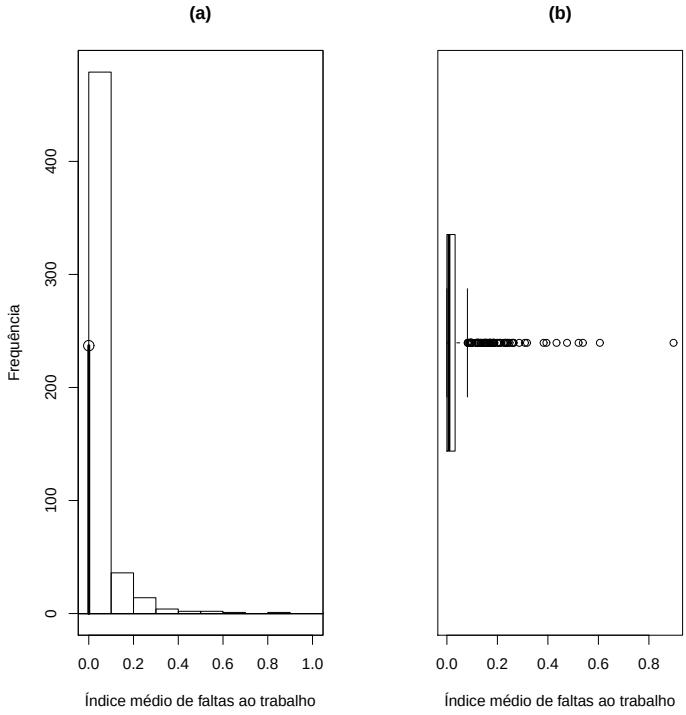

Figura 7: Histograma e box-plot da variável índice médio de faltas ao trabalho com licença médica.

Tabela 4: Estimativas e erros padrão do modelo escolhido

μ	**Estimativas**	**Erro-padrão**	**p-valor**
Intercepto	-4,419	0,2445	5,17E-61
IDADE	0,027	0,0052	1,77E-07
GLI	-0,610	0,1023	3,82E-09
SEXO	1,063	0,1099	5,72E-21
RT	-0,978	0,1041	6,32E-20
HAS2	0,054	0,1133	6,36E-01
HAS3	0,612	0,1577	1,14E-04
CARGO2	-0,158	0,1153	1,70E-01
CARGO3	1,261	0,0791	1,24E-49
TAB2	0,395	0,1038	1,54E-04
TAB3	0,616	0,1596	1,23E-04
AF	-0,360	0,1025	4,72E-04
SONO	0,284	0,1011	5,13E-03
DSISDIG	-0,763	0,1463	2,38E-07
α	**Estimativas**	**Erro-padrão**	**p-valor**
Intercepto	-0,03753	0,2137	8,61E-01
AF	-0,39522	0,1974	4,57E-02
SEXO	-0,59287	0,2105	4,97E-03
CARGO2	-0,04515	0,2155	8,34E-01
CARGO3	-0,76333	0,1895	6,19E-05
ϕ	**Estimativas**	**Erro-padrão**	**p-valor**
Intercepto	4,306	0,3468	2,22E-32
IDADE	-0,031	0,0073	2,05E-05
GLI	0,754	0,1483	4,63E-07
SEXO	-1,325	0,1485	3,51E-18
RT	16,621	0,1546	0,00E+00
HAS2	-0,092	0,1641	5,74E-01
HAS3	-0,937	0,2074	7,19E-06
CARGO2	0,200	0,1567	2,02E-01
CARGO3	-16,831	0,1221	0,00E+00
TAB2	-0,349	0,1461	1,73E-02
TAB3	-0,759	0,2118	3,62E-04
AF	0,360	0,1441	1,26E-02
DSISDIG	1,013	0,2282	1,04E-05

