Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Danielle da Silva Pompeu

Aplicação da distribuição Beta na análise de dados inflacionados de zeros

Maringá
2015

Danielle da Silva Pompeu

Aplicação da distribuição Beta na análise de dados inflacionados de zeros

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção de título de Mestre em Bioestatística.

Orientadora: Dra. Terezinha Aparecida Guedes

Maringá
2015

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

Pompeu, Danielle da Silva
P788a Aplicação da distribuição Beta na análise de dados inflacionados de zeros / Danielle da Silva Pompeu. -- Maringá, 2015.
84 f. : il., figs., tabs.

Orientadora: Profa. Dra. Terezinha Aparecida Guedes.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Programa de Pós-Graduação em Bioestatística, 2015.

1. Distribuição Beta inflacionada. 2. Incidência de cancro cítrico - Modelo de Regressão. I. Guedes,Terezinha Aparecida, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 21.ed.570.15195

DANIELLE DA SILVA POMPEU

Aplicação da distribuição Beta na análise de dados inflacionados de zeros

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Profª. Dra. Terezinha Aparecida Guedes
Universidade Estadual de Maringá - UEM

Prof. Dr. Josmar Mazucheli
Universidade Estadual de Maringá - UEM

Prof. Dr. Afrânio Márcio Corrêa Vieira
Universidade Federal de São Carlos - UFSCar

Maringá, 28 de julho de 2015.

A minha mãe

Embora ninguém possa voltar atrás e fazer um novo começo, qualquer um pode começar agora e fazer um novo fim.

Chico Xavier

Agradecimentos

- A Deus, por ter me concedido o dom da vida, e por ter me dado força e coragem para chegar até aqui;
- A minha mãe Rosangela Silva, pelo amor e apoio incondicional, por ter me ensinado desde cedo o valor da educação, muito obrigada por me fornecer os princípios básicos e fundamentais para minha formação e concretização dos meus objetivos;
- Ao meu amigo Jair de oliveira pelas diversas conversas que incentivaram, distraíram e tornaram melhor o período de desenvolvimento deste trabalho;
- Ao meu namorado, Eudmar Almeida, por seu companheirismo e dedicação, por ter sido o apoio necessário para reerguer-me nos momentos difíceis e pelo permanente estímulo que foi determinante para concretização deste trabalho. Muito obrigada me suportar em meus momentos de estresse com a paciência e dedicação;
- Aos meus amigos Juliana Georgetti, Edilson Gimenes e Paulo Pereira que dividiram comigo momentos incríveis durante nossa jornada na UEM;
- A minha orientadora, Professora Terezinha Guedes pela colaboração, dedicação, incentivo, generosidade, compreensão e firmeza com que norteou meus estudos;
- A todos os professores do programa de Pós-graduação em Bioestatística, pelo conhecimento repassado;
- Em especial aos professores Vanderly Janeiro, Josmar Mazucheli, Afrânio vieira, pela colaboração neste trabalho;
- A Dra. Aline Gonçalves-Zuliane e ao Prof. Dr. William Nunes por terem disponibilizado os dados de incidência de cancro cítrico;
- À CAPES, pelo apoio financeiro;
- Por fim, agradeço a todos que de alguma maneira contribuíram para a realização deste trabalho de forma direta ou indireta.

Lista de Figuras

2.1 Função densidades da distribuição Beta para diferentes valores de μ, com $\phi = 10$ (a) e $\phi = 90$ (b). 10

2.2 Funções de densidades para a distribuição BIZ; $\alpha = 0.5$. 12

2.3 Frequências da incidência de cancro cítico em frutos de laranjas doces. . . 18

2.4 Distribuições acumuladas para Beta inflacionada de zero e Beta. 18

3.1 Box-plot da variável incidência de fungos em folhas laranjas doces variedade Pera por genótipo. 36

3.2 Box-plot da variável incidência de fungos em folhas laranjas doces variedade Pera por Porta-enxerto 37

3.3 Frequência da incidência de fungos em folhas laranjas doces variedade Pera 37

3.4 Gráficos de Resíduos. (a) Valores Ajustados *versus* Quantil Residual; (b) Índice *versus* Quantil Residual; (c) Histograma dos Resíduos; (d) Q-Qplot. 41

3.5 Gráficos de Resíduos. (a) e (b) Componente Discreto; (c) e (d) Componente Contínuo. 41

3.6 Frequência da incidência de fungos em folhas laranjas doces variedade Natal 44

3.7 Frequência da incidência de fungos em folhas laranjas doces variedade Natal 44

3.8 Gráficos de Resíduos. (a) Valores Ajustados *versus* Quantil Residual; (b) Índice *versus* Quantil Residual; (c) Histograma dos Resíduos; (d) Q-Qplot. 46

3.9 Gráficos de Resíduos. (a) e (b) Componente Discreto; (c) e (d) Componente Contínuo. 47

Lista de Tabelas

2.1 Estimativas e erros-padrão para os parâmetros das distribuições BIZ e Beta. 17

3.1 Genótipos de laranja doce, variedade Pera e porta-enxertos. 35

3.2 Estimativas e erros-padrão do modelo de regressão Beta inflacionado de zeros, para a incidência de cancro cítrico em folhas de laranja doce. . . . 39

3.3 Estimativas e erros-padrão do modelo de regressão Beta inflacionado de zeros. 40

3.4 Insumos utilizados para compor os tratamentos do manejo para combate ao cancro cítrico. 43

3.5 Estimativas e erros-padrão do modelo de regressão Beta inflacionado de zeros. 46

Sumário

1 Resumo Geral **1**

1.1 Introdução 1

1.2 Justificativa 2

1.3 Objetivos 3

1.3.1 Objetivo Geral 3

1.3.2 Objetivos específicos: 3

1.4 Materiais e Métodos 3

1.5 Principais Resultados 3

1.6 Conclusão 4

2 Distribuição Beta aplicada a dados inflacionados **6**

2.1 Introdução 7

2.2 Metodologia 8

2.2.1 Distribuição Beta 8

2.2.2 Distribuição Beta Inflacionada 10

2.3 Avaliação dos parâmetros estimados da distribuição BIZ 16

2.4 Conclusão 18

3 Modelo de regressão Beta para dados inflacionados de zeros **20**

3.1 Introdução 21

3.2 Metodologia 24

3.2.1 Distribuição Beta Inflacionada de zeros 24

3.2.2 Modelo de regressão Beta inflacionado de zeros 25

3.2.3 Testes de adequabilidade do modelo 31
3.2.4 Análise residual do modelo RBIZ 32
3.3 Aplicações 34
3.3.1 Incidência de cancro cítrico em folhas de laranja doce variedade Pera 35
3.3.2 Conclusão 42
3.3.3 Incidência de cancro cítrico em folhas de laranja doce variedade Natal 43
3.3.4 Conclusão 47

4 Considerações Finais **49**

Referências Bibliográficas **51**

A Programação **54**

B **57**

B.1 Esperança e variância 57
B.2 Regressão Beta 58
B.3 Esperança e variância da distribuição beta inflacionada no ponto zero (BIZ) 58
B.4 Matriz de informação de Fisher para a distribuição beta inflacionada . . 60

Capítulo 1

Resumo Geral

1.1 Introdução

A análise de regressão é uma técnica estatística utilizada para avaliar a relação funcional entre variáveis que apresente relação de dependência entre si. O modelo de regressão linear normal é bastante utilizado em análises empíricas, no entanto, tal modelo torna-se inapropriado em situações em que a variável resposta esta restrita ao intervalo limitado (0, 1), como ocorre com taxas, frações e proporções.

A forma mais simples de contornar este problema é aplicar uma transformação na variável resposta, para que desta forma, a variável assuma valores na reta real e seja possível ajustar o modelo de regressão linear para a variável transformada. Contudo, a utilização dessas transformações podem trazer alguns inconvenientes, como por exemplo, pode trazer dificuldades na interpretação dos parâmetros do modelo em relação à resposta original e, esta transformação não garante que "a nova variável" tenha distribuição normal e seja homocedástica, sendo estes os pressupostos básicos do modelo de regressão linear normal (PEREIRA, 2012).

Outra maneira de ajustar um modelo de regressão para variável resposta contínua é supondo que a mesma possa ser descrita por uma distribuição de probabilidade. Para essas situações Ferrari e Cribari-Neto (2004), propuseram uma classe de modelos de regressão Beta, em que a variável resposta segue distribuição Beta e, sua estrutura e procedimentos inferenciais do modelo são similares aos dos modelos lineares generaliza-

dos (MCCULLAGH; NELDER, 1989). Na literatura destacam-se diferentes especificações para a regressão Beta, tais como, Paolino (2001), Kieschnick e McCullough (2003), Ferrari e Cribari-Neto (2004), Vasconcellos e Cribari-Neto (2005) e Martínez (2008).

Contudo, em muitos casos o conjunto de dados contém um ou os dois limites do intervalo, ou seja, os valores observados podem estar contidos em um dos intervalos [0, 1), (0, 1] ou [0, 1]. Nestes casos, para descrevê-los será necessário tomar uma distribuição de probabilidade que estime as probabilidades nos pontos 0, 1 ou em ambos.

Ospina e Ferrari (2010) sugerem para os casos em que as observações estão em um dos seguintes intervalos [0, 1), (0, 1] e [0, 1], a utilização de uma distribuição de probabilidade resultante da mistura de uma distribuição contínua e uma distribuição discreta. A distribuição contínua irá estimar a probabilidade para os valores no espaço (0, 1) e a distribuição discreta deve estimar probabilidade para os pontos dos extremos do intervalo no caso de zero e/ou um.

Neste sentido, a proposta deste trabalho consiste em utilizar o modelo de regressão Beta inflacionado, onde assume-se que a variável de estudos, incidência de cancro cítrico, é uma mistura destas distribuições e assume valores no intervalo [0, 1).

1.2 Justificativa

Os conjuntos de observações oriundos de pesquisas no campo da citricultura, com objetivo de estudar a incidência de cancro cítrico, são formados por valores no intervalo [0, 1) e apresentam excesso de zeros. As metodologias utilizadas nas análises destes conjuntos de dados, na maioria das vezes, não são apropriadas. Neste sentido, o tema aqui estudado é justificado, pois será aplicado para analisar dados de incidência de cancro cítrico em folhas e frutos de laranja doce, cujas observações são frequências que ocorrem no intervalo acima citado e apresenta excesso de zeros.

1.3 Objetivos

1.3.1 Objetivo Geral

Utilizar a metodologia de modelos de regressão Beta inflacionado para modelar dados de proporções na presença de zeros.

1.3.2 Objetivos específicos:

- Estudar a distribuição Beta inflacionada e o modelo de regressão Beta inflacionado;

- Aplicar o modelo Beta inflacionado a conjuntos de dados observados da área de agronomia.

1.4 Materiais e Métodos

Para a aplicação da metodologia de modelos inflacionados de zeros, foi utilizado três conjuntos de dados obtidos de planejamentos experimentais implementados na região Noroeste do Estado do Paraná.

1.5 Principais Resultados

No segundo capítulo foi utilizada a distribuição Beta inflacionada de zeros em um conjunto de dados de incidência de cancro cítrico em frutos de laranja doce, variedade Westin. Pode-se perceber que a distribuição Beta inflacionada ajustou-se aos dados de incidência que estão no intervalo $[0, 1)$. No terceiro capítulo, foram aplicados modelos de regressão Beta inflacionado de zeros, em dois conjuntos de dados e, de forma geral pode-se observar que o modelo mostrou-se adequado para descrever a incidência de fungos em folhas de laranja doce.

1.6 Conclusão

Ao longo deste trabalho foram utilizados modelos de regressão Beta inflacionado de zeros, para modelar dados de incidência de cancro cítrico em folhas de laranja doce e, observou-se que o modelo é adequado para descrever tais dados. Esta adequação pode ser vista na análise residual, bem como nas estimativas encontradas no modelo. Na primeira aplicação do modelo de regressão, observou-se que o componente discreto do modelo explicou o comportamento dos porta-enxertos. Pelo componente contínuo, verificou-se quais genótipo mais contribuiu para a redução de cancro cítrico em folhas de laranja doce. Para a segunda aplicação, observou-se que todos os tratamentos contribuem para o não aparecimento da doença. Sendo assim, pode-se afirmar que esta metodologia é adequada para descrever dados de proporção com excesso de zeros.

Capítulo 2

Distribuição Beta aplicada a dados inflacionados

Resumo do capítulo

A distribuição Beta é utilizada para modelar dados que são medidos na forma de taxas, frações ou proporções, restritos ao intervalo contínuo (0, 1). Contudo, é comum observar situações onde observa-se em um ou em ambos os extremos do intervalo ([0, 1) , (0, 1] e [0, 1]). Neste casos, a distribuição Beta torna-se inadequada, sendo necessário utilizar um modelo de mistura entre uma distribuição discreta que capture a massa de probabilidade zero ou um e uma distribuição contínua. Logo, o objetivo deste trabalho foi estudar a distribuição Beta inflacionada no ponto zero aplicada a um conjunto de dados de incidências de fungos em frutos de laranja doce.

Abstract

The Beta distribution is used to model data which are measured as rates, ratios or fractions, restricted to the continuous interval (0, 1). However, it is common to find situations where the values are at one or both extremes of the range ([0, 1), (0, 1] and [0, 1]). In this case, the Beta distribution becomes inadequate, being necessary to use a mixture model of a discrete distribution, that captures the probability mass zero or one, and a continuous distribution. Therefore, the aim of this study was to model fungi incidences in sweet orange fruit using the zero-inflated Beta distribution.

2.1 Introdução

Devido ao grande número de pesquisas que resultam em observações que são mensurados na forma de taxas, frações ou proporções a distribuição de probabilidade Beta é apropriada, devido sua flexibilidade para modelar dados que assumem valores no intervalo (0, 1). Contudo, existem situações em que observa-se valores em um ou em ambos os extremos do intervalo. Nestes casos, o uso da distribuição Beta torna-se inapropriada, sendo assim, o uso de uma distribuição de mistura entre uma distribuição discreta e uma distribuição contínua é mais apropriado.

A distribuição obtida a partir da mistura destas distribuições faz parte da classe de distribuições inflacionadas. O termo inflacionado indica que a frequência observada em alguns pontos (zero e/ou um), por exemplo, é maior do que o esperado pela distribuição, sendo assim, nos trabalhos em que fazem uso de distribuições inflacionadas é suposto que a variável dependente é uma mistura de duas distribuições, sendo uma degenerada no ponto com maior frequência (zero) e outra distribuição conhecida. O precursor dos modelos de mistura foi Aitchison (1955), utilizando uma distribuição degenerada em zero e uma distribuição lognormal denominada de *delta*.

A partir do trabalho de Aitchison (1955), inúmeros trabalhos foram desenvolvidos com diferentes distribuições inflacionadas, dentre os quais destacam-se, o modelo de regressão proposto por Feuerverger (1979), com o intuito de estudar as precipitações

pluviométricas utilizando a distribuição Gama. Lambert (1992), foi a pioneira no uso de modelos de regressão para dados de contagem utilizado uma distribuição inflacionada de zeros.

Fumes e Corrente (2010), comparam diferentes modelos inflacionados de zero, tais como, Poisson inflacionado de zero (ZIP), Binomial Negativa Inflacionada de Zero (ZINB), Poisson e Binomial Negativa, com o objetivo de verificar qual modelo melhor se ajusta aos dados, segundo o percentual de zeros encontrados e os fatores que influenciam tanto o consumo como o não consumo de alguns alimentos. Destacam-se também os modelos Binomial (HALL, 2000), Normal Inversa (HELLER et al., 2006), Beta ((MARTÍNEZ, 2008); (PEREIRA; SOUZA; CRIBARI-NETO, 2014) (VIEIRA; HINDE; DEMÉTRIO, 2000)), entre outros.

Este capítulo tem como objetivo estudar a distribuição Beta inflacionada no ponto zero (BIZ). Além disso, a metodologia será aplicada a um conjunto de observações de incidência de fungos em frutos de laranjas doces, cujas mudas foram enxertadas em outras variedades de frutas cítricas. A área experimental foi localizada no município de Paranavaí, região Noroeste do Estado do Paraná.

2.2 Metodologia

2.2.1 Distribuição Beta

A distribuição Beta pertence à família de distribuições contínuas, cujo suporte é restrito ao intervalo aberto (0, 1). Segundo Mood, Graybill e Boes (1974), uma variável aleatória Y segue uma distribuição Beta com parâmetros $a, b > 0$ quando sua função de densidade de probabilidade for escrita na forma:

$$f_y(y) = f_y(y; a,b) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} \mathrm{I}_{(0,1)}(y). \tag{2.1}$$

A função Beta é dada pela integral $B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}dy$. A média e variância

(Apêndice B.1) de Y são dadas, respectivamente, por:

$$E[Y] = \frac{a}{a+b} \qquad \text{e} \qquad Var[Y] = \frac{ab}{(a+b+1)(a+b)^2}.$$

A função Beta esta relacionada com a função gama de acordo com a seguinte expressão:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Sendo assim, a função densidade da variável Y que segue uma distribuição Beta pode ser reescrita como:

$$f_y(y) = f_y(y; a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} \tag{2.2}$$

em que, $\Gamma(\cdot)$ é a função gama, que é definida no ponto k como $\Gamma(k) = \int_0^\infty y^{k-1}e^{-y}dy$. Se $a = 0$ e $b = 1$, a distribuição Beta assume sua forma padrão, sendo esta a mais utilizada em diversas aplicações. O seu uso pode ser visto em Bury (1999), que aplica a distribuição Beta em problemas relacionados a engenharia e Pereira (2010) que aplica a distribuição em uma análise das eficiências administrativas dos municípios do estado de São Paulo.

Uma outra parametrização da distribuição Beta foi sugerida por Ferrari e Cribari-Neto (2004), que desenvolveram uma especificação mais estruturada e melhor formalizada da distribuição Beta, com objetivo de descrever os parâmetros de posição e precisão, representados por μ e ϕ respectivamente. Esta nova parametrização permite modelar de forma direta a média usando um preditor linear e uma função de ligação, sendo assim, pode-se dizer que esta especificação é semelhante aos modelos lineares generalizados (MCCULLAGH; NELDER, 1989).

Sendo que, $\mu = \dfrac{a}{a+b}$ e $\phi = a+b$ (ver Apêndice B.2) e a função de densidade de Y pode ser escrita como

$$f_y(y; \mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{2.3}$$

onde $0 < y < 1$, $0 < \mu < 1$ e $\phi > 0$.

Sendo assim, Y segue uma função de distribuição Beta com média μ e precisão ϕ e denotada por $Y \sim \mathfrak{B}(\mu,\phi)$. A média e variância de Y nesta parametrização são respectivamente, $E[Y] = \mu$ e $Var[Y] = \dfrac{V(\mu)}{1+\phi}$, em que $V(\mu) = \mu(1-\mu)$ e representa a função de variância.

A função de densidade da variável aleatória Y, que tem distribuição de probabilidade Beta, é muito flexível podendo assumir diversas formas, dependendo dos valores atribuído aos parâmetros μ e ϕ, por exemplo, na forma de 'U', quando têm-se $(1-\mu)\phi < 1$ e $\mu\phi < 1$, 'J', quando têm-se $(1-\mu)\phi - 1(\mu\phi - 1) < 0$ e, ainda na forma de 'J' invertido, quando têm-se $(1-\mu)\phi - 1(\mu\phi - 1) > 0$ (MARTÍNEZ, 2008). Deste modo, a distribuição pode ficar na forma simétrica quando tem-se $\mu = \dfrac{1}{2}$ e assimétrica quando tem-se $\mu \neq \dfrac{1}{2}$ (FERRARI; CRIBARI-NETO, 2004). Podendo ser observado na Figura 2.1.