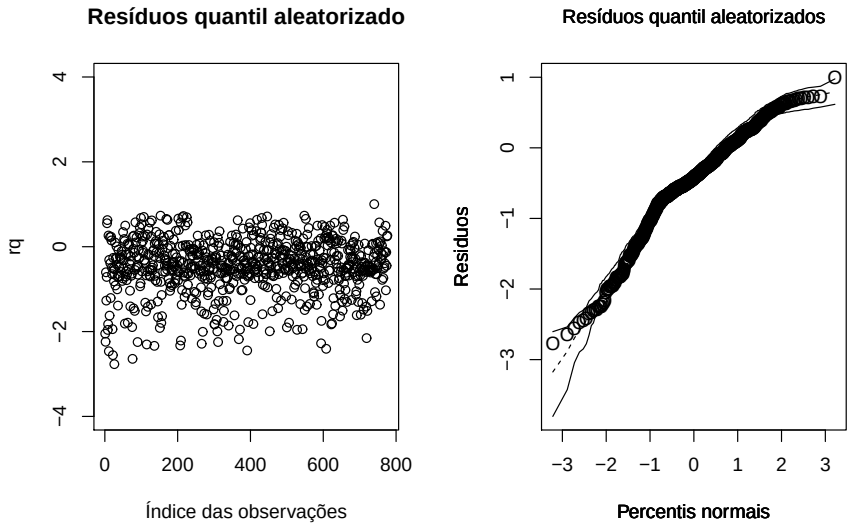

Figura 8: Gráficos dos resíduos do modelo Beta inflacionado em zero para os dados de absenteísmo.

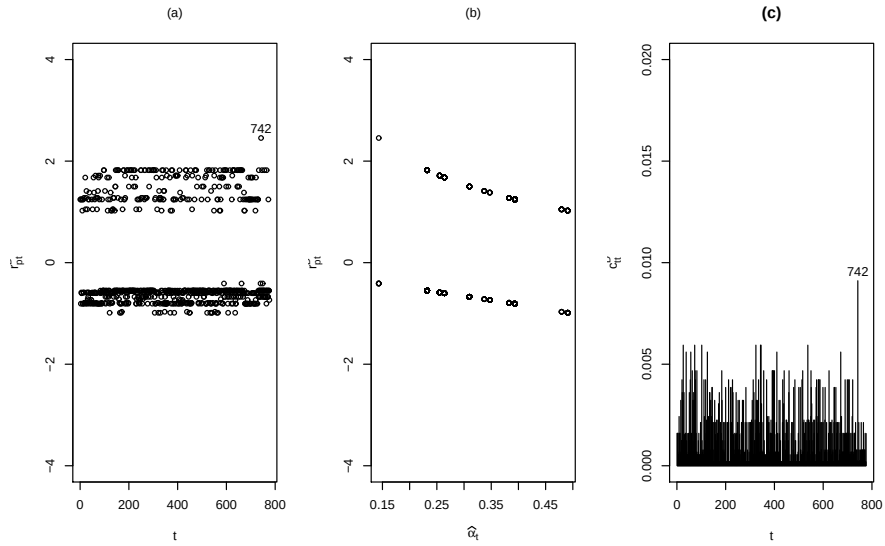

Figura 9: Gráficos de diagnóstico para o componente discreto.