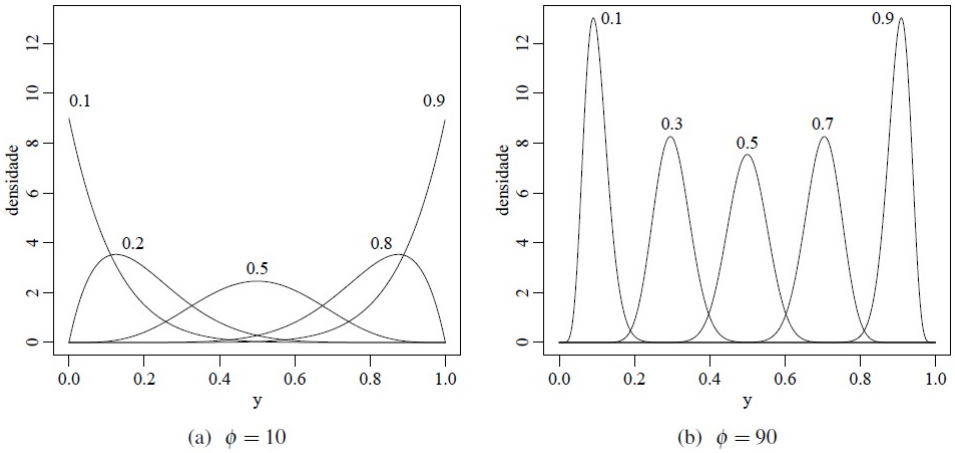

Figura 2.1: Função densidades da distribuição Beta para diferentes valores de μ, com $\phi = 10$ (a) e $\phi = 90$ (b).

2.2.2 Distribuição Beta Inflacionada

A distribuição Beta inflacionada surgiu da necessidade de descrever conjuntos de observações que assumem valores em um dos seguintes intervalos: [0, 1), (0, 1] e [0, 1]. Para este tipo de situação, Ospina e Ferrari (2010) introduziram uma família de distribuições, conhecida como distribuição Beta inflacionada, que são misturas de uma

distribuição Beta e uma distribuição de Bernoulli degenerada em zero e/ou um, a fim de estimar a massa de probabilidade de 0 e/ou 1.

A função de distribuição acumulada do modelo de mistura da distribuição Beta com a degenerada em zero, um ou ambos é dada por:

$$BI_c(y;\alpha,\mu,\phi) = \alpha I_{\{c\}}(y) + (1-\alpha)F(y;\mu,\phi), \tag{2.4}$$

em que, $I_A(y)$ é uma função indicadora, que assume valor 1 se $y \in A$ e 0 caso contrário, sendo A o conjunto de elementos, em que encontra-se o valor de $y = c$; $F(\cdot;\mu,\phi)$ é a função de distribuição acumulada da distribuição Beta; $\alpha = P(y = c)$ é o parâmetro de mistura da distribuição $(0 < \alpha < 1)$. Como a função BI_c tem um ponto de massa em $y = c$, então não pode ser considerada completamente contínua. Note que, com uma probabilidade α, a variável Y é selecionada a partir de um distribuição degenerada em c e, quando a probabilidade é de $(1-\alpha)$, a variável é selecionada de uma distribuição Beta. Isto é, a função de densidade de probabilidade da variável Y é dada pelo valor gerado pela mistura, e escrita na forma:

$$bi_c(y;\alpha,\mu,\phi) = \{\alpha^{I_{\{c\}}(y)}(1-\alpha)^{1-I_{\{c\}}y}\}\{f(y;\mu,\phi)^{1-I_{\{c\}}y}\} \tag{2.5}$$

em que $\alpha > 0$, $0 < \mu < 1$ e $\phi > 0$, sabe-se que μ e ϕ são os parâmetros da distribuição Beta e $f(y;\mu,\phi)$ é a função densidade apresentada na Equação (2.3). A função (2.5) é a função de densidade de probabilidade de uma distribuição Beta inflacionada no ponto de massa c, para $c = 0$ ou $c = 1$. Se $\alpha > 0$, a massa de probabilidade da distribuição Beta no ponto $y = c$ é excedida. A probabilidade de se observar $y = 0$ ou $y = 1$ é $\alpha = P(y = c)$. Observe que o primeiro termo da distribuição apresentada na Equação (2.5) depende apenas de α e o segundo termo depende de (μ,ϕ), pois envolve a parte contínua da variável resposta (MARTÍNEZ, 2008).

Neste trabalho, será discutido o caso em que valores observados estão no intervalo [0,1) $(0 \leq y < 1)$. A distribuição apresentada na Equação (2.5) para y no intervalo [0,1) é denominada de distribuição Beta inflacionada no ponto zero (BIZ), denotada

por $y \sim BIZ(\alpha, \mu, \phi)$. A função de densidade da distribuição BIZ é escrita na forma:

$$bi_c(y; \alpha, \mu, \phi) = \{\alpha^{\mathrm{I}_{\{0\}}(y)}(1-\alpha)^{1-\mathrm{I}_{(0,1)}(y)}\}\{f(y; \mu_{it}, \phi_{it})^{1-\mathrm{I}_{(0,1)}(y)}\}. \tag{2.6}$$

Em geral a esperança e variância de uma variável Y com distribuição Beta inflacionada (2.5) é dada por

$$E(Y) = c\alpha + (1-\alpha)\mu \quad \text{e} \quad Var(Y) = (1-\alpha)\frac{V(\mu)}{\phi+1} + \alpha(1-\alpha)(c-\mu)^2. \tag{2.7}$$

Para Y com distribuição Beta inflacionada no ponto zero (2.6), sua média e variância são dadas por

$$E(Y) = (1-\alpha)\mu \qquad \text{e} \qquad Var(Y) = (1-\alpha)\frac{V(\mu)}{\phi+1} + \alpha(1-\alpha)\mu^2.$$

A variância de Y, com distribuição BIZ, será tanto menor quanto maior for o valor do parâmetro de dispersão ϕ.

Observa-se na Figura (2.2) o gráfico da distribuição BIZ, que independentemente do valor escolhido para os parâmetros μ e ϕ a distribuição se apresentar de forma assimétrica. Isso ocorre, devido o ponto de massa (zero), representado por $\alpha = P(y = 0)$, este ponto pode ser observado na linha vertical do gráfico.

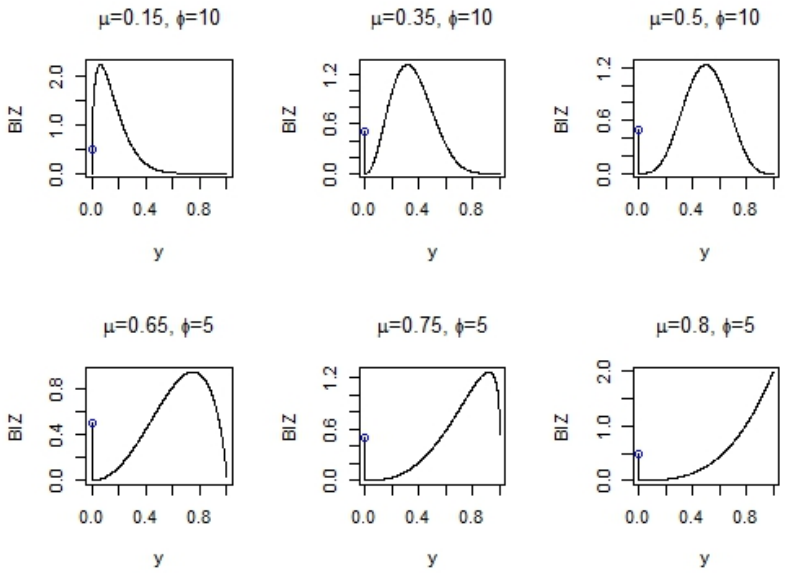

Figura 2.2: Funções de densidades para a distribuição BIZ; $\alpha = 0.5$.

A proposição a seguir foi apresentada por Ospina e Ferrari (2010) para mostrar que uma distribuição Beta inflacionada de zeros ou de uns pertence a família exponencial.

Proposição 2.3.1: A distribuição Beta inflacionada no ponto c apresentada na Equação (2.5) pertence à família exponencial de dimensão 3 de posto completo.

Demonstração: Seja $\eta = (\eta_1,\eta_2,\eta_3)$, onde $\eta_1 = [log(\alpha/(1-\alpha)) + b(\eta_2,\eta_3)]$, $\eta_2 = \mu\phi$ e $\eta_3 = (1-\mu)\phi$, sendo $b(\eta_2,\eta_3) = log(\Gamma(\eta_2)\Gamma(\eta_3)/\Gamma(\eta_2+\eta_3))$. Considere também o vetor de estatísticas dado por $T(y) = (t_1(y),t_2(y),t_3(y))$, sendo que:

$$
\begin{aligned}
t_1(y) &= \begin{cases} 1, \text{ se } \quad y = c, \\ 0, \text{ se } \quad y \in (0,1). \end{cases} \\
t_2(y) &= \begin{cases} \log(y), \text{ se } \quad y \in (0,1), \\ 0, \text{ se } \quad y = c. \end{cases} \\
t_3(y) &= \begin{cases} \log(1-y), \text{ se } \quad y \in (0,1), \\ 0, \text{ se } \quad y = c. \end{cases}
\end{aligned} \tag{2.8}
$$

Desta forma, a distribuição Beta inflacionada no ponto c apresentada na Equação (2.5) pode ser escrita como:

$$bi_c(y,\alpha,\mu,\phi) = \exp\{\eta^{\intercal}\mathrm{T}(y) - B^*(\eta)\}h(y), \tag{2.9}$$

sendo que $B^*(\eta) = \log\{1 + exp[\eta_1 - \mathrm{b}(\eta_2,\eta_3)]\} + b(\eta_2,\eta_3)$ é uma função de valores reais em η e $h(y)$ é uma função positiva definida no conjunto $(0,1) \cup c$, sendo escrita como:

$$h(y) = \begin{cases} \dfrac{1}{\{y(1-y)\}}, \text{ se } \quad y \in (0,1), \\ 0, \text{ se } \quad y = c. \end{cases}$$

Sendo assim, a distribuição Beta inflacionada apresentada na Equação (2.5) pertence a família exponencial. Segundo Martínez (2008), a parametrização de η é uma transformação bijetora que leva $\mathfrak{X} = \{(\alpha,\mu,\phi) : (\alpha,\mu,\phi) \in (0,1) \times (0,1) \times \mathbb{R}^+\}$ isto é, o jacobiano da transformação é diferente de zero para todo $\eta \in \mathfrak{D}$, um subconjunto aberto de $\mathbb{R}^3$.

Além disso, os $t's$ e os $\eta's$ não satisfazem contrastes lineares e o espaço paramétrico contém um retângulo tridimensional. Deste modo, a Expressão (2.9) é uma representação canônica da distribuição Beta inflacionada no ponto c na família exponencial de dimensão 3 de posto completo.

A partir da proposição (2.3.1) tem-se o vetor de estatísticas $\sum_{t=1}^{n} \mathrm{T}(y_t) = (\mathrm{T}_1, \mathrm{T}_2, \mathrm{T}_3)$, em que:

$$\begin{aligned} \mathrm{T}_1 &= \sum_{t=1}^{n} \mathrm{I}_{\{c\}}(y_t), \\ \mathrm{T}_2 &= \sum_{t:y_t \in (0,1)} \log y_t, \\ \mathrm{T}_3 &= \sum_{t:y_t \in (0,1)} \log(1 - y_t), \end{aligned} \tag{2.10}$$

é uma estatística suficiente completa (LEHMANN; CASELLA, 1998).

Considerando uma amostra aleatória $y_1,...,y_n$, em que cada unidade amostral tem função densidade apresentada na Equação (2.5). A função de verosimilhança para o parâmetro $\theta = (\alpha, \mu, \phi)$ é dada por:

$$L(\alpha, \mu, \phi, y) = \prod_{t=1}^{n} bi_c(y_t; \alpha, \mu, \phi) = L_1(\alpha) L_2(\mu, \phi) \tag{2.11}$$

sendo que

$$L_1(\alpha) = \prod_{t=1}^{n} \alpha^{\mathrm{I}_{\{c\}}(y_t)} (1-\alpha)^{1-\mathrm{I}_{\{c\}}(y_t)} = \alpha^{T_1} (1-\alpha)^{n-T_1},$$

$$L_2(\mu, \phi) = \prod_{t=1}^{n} f(y_t, \mu, \phi)^{1-\mathrm{I}_{\{c\}}(y_t)}.$$

Como pode ser observado, o primeiro componente da função de verossimilhança $L(\alpha, \mu, \phi, y)$, $L_1(\alpha)$, depende apenas do parâmetro α. O segundo componente da função, $L_2(\mu, \phi)$ depende dos parâmetros (μ, ϕ).

O logaritmo da função de verossimilhança para $\theta = (\alpha, \mu, \phi)$ é dado por:

$$l(\alpha, \mu, \phi, y) = \log(L_1(\alpha) L_2(\mu, \phi)) = l_1(\alpha) + l_2(\mu, \phi) \tag{2.12}$$

em que

$$l_1(\alpha) = \mathrm{T}_1 \log(\alpha) + (n - \mathrm{T}_1)\log(1-\alpha)$$

$$\begin{aligned} l_2(\mu,\phi) =& (n - \mathrm{T}_1)\log(f(y_t;\mu,\phi)) \\ =& (n - \mathrm{T}_1)\log(\Gamma(\phi)) - \log(\Gamma(\mu,\phi)) + \log(\Gamma((1-\mu)\phi)) \\ &+ \mathrm{T}_2(\mu\phi - 1) + \mathrm{T}_3((1-\mu)\phi - 1). \end{aligned}$$

Derivando o logaritmo da função de verossimilhança em relação a θ, obtém-se as funções escores.

$$\begin{aligned} U_1 &= \frac{\partial l_1(\alpha)}{\partial \alpha} = \textstyle\sum_{t=1}^{n} \frac{\mathrm{I}_{(y_t)}}{\alpha} - \frac{(1-\mathrm{I}_{(y_t)})}{1-\alpha} \\ U_2 &= \frac{\partial l_2(\mu\phi)}{\partial \mu} = \textstyle\sum_{t=1}^{n} [-\psi(\mu\phi)\phi - \psi((1-\mu)\phi)\phi + \phi\log(y_t) - \phi\log(1-y_t)] \\ U_3 &= \frac{\partial l_2(\mu\phi)}{\partial \phi} = \textstyle\sum_{t=1}^{n} [(1-\mathrm{I}_{(y_t)})\psi(\phi) - \psi(\mu\phi)\mu + \psi((1-\mu)\phi)(1-\mu) \end{aligned}$$

ou, utilizando as estatísticas apresentadas na Equação (2.10) tem-se

$$\begin{aligned} U_1 &= \frac{\partial l_1(\alpha)}{\partial \alpha} = \frac{\mathrm{T}_1}{\alpha} - \frac{n-\mathrm{T}_1}{1-\alpha}, \\ U_2 &= \frac{\partial l_2(\mu\phi)}{\partial \mu} = \phi(n-\mathrm{T}_1)[\psi((1-\mu)\phi) - \psi(\mu\phi) + \mathrm{T}_2 + \mathrm{T}_3, \\ U_3 &= \frac{\partial l_2(\mu\phi)}{\partial \phi} = (n-\mathrm{T}_1)[\psi(\phi) - \mu\psi(\mu\phi) - (1-\mu)\phi((1-\mu)\phi)] + n\mathrm{T}_2 - (1-\mu)\mathrm{T}_3. \end{aligned}$$

sendo que $\psi(\cdot)$ é a função digama obtida a partir da primeira derivada da Γ.

Dada a separabilidade apresentada na Equação (2.11) é possível obter a função escore para cada um dos parâmetros de forma independente. Sendo assim, o estimador de máxima verossimilhança de α é obtido a partir da solução do sistema $U_1(\alpha) = 0$, este por sua vez possui solução analítica sendo possível obter o EMV de α dado por $\widehat{\alpha} = \mathrm{T}_1/n$; $\widehat{\alpha}$ é uma função de estatística suficiente completa e é um estimador não viesado de α (LEHMANN; CASELLA, 1998). Os estimadores de μ e ϕ não possuem forma fechada e devem ser obtidos numericamente pela maximização da função de log verossimilhança usando um algoritmo de otimização não-linear, tal como um algoritmo de Newton ou um algoritmo quasi-Newton (GOURIEROUX; MONFORT, 1995).

A matriz de informação de Fisher para a distribuição Beta inflacionada é dada por:

$$J(\widehat{\theta}) = \begin{bmatrix} \frac{\partial^2 l_1}{\partial^2 \alpha} & 0 & 0 \\ 0 & \frac{\partial^2 l_2}{\partial^2 \mu} & \frac{\partial^2 l_2}{\partial^2 \mu\phi} \\ 0 & \frac{\partial^2 l_2}{\partial^2 \mu\phi} & \frac{\partial^2 l_2}{\partial^2 \phi} \end{bmatrix}$$

onde,

$$\begin{aligned}
\frac{\partial^2 l_1(\theta)}{\partial^2 \alpha} &= \frac{n}{\alpha(1-\alpha)}, \\
\frac{\partial^2 l_2(\theta)}{\partial^2 \mu} &= n\phi^2(1-\alpha)[\psi'((1-\mu)\phi) + \psi'(\mu\phi)], \\
\frac{\partial^2 l_2(\theta)}{\partial^2 \phi} &= n(1-\alpha)\{-\psi'(\phi) + \mu^2\psi'(\mu\phi) - (1-\mu)^2\psi'((1-\mu)\phi)\}, \\
\frac{\partial^2 l_2(\theta)}{\partial \mu \partial \phi} &= (1-\alpha)\phi\{\psi'(\mu\phi)\mu - \psi'((1-\mu)\phi)(1-\mu)\}
\end{aligned}$$

e, $\psi'(\cdot)$ é a função trigama definida como $\psi'(j) = \frac{\partial^2 \log \Gamma(j)}{\partial^2 (j)}$

Observa-se que na matriz de informação de Fisher os termos $\frac{\partial^2 l_1}{\partial^2 \alpha\mu} = \frac{\partial^2 l_1}{\partial^2 \alpha\phi} = \frac{\partial^2 l_2}{\partial^2 \mu\alpha} = \frac{\partial^2 l_2}{\partial^2 \phi\alpha} = 0$, indicando que o parâmetro α é ortogonal ao vetor de parâmetros (μ, ϕ), ou seja, os respectivos componentes do vetor escore são não correlacionados e $\widehat{\alpha}$ e $(\widehat{\mu}$,$\widehat{\phi})$ são assintoticamente independentes. Além disso, pela proposição (2.3.1) pode ser obtido a normalidade assintótica do estimador de máxima verossimilhança (EMV) e pela normalidade assintótica é dada a consistência do EMV de θ, sendo possível a construção dos intervalos de confiança.

2.3 Avaliação dos parâmetros estimados da distribuição BIZ

O conjunto de dados utilizado para ilustrar a utilização da distribuição Beta inflacionada, cujos valores se encontram no intervalo [0, 1), foi obtido de um planejamento conduzido no município de Paranavaí, região Noroeste do Estado do Paraná. Foram utilizados como material vegetal nove genótipos de laranja doce, variedade Westin, enxertados sob quatro porta-enxertos diferentes. Foram coletados 100 frutos aleatoria-

mente de cada variedade para estimar a incidência do cancro cítrico. Para a incidência foi feita a contagem de frutos doentes em relação ao total de frutos coletados.

Para análise exploratória, foram formandos dois conjuntos de dados, um com todas as observações para a estimação dos parâmetros da distribuição BIZ e outro retirando os zeros para a estimação dos parâmetros da distribuição Beta.

Na Tabela (2.1), foram obtidos os valores estimados para os parâmetros das distribuições BIZ e Beta, utilizando o método de estimação de máxima verossimilhança.

Tabela 2.1: Estimativas e erros-padrão para os parâmetros das distribuições BIZ e Beta.

Distribuição BIZ			Distribuição Beta		
Parâmetros	Estimativa	Erro-padrão	Parâmetros	Estimativa	Erro-padrão
$\widehat{\alpha}$	0,668	0,059	$\widehat{a}$	0,174	0,0375
$\widehat{\mu}$	0,174	0,004	$\widehat{b}$	0,327	0,0407
$\widehat{\phi}$	8,360	0,242	-	-	-
E(Y)	0,058		E(Y)	0,174	
Var(Y)	0,345		Var(Y)	0,471	

Para verificar a frequência de zeros no conjunto de dados foi construído o histograma apresentado na Figura (2.3), pode-se observar que a distribuição da variável Y é assimétrica. Além disso, a linha vertical com um ponto acima no histograma representa a quantidade de zeros na amostra, que corresponde a 66,79% dos dados.

Nos gráficos das distribuições acumuladas, Figura (2.4), pode-se observar que os valores ajustados pelas distribuições BIZ e Beta não se distanciam dos valores observados da variável incidência. Desta forma, pode-se dizer que ambas distribuições se ajustam aos dados.