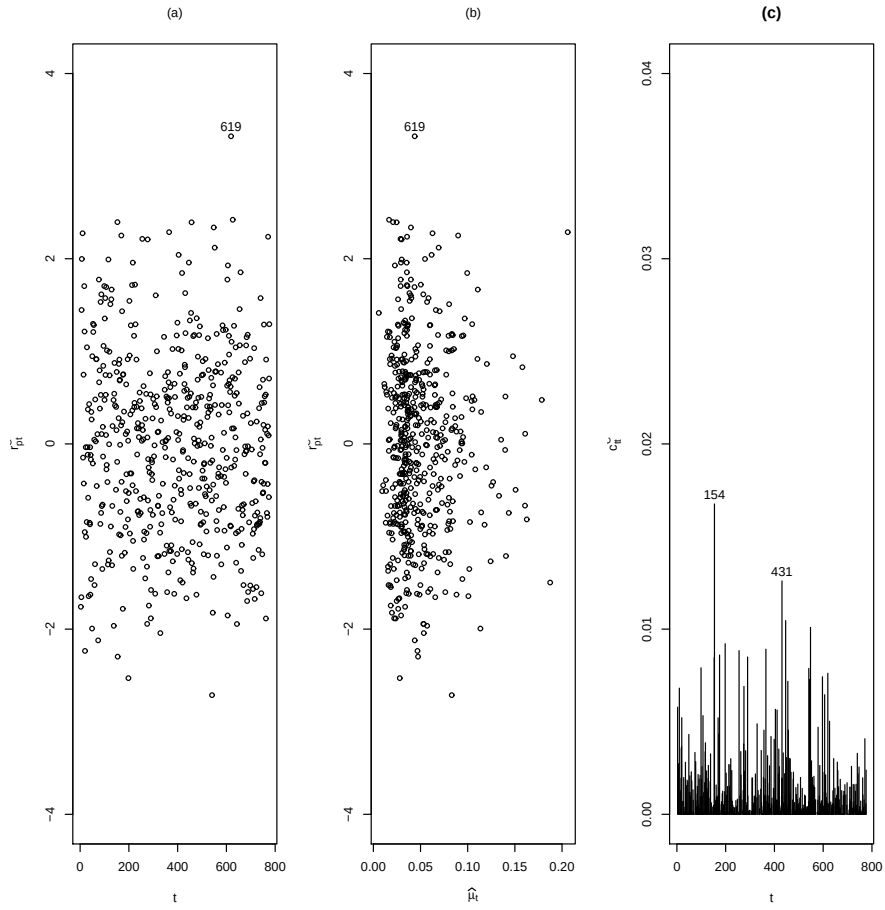

Figura 10: Gráficos de diagnóstico para o componente contínuo.

Tabela 5: Modelo com exclusão da observação 742

μ	**EST**	**S.e**	**p**	**Rel. est.**	**Rel. s.e.**
Intercepto	-4,419	0,2445	5,32E-61	0,000	0,000
IDADE	0,027	0,0052	1,77E-07	0,000	0,000
GLI	-0,610	0,1023	3,83E-09	0,000	0,000
SEXO	1,063	0,1099	5,74E-21	0,000	0,000
RT	-0,978	0,1041	6,34E-20	0,000	0,000
HAS2	0,054	0,1133	6,36E-01	0,000	0,000
HAS3	0,612	0,1577	1,14E-04	0,000	0,000
CARGO2	-0,158	0,1153	1,70E-01	0,000	0,000
CARGO3	1,261	0,0791	1,26E-49	0,000	0,000
TAB2	0,395	0,1038	1,54E-04	0,000	0,000
TAB3	0,616	0,1596	1,23E-04	0,000	0,000
AF	-0,360	0,1025	4,72E-04	0,000	0,000
SONO	0,284	0,1011	5,13E-03	0,000	0,000
DSISDIG	-0,763	0,1463	2,38E-07	0,000	0,000
α	**EST**	**S.e**	**p**	**Rel. est.**	**Rel. s.e.**
Intercepto	-0,02035	0,214	9,24E-01	-45,777	0,140
AF	-0,40217	0,1977	4,23E-02	1,759	0,152
SEXO	-0,63075	0,2117	2,98E-03	6,389	0,570
CARGO2	-0,03802	0,2158	8,60E-01	-15,792	0,139
CARGO3	-0,78865	0,1901	3,71E-05	3,317	0,317
ϕ	**EST**	**S.e**	**p**	**Rel. est.**	**Rel. s.e.**
Intercepto	4,306	0,3468	2,24E-32	0,000	0,000
IDADE	-0,031	0,0073	2,05E-05	0,000	0,000
GLI	0,754	0,1483	4,63E-07	0,000	0,000
SEXO	-1,325	0,1485	3,52E-18	0,000	0,000
RT	16,621	0,1546	0,00E+00	0,000	0,000
HAS2	-0,092	0,1641	5,74E-01	0,000	0,000
HAS3	-0,937	0,2074	7,20E-06	0,000	0,000
CARGO2	0,200	0,1567	2,02E-01	0,000	0,000
CARGO3	-16,831	0,1221	0,00E+00	0,000	0,000
TAB2	-0,349	0,1461	1,73E-02	0,000	0,000
TAB3	-0,759	0,2118	3,62E-04	0,000	0,000
AF	0,360	0,1441	1,26E-02	0,000	0,000
DSISDIG	1,013	0,2282	1,04E-05	0,000	0,000