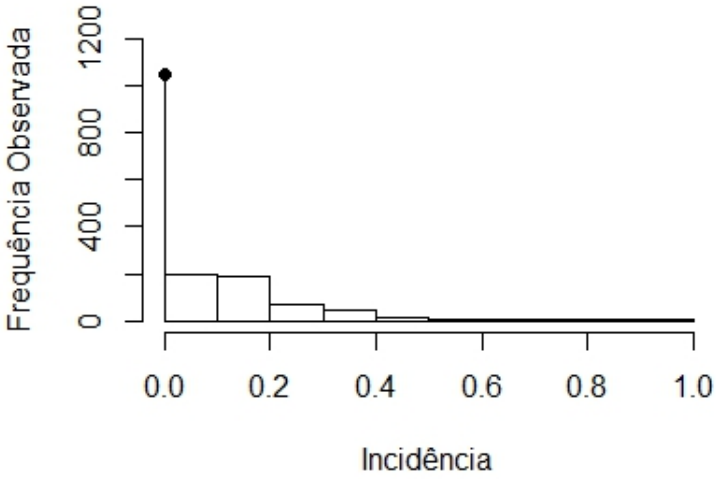

Figura 2.3: Frequências da incidência de cancro cítico em frutos de laranjas doces.

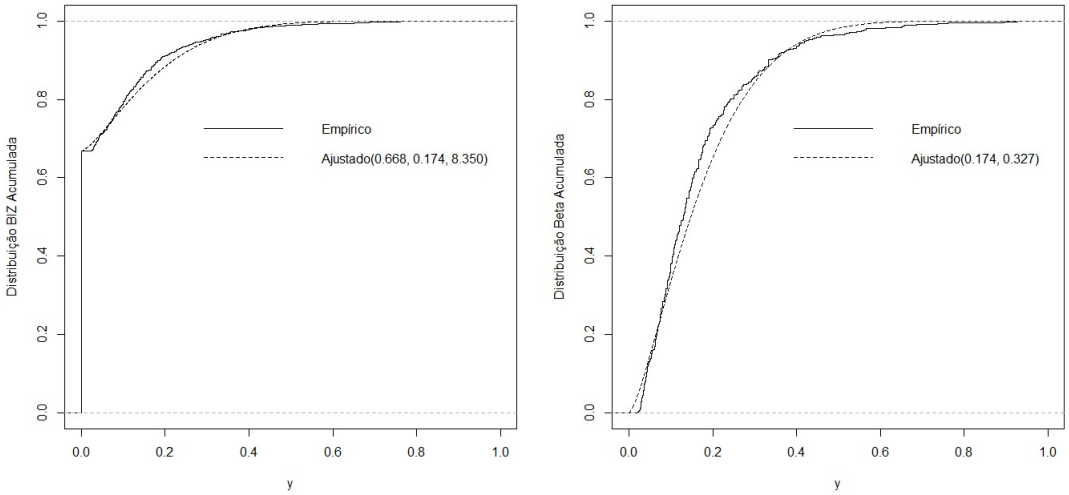

Figura 2.4: Distribuições acumuladas para Beta inflacionada de zero e Beta.

2.4 Conclusão

A distribuição Beta é apropriada para modelar dados que estão no intervalo contínuo (0,1), contudo, a mesma torna-se inapropriada quando existem valores em um dos extremos, sendo assim, é recomendado utilizar um modelo de mistura. Neste capítulo foi discutido o uso de modelos de mistura entre a distribuição Beta e uma distribuição degenera em zero (Bernoulli) com o objetivo de modelar dados no intervalo [0,1).

Capítulo 3

Modelo de regressão Beta para dados inflacionados de zeros

Resumo do capítulo

A regressão Beta inflacionada, é um importante modelo utilizado para descrever taxas, proporções e outras variáveis que assumem valores nos intervalos [0, 1), (0, 1] e [0, 1], e que dependem de uma ou mais variáveis explicativas. Neste estudo, foi utilizados modelos de regressão Beta inflacionado de zeros para analisar dois conjuntos de dados resultantes de experimentos com citros na região noroeste do Paraná. O primeiro conjunto de dados consistiu das observações obtidas de um experimento em que foram empregados como material vegetal, nove genótipos de Laranja doce, variedade Pera, enxertados sob quatro porta-enxertos diferentes. Para o componente discreto, observou-se que a variável porta-enxerto, foi significativa para o não aparecimento da doença. Para a componente contínuo, a variável genótipo mostrou ser significativa para explicar a incidência de cancro cítrico. O segundo conjunto de dados consiste em observações obtidas a partir de um experimento com nove manejos combinados a partir de diferentes insumos, a fim de determinar a influência dessas combinações de insumos sobre a incidência de cancro cítrico em folhas de laranjas doces, variedade Natal. Para o componente discreto, observou-se que todos os manejos foram significativos para o não aparecimento da doença. Para o componente contínuo, quatro dos nove manejos

destaram-se na diminuição do cancro cítrico. As ferramentas de diagnóstico mostram que os modelos foram adequados e desta forma, pode-se dizer que esta metodologia é adequada para modelar dados de proporção com excesso de zeros.

Abstract

The inflated Beta regression, is an important model used to describe rates, ratios, and other variables which take values in the intervals [0, 1), (0, 1] and [0, 1], and that depend on one or more independent variables. In this study, we used a zero-inflated Beta model to analyze two datasets resulting from experiments with citrus in the northwestern region of Paraná. The first dataset consisted of observations obtained from an experiment in which nine sweet orange genotypes from the Pera variety (*citrus sinensis*), grafted on four different rootstocks, were employed as vegetable materials. For the discrete component, we observed that the explanatory variable, rootstock, was significant, inhibiting the onset of the disease. For the continuous component, the variable genotype was shown to be significant to explain the incidence of citrus canker. The second dataset consists of observations obtained from an experiment with nine managements combined from different inputs. In order to determine the influence of these input combinations on the incidence of citrus canker in leaves of sweet oranges, Natal variety, five evaluations were performed at different times. For the discrete component, all managements were significant and shown to inhibit the onset of the disease. For the continuous component, four of the nine managements stood out in decreasing the citrus canker in leaves. Diagnostic tools showed that the model was a good fit to this zero-inflated data.

3.1 Introdução

A análise de regressão é uma técnica estatística usualmente utilizada em diversas áreas do conhecimento científico. Tem por objetivo analisar a relação funcional entre variáveis que apresente relação de dependência, de maneira que seja possível a estimação

ou previsão de uma variável resposta por meio de uma ou mais variáveis preditoras (NETER et al., 1996).

Na modelagem estatística, muitos modelos de regressão tem sido propostos, mas o mais utilizado é o modelo de regressão clássico, conhecido também como modelo de regressão linear. Nesta forma de modelagem, a relação entre as variáveis é descrita por uma função linear com a suposição de independência e normalidade dos erros.

Estes modelos tornam-se inapropriados quando a variável dependente é mensurada na forma de taxas, frações ou proporções, cujos valores estão contidos em um dos seguintes intervalos limitados ([0, 1), (0, 1] e [0, 1]). Nestes casos, os valores estimados pelo modelo de regressão clássico podem ultrapassar os limites desses intervalo. Então, para que isso não ocorra recomenda-se fazer uma transformação na variável resposta de modo que seus valores pertençam a reta real. Esta transformação, no entanto, pode trazer dificuldades na interpretação dos parâmetros do modelo em relação à resposta original.

Outra maneira de ajustar um modelo de regressão para variável resposta contínua é supondo que possa ser descrita por uma distribuição de probabilidade. A distribuição Beta é uma das distribuições apropriadas para modelar dados que assumem valores no intervalo aberto (0, 1).

Na literatura destacam-se vários trabalhos que utilizam modelos de regressão para variáveis com distribuição Beta. Dentre estes trabalhos destaca-se Ferrari e Cribari-Neto (2004) que desenvolveram um modelo de regressão no qual a distribuição da variável resposta é Beta e a resposta média está relacionada com um preditor linear por meio de uma função de ligação. Para facilitar a interpretação dos parâmetros do modelo, os autores propuseram uma nova parametrização para a função de distribuição Beta, a qual utiliza o princípio da lei Beta que é indexado por parâmetros de média e dispersão. Desta forma, o modelo torna-se útil em situações em que a variável dependente tem seus valores no intervalo contínuo (0, 1) e está relacionada com outras variáveis.

Além do trabalho de Ferrari e Cribari-Neto (2004), outros autores empregaram a regressão Beta. Paolino (2001) comparou as estimativas encontradas para os modelos de

regressão Beta e regressão linear, com e sem transformação na variável dependente. Ele concluiu que as estimativas encontradas pela distribuição Beta apresentaram vantagens significativas sobre as estimativas encontradas pelo modelo linear, nos casos em que a variável de estudo assume valores no intervalo (0, 1). Pereira (2010) desenvolveu testes de erros de especificação para os modelos de regressão Beta e correções do teste de razão de verossimilhança para amostras pequenas, além de realizar aplicações e inferências em modelos de regressão Beta para situações em que a dispersão foi considerada fixa e, também, variável.

Existem situações em que a variável de interesse assume valores em um ou em ambos os extremos do intervalo, ou seja, [0, 1), (0, 1] e [0, 1]. Nestes casos, o uso da distribuição Beta torna-se inviável e o emprego de um modelo de mistura entre uma distribuição discreta e uma contínua tem sido recomendado. A distribuição discreta deve estimar a massa de probabilidade em zero e/ou um e a distribuição contínua deve descrever a parte contínua do conjunto de dados. Este tipo de modelagem é denominado de modelos inflacionados.

O primeiro modelo de regressão inflacionado foi proposto por Feuerverger (1979) com o objetivo de estudar as precipitações pluviométricas, utilizando a distribuição gama. Desde então foram desenvolvidos modelos inflacionados para diversas distribuições como, por exemplo, Poisson (LAMBERT, 1992), binomial (HALL, 2000), binomial negativa (FUMES; CORRENTE, 2010), Normal Inversa (HELLER et al., 2006), Beta ((COOK; KIESCHNICK; MCCULLOUGH, 2008); (MARTÍNEZ, 2008)), entre outros.