Tabela 6: Modelo com exclusão das observações 154, 431 e 619

μ	**EST**	**S.e**	**p**	**Rel. est.**	**Rel. s.e.**
(Intercept)	-4,506	0,239	2,83E-65	1,981	-2,213
IDADE	0,024	0,005	3,44E-06	-12,842	-1,807
GLI	-0,560	0,102	4,92E-08	-8,186	-0,666
SEXO	1,077	0,109	1,20E-21	1,272	-0,567
RT	-1,016	0,104	2,10E-21	3,841	-0,343
HAS2	0,120	0,113	2,89E-01	123,074	-0,633
HAS3	0,581	0,158	2,43E-04	-5,007	-0,092
CARGO2	-0,158	0,114	1,68E-01	-0,259	-0,717
CARGO3	1,237	0,078	1,61E-49	-1,951	-1,845
TAB2	0,440	0,103	2,21E-05	11,436	-0,712
TAB3	0,596	0,160	2,13E-04	-3,357	0,264
AF	-0,126	0,096	1,91E-01	-64,889	-5,864
SONO	0,317	0,099	1,48E-03	11,744	-1,657
DSISDIG	-0,785	0,146	1,01E-07	2,927	-0,200
α	**EST**	**S.e**	**p**	**Rel. est.**	**Rel. s.e.**
(Intercept)	-0,033	0,214	8,78E-01	-12,150	0,094
AF	-0,401	0,200	4,31E-02	1,394	0,152
SEXO	-0,593	0,211	4,96E-03	0,042	0,000
CARGO2	-0,045	0,216	8,34E-01	0,045	0,000
CARGO3	-0,751	0,190	8,17E-05	-1,605	0,053
ϕ	**EST**	**S.e**	**p**	**Rel. est.**	**Rel. s.e.**
(Intercept)	4,567	0,349	2,24E-35	6,068	0,751
IDADE	-0,026	0,007	3,85E-04	-16,383	0,496
GLI	0,702	0,150	3,41E-06	-6,870	1,212
SEXO	-1,362	0,149	5,31E-19	2,857	0,272
RT	16,686	0,155	0,00E+00	0,388	0,298
HAS2	-0,226	0,165	1,69E-01	145,435	0,243
HAS3	-0,940	0,210	8,87E-06	0,175	1,197
CARGO2	0,209	0,157	1,84E-01	4,382	0,232
CARGO3	-16,758	0,123	0,00E+00	-0,431	0,709
TAB2	-0,437	0,146	2,91E-03	25,359	0,153
TAB3	-0,664	0,219	2,49E-03	-12,434	3,387
AF	-0,122	0,148	4,08E-01	-133,945	2,540
DSISDIG	1,001	0,232	1,86E-05	-1,168	1,816

4 CONCLUSÃO

No presente trabalho foram revisadas a distribuição de probabilidade Beta e algumas de suas propriedades. Foi mostrada também a parametrização proposta por Ferrari & Cribari-Neto (2004), que teve como objetivo definir um modelo de regressão para variáveis aleatórias com distribuição Beta. Em seguida, a distribuição Beta inflacionada, proposta por Ferrari & Ospina (2010), e o modelo de regressão Beta inflacionado, proposto por Ospina & Ferrari (2012), foram apresentados.

Para verificar a bondade de ajuste do modelo de regressão Beta inflacionado verificamos que é muito importante que seja realizada uma análise dos resíduos do modelo ajustado. Esta análise foi realizada através do estudo dos componentes discreto e contínuo, separadamente e simultaneamente. A verificação de pontos influentes foi realizada através de influência local nos modelos. Neste trabalho, utilizamos a distância de Cook para detecção destes pontos.

No capítulo Aplicação, o modelo de regressão Beta inflacionado em zero foi ajustado aos dados de absenteísmo com licença médica dos trabalhadores de uma empresa de petróleo. Entretanto, o ajuste realizado não se mostrou adequado devido a influência das observações discrepantes, os quais causaram um impacto significativo nas estimativas dos parâmetros. Deste modo, faz-se necessário a busca de uma outra metodologia de análise para estes dados. Uma alternativa a ser investigada como trabalho futuro é o modelo de quasiverossimilhança abordado por Ramalho, Ramalho & Murteira (2011) como um método alternativo para avaliar os dados, o qual nos permite modelar os dados sem assumir qualquer distribuição de probabilidade, utilizando uma função que relaciona a média e a variância.

REFERÊNCIAS

[1] Ferrari, S. L. P. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 7, 799-815.

[2] Ospina, R. & Ferrari, S. L. P. (2010). Inflated beta distributions. *Stat Papers*, 51, 111-126.

[3] Ospina, R. & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models *Computational Statistics and Data Analysis*, 56, 1609-1623.

[4] Oenning, N. S. X., Carvalho, F. M. & Lima, V. M. C. (2014). Fatores de risco para absenteísmo com licença médica em trabalhadores da indústria de petróleo. *Rev. Saúde Pública,* 48(1), 103-112.

[5] Oetiker, T., Partl, H., Hyna, I. & Schilegl, E. (2006). The Not So Short Introduction to LATEX2. Inc. 675 Mass Ave, Cambridge, MA 02139, USA.

[6] Ramalho, E. A., Ramalho, J. J. S. & Murteira, J. M. R. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, 25, 19–68.

[7] Ridout, M., Demétrio, C. G.B. & Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*, Cape Town, 1998.

[8] Stasinopoulos, D. M. & Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software.* Volume 23, Issue 7.