O objetivo deste capítulo foi utilizar a metodologia de modelo de regressão inflacionado para descrever a incidência de cancro cítrico em folhas de laranja doce, variedade Pera, cujos genótipos foram enxertados em outras variedades de frutas cítricas, além de possibilitar que os diferentes genótipos e porta-enxertos sejam comparados em relação a vulnerabilidade ao fungo.

3.2 Metodologia

3.2.1 Distribuição Beta Inflacionada de zeros

A distribuição Beta inflacionada surgiu da necessidade de descrever conjuntos de observações que assumem valores em um dos seguintes intervalos: [0, 1), (0, 1] e [0, 1]. Para este tipo de situação, Ospina e Ferrari (2010) introduziram uma família de distribuições, conhecida como distribuição Beta inflacionada, que são misturas de uma distribuição Beta e uma distribuição de Bernoulli, a fim de estimar a massa de probabilidade de zero e/ou um. Esta família de distribuições é formada pelas diferentes escolhas do parâmetros da distribuição Beta e pela escolha da distribuição que vai descrever a parte discreta.

A função de distribuição acumulada do modelo de mistura da distribuição Beta com a degenerada em zero, um ou ambos é dada por:

$$BI_c(y;\alpha,\mu,\phi) = \alpha I_{\{c\}}(y) + (1-\alpha)F(y;\mu,\phi), \tag{3.1}$$

em que, $I_A(y)$ é uma função indicadora, que assume valor 1 se $y \in A$ e 0 caso contrário, sendo A o conjunto de elementos, em que encontra-se o valor de $y = c$; $F(\cdot;\mu,\phi)$ é a função de distribuição acumulada da distribuição Beta; $\alpha = P(y = c)$ é o parâmetro de mistura da distribuição $(0 < \alpha < 1)$. Como a função BI_c tem um ponto de massa em $y = c$, então não pode ser considerada completamente contínua. Note que, com uma probabilidade α, a variável Y é selecionada a partir de um distribuição degenerada em c e, quando a probabilidade é de $(1-\alpha)$, a variável é selecionada de uma distribuição Beta. Isto é, a função de densidade de probabilidade da variável Y é dada pelo valor gerado pela mistura, e escrita na forma:

$$bi_c(y;\alpha,\mu,\phi) = \{\alpha^{I_{\{c\}}(y)}(1-\alpha)^{1-I_{\{c\}^y}}\}\{f(y;\mu,\phi)^{1-I_{\{c\}^y}}\}, \tag{3.2}$$

em que $\alpha > 0$, $0 < \mu < 1$, $\phi > 0$ e $f(y;\mu,\phi)$ é a função densidade apresentada na Equação (2.3) (FERRARI; CRIBARI-NETO, 2004). Se $\alpha > 0$, a massa de probabilidade da

distribuição Beta no ponto $y = c$ é excedida, ou seja, a probabilidade de se observar $y = 0$ ou $y = 1$ é $\alpha = P(y = c)$. Observe que o primeiro termo da distribuição apresentada na Equação (3.2) depende apenas de α e o segundo termo depende de (μ, ϕ), pois envolve a parte contínua da variável resposta (MARTÍNEZ, 2008).

A esperança e a variância da variável Y que segue a distribuição Beta inflacionada são dadas, respectivamente, por: $E(y) = \alpha c + (1 - \alpha)\mu$ e $Var(y) = (1 - \alpha)\frac{\mu(1-\mu)}{1+\phi}$. Sendo assim, a partir da $\widehat{E}(y) = \widehat{\alpha}c + (1 - \widehat{\alpha})\widehat{\mu}$ a resposta do modelo Beta inflacionado é estimado. A distribuição apresentada na Equação (3.2) para Y no intervalo [0, 1) é denominada de distribuição Beta inflacionada no ponto zero (BIZ), denotada por $y \sim BIZ(\alpha, \mu, \phi)$. Caso o ponto de massa for igual a um a distribuição apresentada na Equação (3.2) será denominada de distribuição Beta inflacionada de um (BIU) e escrita $y \sim BIU(\alpha, \mu, \phi)$. Neste trabalho, será discutido o caso em que os valores observados estão no intervalo [0, 1) $(0 \leq y < 1)$.

3.2.2 Modelo de regressão Beta inflacionado de zeros

Sejam Y_1,...,Y_n variáveis aleatórias independentes, em que cada Y_i, i=1,...,n, segue a distribuição Beta inflacionada no ponto c (c = 0 ou c = 1) conforme Equação (3.1), isto é $Y_t \sim BI_c(\alpha_t, \mu_t, \phi)$. O modelo de regressão Beta inflacionado (RBI_c) é definido pelos componentes sistemáticos:

$$\begin{aligned} h(\alpha_t) &= \sum_{i=1}^{M} z_{ti}\gamma_i = \zeta_t \\ g(\mu_t) &= \sum_{i=1}^{m} x_{ti}\beta_i = \eta_t \end{aligned} \tag{3.3}$$

sendo que z_{t1},...,z_{tM} e x_{t1},...,x_{tm} são observações de variáveis regressoras conhecidas e $M + m < n$. Essas variáveis regressoras, podem coincidir totalmente ou parcialmente. Os vetores $\gamma = (\gamma_1,...,\gamma_M)^\intercal$ e $\beta = (\beta_1,...,\beta_m)^\intercal$ são vetores de parâmetros de regressão desconhecidos e pertencem a $\mathbb{R}^M$ e $\mathbb{R}^m$, respectivamente.

As funções de ligação $h : (0, 1) \to \mathbb{R}$ e $g : (0, 1) \to \mathbb{R}$ são estritamente monótonas e duas vezes diferenciáveis, para α e μ. Entre as funções de ligação mais utilizadas para

α e μ estão a função logit: $g(\mu_i) = \log \dfrac{\mu_i}{1-\mu_i}$, a função probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$, em que $\Phi(\cdot)$ é a função de distribuição normal padrão, a complementar log-log: $g(\mu_i) = \log\{-log(1-\mu_i)\}$, a função log-log: $g(\mu_i) = -log\{-log(\mu_i)\}$, entre outras. Uma discussão mais detalhada pode ser observada em (ANDRADE, 2007).

No conjunto de dados de incidência de cancro cítrico em folhas de laranja doce, variedade Pera, utilizou-se para a construção do modelo de regressão da parte discreta a função de ligação *probit*, no qual sua inversa é dada por: $\alpha_t = \Phi\{\zeta_t\}$ e, para o modelo contínuo, usou-se a função de ligação *log*, em que sua inversa é dada por: $\mu_t = \exp\{\eta_t\}$. Já para o conjunto de dados de incidência de cancro cítrico em folhas de laranja doce, variedade Natal, foi utilizado para os modelos discreto e contínuo a função de ligação *logit* e, sua inversa é escrita: $\alpha_t = \frac{\exp(\zeta_t)}{1+\exp(\zeta_t)}$ e $\mu_t = \frac{\exp(\eta_t)}{1+\exp(\eta_t)}$, para os respectivos modelos.

Observe que μ_t é a média condicional de y_t para $y \in (0,1)$ e ϕ o parâmetro de dispersão, neste trabalho, considerado constante para todas as observações.

Para o modelo RBI_c a estimação do vetor de parâmetros $\theta = (\gamma^\intercal, \beta^\intercal, \phi)^\intercal$, pode ser realizada pelo método de máxima verossimilhança cuja função de verossimilhança é dada por:

$$L(\theta) = \prod_{t=1}^{n} bi_c(y_t; \alpha_t, \mu_t, \phi) = L_1(\gamma) L_2(\beta, \phi) \tag{3.4}$$

sendo que

$$L_1(\gamma) = \prod_{t=1}^{n} \alpha_t^{\mathrm{I}_{\{c\}}(y_t)} (1-\alpha_t)^{1-\mathrm{I}_{\{c\}}(y_t)}.$$

$$L_2(\beta, \phi) = \prod_{t:y_t \in (0,1)} f(y_t, \mu_t, \phi).$$

Os parâmetros α_t e μ_t são definidos como funções de γ e β, a partir da Equação (3.3), ou seja, $\alpha_t = h^{-1}(\zeta_t)$ e $\mu_t = g^{-1}(\eta_t)$. Sendo que $\mathrm{I}_A(y)$ é uma função indicadora, em que irá assumir o valor 1 se $y \in A$ e 0 caso contrario e A é o conjunto de elementos onde encontra-se o valor de $y = c$. Além disso, observa-se que a função de verossimilhança $L(\theta)$ foi fatorada em dois termos, o primeiro termo $L_1(\gamma)$, depende do vetor de parâmetro γ é o componente discreto, o qual envolve os parâmetros utilizados para modelar a probabilidade de ocorrência de zero ou de um. O segundo termo $L_2(\beta, \phi)$, depende do vetor de parâmetros β e de ϕ e, envolve os parâmetros usados para modelar a distribuição

condicional da variável resposta pertence ao intervalo (0, 1). Sendo assim, os vetores de parâmetros γ e $(\beta^\intercal, \phi)^\intercal$ são separáveis, com isso, é possível obter as funções escore tanto para γ quanto para $(\beta^\intercal, \phi)^\intercal$ de forma independente.

Considerando a Equação (3.4), o logaritmo da função de verossimilhança para $\theta = (\gamma^\intercal, \beta^\intercal, \phi)^\intercal$ é dado por

$$l(\theta) = l_1(\gamma) + l_2(\beta, \phi), \tag{3.5}$$

onde

$$l_1(\gamma) = \sum_{t=1}^{n} l_t(\alpha_t),$$

$$l_2(\beta,\phi) = \sum_{t:y_t \in (0,1)} l_t(\mu_t,\phi),$$

em que

$$l_t(\alpha_t) = \mathrm{I}_{\{c\}}(y_t)\log(\alpha_t) + (1 - \mathrm{I}_{\{c\}}(y_t))\log(1 - \alpha_t),$$

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

$l_t(\alpha_t)$ é a função de log-verossimilhança de um modelo linear generalizado cuja resposta é binaria, enquanto, $l_t(\mu_t, \phi)$ é a função de log-verossimilhança de um modelo de regressão Beta cuja variável resposta pertence a um intervalo contínuo e aberto (0, 1) (FERRARI; CRIBARI-NETO, 2004). $\mathrm{I}_{\{c\}}(y_t)$ é uma variável aleatória que segue distribuição Bernoulli, com $t = 1,...,n$; $P(\mathrm{I}_{\{c\}}(y_t) = 1) = \alpha_t$ esta associada às covariáveis a partir de um preditor linear ζ_t e uma função de ligação h apresentada na Equação (3.3).

Para obtenção da função escore é necessário fazer uma diferenciação da função de log-verossimilhança em relação a cada um dos parâmetros, além disso, devido a separabilidade dos vetores de parâmetros γ e $(\beta^\intercal,\phi)^\intercal$, é possível obter de forma independente as funções escores.

Desta forma, para $R = 1,...,M$ a função escore para γ é dada por

$$U_R = \frac{\partial l_1(\gamma)}{\partial \gamma_R} = \sum_{t=1}^{n} \frac{\partial l_t(\alpha_t)}{\partial \alpha_t} \frac{\mathrm{d}\alpha_t}{\mathrm{d}\zeta_t} \frac{\partial \zeta_t}{\partial \gamma_R},$$

sendo,

$$\frac{\partial l_t(\alpha_t)}{\partial \alpha_t} = \frac{\mathrm{I}_{\{c\}}(y_t) - \alpha_t}{\alpha_t(1-\alpha_t)},$$

$$\frac{\mathrm{d}\alpha_t}{\mathrm{d}\zeta_t} = \frac{\mathrm{d}h^{-1}(\zeta_t)}{\mathrm{d}\zeta_t} = \frac{1}{h'(\alpha_t)},$$

e

$$\frac{\partial \zeta_t}{\partial \gamma_R} = z_{tR}$$

logo a função escore para γ pode ser escrita como

$$U_R = \frac{\partial l_1(\gamma)}{\partial \gamma_R} = \sum_{t=1}^{n} \frac{\mathrm{I}_{\{c\}}(y_t) - \alpha_t}{\alpha_t(1-\alpha_t)} \frac{1}{h'(\alpha_t)} z_{tR}, \tag{3.6}$$

e, para $r = 1,...,m$ a função escore para β é dada por

$$U_r = \frac{\partial l_2(\beta,\phi)}{\partial \beta_r} = \sum_{t:y_t\in(0,1)} \frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t} \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} \frac{\partial \eta_t}{\partial \beta_r},$$

onde

$$\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t} = \sum_{t:y_t\in(0,1)} \phi\Big[\log\Big(\frac{y_t}{1-y_t}\Big) - \{\psi(\mu_t\phi) - \psi((1-\mu_t)\phi)\}\Big],$$

$$\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} = \frac{\mathrm{d}g^{-1}(\eta_t)}{\mathrm{d}\eta_t} = \frac{1}{g'(\mu_t)},$$

e

$$\frac{\partial \eta_t}{\partial \beta_r} = x_{tr},$$

sendo que ψ é resultado da primeira derivada de Γ.

Segundo Martínez (2008), definindo

$$y_t^* = \begin{cases} \log(\frac{y_t}{1-y_t}), \text{ se } \quad y_t \in (0,1), \\ 0, \text{ caso contrário.} \end{cases} \tag{3.7}$$

e

$$\mu_t^* = E(y_t^* | \mathrm{I}_{\{c\}}(y_t) = 0) = \psi(\mu_t\phi) - \psi((1-\mu_t)\phi), \tag{3.8}$$

obtêm-se a função escore para o parâmetro β,

$$U_r = \phi \sum_{t=1}^{n} (1 - \mathrm{I}_{\{c\}}(y_t))(y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} x_{tr}. \tag{3.9}$$

Finalmente, a função escore para o parâmetro de precisão ϕ é obtido independentemente de γ, por

$$U_\phi = \frac{\partial l_2(\beta,\phi)}{\partial \phi} = \sum_{t:y_t \in (0,1)} \frac{\partial l_t(\mu_t,\phi)}{\partial \phi} =$$

$$\sum_{t:y_t \in (0,1)} \left\{ \mu_t \left(\log\left(\frac{y_t}{1-y_t}\right) - [\psi(\mu_t\phi) - \psi((1-\mu_t)\phi)] \right) + \log(1-y_t) + \psi(\phi) - \phi((1-\mu_t)\phi) \right\}$$

Para Martínez (2008), definindo

$$s(y_t) = \begin{cases} \log(1-y_t), \text{ se } \quad y_t \in (0,1), \\ 0, \text{ se } \quad y = c, \end{cases}$$

a função escore para ϕ pode ser escrita como:

$$U_\phi = \sum_{t:y_t \in (0,1)} (1 - \mathrm{I}(y_t)) \{ \mu_t(y_t^* - \mu_t^*) + s(y_t) + \psi(\phi) - \psi((1-\mu_t)\phi) \}. \tag{3.10}$$

Os vetores escore de γ e β e a função escore para o parâmetro de precisão ϕ, podem ser escritos, respectivamente, na forma matricial, como:

$$U_\gamma(\gamma) = Z^\intercal PG(y^c - \alpha^*),$$

$$U_\beta(\beta,\phi) = \phi X^\intercal TH(y^* - \mu^*),$$

$$U_\phi(\beta,\phi) = tr(HD^*).$$

Sendo Z uma matriz de valores fixos conhecidos de dimensão $(n \times M)$ onde a $t-$ésima linha é dada por: $z_t^\intercal = (z_{t1},...,z_{tM})$ e, X é uma matriz de valores fixos conhecidos de dimensão $(n \times m)$ em que a $t-$ésima linha é definida como $x_t^\intercal = (x_{t1},...,x_{tm})$. As matrizes diagonais são: $P = \mathrm{diag}\left\{ \frac{1}{\alpha_1(1-\alpha_1)},...,\frac{1}{\alpha_n(1-\alpha_n)} \right\}$, $G = \mathrm{diag}\left\{ \frac{\mathrm{d}\alpha_1}{\mathrm{d}\zeta_1},...,\frac{\mathrm{d}\alpha_n}{\mathrm{d}\zeta_n} \right\}$, T =

$\mathrm{diag}\left\{\frac{\mathrm{d}\mu_1}{\mathrm{d}\eta_1},\ldots,\frac{\mathrm{d}\mu_n}{\mathrm{d}\eta_n}\right\}$ e $H = \mathrm{diag}\{1-\mathrm{I}_{\{c\}}(y_1),\ldots,1-\mathrm{I}_{\{c\}}(y_n)\}$ e os vetores $y^* = (y_1^*,\ldots,y_n^*)^\intercal$, $\mu^* = (\mu_1^*,\ldots,\mu_n^*)^\intercal$, $y^c = (\mathrm{I}_{\{c\}}(y_1),\ldots,I_{\{c\}}(y_n))^\intercal$, $\alpha^* = (\alpha_1,\ldots,\alpha_n)^\intercal$ e $z^* = ((y^* - \mu^*)^\intercal,1)^\intercal$.

Como decorrência da ortogonalidade do vetor de parâmetros γ e $(\beta^\intercal,\phi)^\intercal$, o estimador de máxima verossimilhança de γ é assintoticamente independente dos estimadores dos parâmetros β e ϕ. O estimador de γ é obtido com a solução do sistema não-linear $U_\gamma(\gamma) = 0$. Já o estimador de $(\beta^\intercal,\phi)^\intercal$, é obtido com a solução do sistema não-linear $(U_\beta(\beta^\intercal,\phi)^\intercal)^\intercal, U_\phi(\beta^\intercal,\phi)^\intercal))^\intercal = 0$. Contudo, estes estimadores não possuem forma fechada, sendo assim, podem ser obtidos pela maximização da função de probabilidade logarítmica utilizando um algoritmo de optimização não linear, tal como um algoritmo de Newton ou um algoritmo quasi-Newton (FERRARI; CRIBARI-NETO, 2004).

A matriz de informação de Fisher para o modelo Beta inflacionado é dada por

$$K(\theta) = \begin{pmatrix} K_\gamma(\gamma) & 0 \\ 0 & K_\vartheta(\vartheta) \end{pmatrix} \tag{3.11}$$

em que $K_\gamma(\gamma) = K_{\gamma\gamma} = Z^\intercal QZ$ é uma matriz de informação de Fisher para γ e $Q = GPG = \mathrm{diag}\{q_1,\ldots,q_n\}$ é uma matriz diagonal, com $q_t = p_t(\mathrm{d}\alpha_t/\mathrm{d}\zeta)^2$ e $p_t = 1/[\alpha_t(1-\alpha_t)]$. Adicionalmente,

$$K_\vartheta(\vartheta) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix} \tag{3.12}$$

é a matriz de informação de Fisher para $\vartheta = (\beta^\intercal,\phi)^\intercal$. Sendo que $K_{\beta\beta} = \phi^2 X^\intercal \triangle \mathrm{TW}X$, $K_{\beta\phi} = X^\intercal \triangle \mathrm{T}\varsigma$, $K_{\phi\beta} = K_{\beta\phi}^\intercal$ e $K_{\phi\phi} = tr(\triangle\mathrm{D})$. Em que $\triangle = \mathrm{diag}\{\delta_1,\ldots,\delta_n\}$, $\mathrm{W} = \mathrm{diag}\{\mathrm{w}_1,\ldots,\mathrm{w}_n\}$ e $\mathrm{D} = \mathrm{diag}\{\mathrm{d}_1,\ldots,\mathrm{d}_n\}$ são definidas como matrizes diagonais e o vetor $\varsigma = (\varsigma_1,\ldots,\varsigma_n)^\intercal$. Para $t = 1,\ldots,n$, tem-se $\delta_t = 1-\alpha_t$, $\mathrm{w}_t = \psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)$, $\mathrm{d}_t = \mu_t^2\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)(1-\mu_t)^2 - \psi'(\phi)$ e $\varsigma_t = \phi[\mu_t\psi'(\mu_t\phi) - \psi'((1-\mu_t)\phi)(1-\mu_t)]$.

A inversa da matriz de informação de Fisher é

$$K(\theta)^{-1} = \begin{pmatrix} K_\gamma(\gamma)^{-1} & 0 \\ 0 & K_\vartheta(\vartheta)^{-1} \end{pmatrix} = \begin{pmatrix} K_{\gamma\gamma} & 0 & 0 \\ 0 & K_{\beta\beta} & K_{\beta\phi} \\ 0 & K_{\phi\beta} & K_{\phi\phi} \end{pmatrix} \tag{3.13}$$

onde

$$K_{\gamma\gamma} = (\mathrm{Z}^{\intercal}\mathrm{QZ})^{-1},$$

$$K_{\beta\beta} = (\mathrm{X}^{\intercal}\mathrm{W}_{\beta\beta}\mathrm{X})^{-1}\left\{\mathrm{I}_m + \frac{\mathrm{X}^{\intercal}\mathrm{T}\varsigma\varsigma^{\intercal}\mathrm{T}^{\intercal}\mathrm{X}(\mathrm{X}^{\intercal}\mathrm{W}_{\beta\beta}\mathrm{X})^{-1}}{\mathrm{tr(D)} - \varsigma^{\intercal}\mathrm{T}^{\intercal}\mathrm{X}(\mathrm{X}^{\intercal}\mathrm{W}_{\beta\beta}\mathrm{X})^{-1}\mathrm{X}^{\intercal}\mathrm{T}\varsigma}\right\},$$

$$K_{\beta\phi} = (K_{\phi\beta})^{\intercal} = -[\mathrm{tr(D)} - \varsigma^{\intercal}\mathrm{T}^{\intercal}\mathrm{X}(\mathrm{X}^{\intercal}\mathrm{W}_{\beta\beta}\mathrm{X})^{-1}\mathrm{X}^{\intercal}\mathrm{T}\varsigma]^{-1}(\mathrm{X}^{\intercal}\mathrm{W}_{\beta\beta}\mathrm{X})^{-1}\mathrm{X}^{\intercal}\mathrm{T}\varsigma,$$

$$K_{\phi\phi} = [\mathrm{tr(D)} - \varsigma^{\intercal}\mathrm{T}^{\intercal}\mathrm{X}(\mathrm{X}^{\intercal}\mathrm{W}_{\beta\beta}\mathrm{X})^{-1}\mathrm{X}^{\intercal}\mathrm{T}\varsigma]^{-1}.$$

Os algoritmos de maximização requerem a especificação de valores iniciais para o esquema iterativo. Para isto, a obtenção das estimativas pontuais para os parâmetros do modelo RBIZ será utilizado o pacote *gamlss* do programa estatístico R.

3.2.3 Testes de adequabilidade do modelo

a) Pseudo R^2 de McFadden

Existem várias medidas da "qualidade do ajuste" que podem ser utilizadas. Entre estas, a avaliação do ajuste do modelo pode ser feita com base nos valores estimados para a máxima verossimilhança da amostra. Uma destas, é a pseudo R^2 de McFadden (MCFADDEN, 1973) dada por:

$$\widehat{\rho}^2 = 1 - \frac{l_\theta}{l_0},$$

em que, l_θ é a função da log-verossimilhança do modelo ajustado e l_0 a função da log-verossimilhança do modelo nulo, ou seja, o modelo sem a estrutura de regressão. Segundo Louviere, Hensher e Swait (2000), um modelo é considerando bem ajustado, quando o valor do $\widehat{\rho}^2$ estiver entre 0,2 a 0,4. Domencich e McFadden (1975), realizaram simulações com objetivo de comparar o intervalo do $\widehat{\rho}^2$ com o intervalo do coeficiente de correlação múltipla (R), sendo verificado que o intervalo de $\widehat{\rho}^2$ (0,2 a 0,4) é equivalente ao intervalo de R (0,7 a 0,9).

b) Teste de Hipóteses

Os testes de hipóteses sobre os parâmetros do modelo RBI_c podem ser obtido utilizando-se as propriedades assintóticas do EMV (MARTÍNEZ, 2008). Após o ajuste do modelo RBI_c pode-se realizar teste de hipótese para os paramétrios do modelo de regressão. Para isso, separa-se os vetores de parâmetros $\gamma = (\gamma_1^\intercal,\gamma_k^\intercal)^\intercal$ e $\beta = (\beta_1^\intercal,\beta_k^\intercal)^\intercal$, sendo que $k = 1,..,n$, $\gamma_1 = (\gamma_1,...,\gamma_{M_1})^\intercal$, $\gamma_k = (\gamma_{M_1+1},...,\gamma_M)^\intercal$, $\beta_1 = (\beta_1,...,\beta_{m_1})^\intercal$ e $\beta_k = (\beta_{m_1+1},...,\beta_m)^\intercal$. Para testar a hipótese de que os modelos são iguais, as hipóteses do teste são:

$$\begin{aligned} H_0: &\quad \gamma_1 = \gamma_k = \beta_1 = \beta_k = 0; \\ H_1: &\quad \text{Pelo menos uma das igualdades é diferente.} \end{aligned} \tag{3.14}$$

A estatística do teste da razão de log-verossimilhanças (RV) é dada por

$$RV = 2\{l(\widehat{\gamma},\widehat{\beta},\widehat{\phi}) - l(\widetilde{\gamma},\widetilde{\beta},\widetilde{\phi})\}$$

em que $l(\gamma, \beta, \phi)$ é a função de log-verossimilhança observada na Equação (3.5) e, $(\widetilde{\gamma}^\intercal, \widetilde{\beta}^\intercal, \widetilde{\phi})^\intercal$ é o valor maximizado da função de verossimilhança restrita de $(\gamma^\intercal, \beta^\intercal, \phi)^\intercal$, sendo obtida a partir da hipótese nula (H_0). Em condições usuais de regularidade, tem-se que, sob H_0, a RV segue distribuição qui-quadrado (χ^2) com $M_1 + m_1$ graus de liberdade, o teste RV pode ser realizado usando valores críticos aproximados obtidos dessa distribuição.

3.2.4 Análise residual do modelo RBIZ

Após ter definido o modelo é importante testar a sua validade. Para verificar se um modelo é adequado para a predição, é necessário averiguar algumas suposições. Segundo Levine e Stephan (2003), destacam-se como principais suposições: normalidade, homocedasticidade e independência dos resíduos. Estas suposições são fundamentais, uma vez que toda a inferência estatística aplicada no modelo de regressão baseiam-se nelas, sendo estas não satisfeitas, os resultados do modelo ajustado tornam-se não confiáveis.

Nesta direção, técnicas de diagnósticos são utilizadas para a verificação das suposições, sendo os gráficos de resíduos, os mais utilizados para observar se os dados se comportam de maneira homogênia. Os gráficos também ajudam na avaliação da estabilidade e robustez de resultados inferenciais, visto que os resíduos medem a discrepância entre o modelo ajustado e o conjunto de dados.

A análise dos resíduos do modelo de regressão Beta inflacionado de zeros, segundo Ospina e Ferrari (2012), deve ser dividida em duas partes: uma que irá avaliar separadamente os resíduos dos componentes discreto (r_{pt}^{D}) e contínuo (r_{pt}^{C}) do modelo, em que os autores propõem os resíduos de Pearson padronizados com base nos escores do algoritmo iterativo de Fisher utilizado para estimar os parâmetros, sendo escrito como:

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

em que,

$$r_{pt}^{(D)} = \frac{\mathrm{I}_{\{c\}}(y_t) - \widehat{\alpha}_t}{\sqrt{\widehat{\alpha}_t(1-\widehat{\alpha}_t)(1-\widehat{h}_{1tt}^{*})}} \tag{3.15}$$

e

$$r_{pt}^{(C)} = \frac{y_t^{*} - \widehat{\mu}_t^{*}}{\sqrt{\widehat{w}_t(1-\widehat{\alpha}_t)(1-\widehat{h}_{2tt}^{*})}}, \tag{3.16}$$

sendo que y_t^{*} e μ_t^{*} foram definidos respectivamente nas Expressões (3.7) e (3.8) e, $\widehat{h}_{1tt}^{*}$ e $\widehat{h}_{2tt}^{*}$ são elementos da diagonal principal das matrizes de projeção, para $t = 1,...,n$.

a) Teste de normalidade dos resíduos

A distribuição normal é uma das distribuições mais importantes na estatística, além disso, esta distribuição apresenta alguns atributos matemáticos interessantes que permitem fazer conclusões importantes em diversos resultados teóricos (MEYER, 2000). A normalidade dos resíduos é uma suposição essencial para que os resultados do ajuste do modelo de regressão sejam confiáveis. De acordo com Martínez (2008), para o modelo de RBIZ, os resíduos são obtidos a partir da função $r(y_t, \widehat{E(y_t)})$, definição esta sugerida por (COX; SNELL, 1968).

b) Homocedasticidade dos Resíduos

A presença de homocedasticidade será apresentada a partir do gráfico de resíduos padronizados *versus* valores ajustados para $\widehat{\alpha}$ e $\widehat{\mu}$. Este gráfico é apropriado para examinar esta suposição. Geralmente, a falta de homogeneidade de variâncias tende a produzir um gráfico com forma tendencialmente crescente ou decrescente.

c) Independência dos Resíduos

Souza, Jacobi e Pereira (2005) comentam que o processo de amostragem deve respeitar a casualidade na hora da coleta dos dados para que os resíduos não apresentem correlação por conta do período de tempo. Isto resulta em verificar se os resíduos padronizados são variáveis aleatórias independentes, ou seja, que as observações são independentes uma das outras. Para a sua verificação se faz necessário um gráfico que contenha os valores dos resíduos padronizados e a ordem em que as observações foram coletadas. Espera-se que os valores dos resíduos padronizados estejam em torno de uma media zero.

Quando os resíduos padronizados não se comportam de forma aleatória, ou seja, seguem um padrão acima ou abaixo do resíduo zero, pode-se dizer que os dados não são independentes. Para averiguação da independência do resíduos foi utilizado o Gráfico de resíduos padronizados *versus* ordem das observações

3.3 Aplicações

Para ilustrar a aplicação da regressão Beta inflacionada de zeros (RBIZ), em dados reais, utilizou-se dois conjuntos de dados obtido de planejamentos experimentais implementados na região Noroeste do Estado do Paraná. O primeiro conjunto de dados é referente a incidência de cancro cítrico em folhas de laranja doce, variedade Pera, cujos genótipos foram enxertados em outras variedades de cítricos, este experimento foi iniciado em novembro de 2010.

O segundo conjunto de dados é referente a incidência de cancro cítrico em folhas de laranja doce, variedade Natal, o estudo ocorreu no período de agosto de 2012 a junho de 2013, quando as plantas estavam com seis anos de idade, sendo que as mesmas, já apresentavam incidência de cancro cítrico quando o estudo foi iniciado.

3.3.1 Incidência de cancro cítrico em folhas de laranja doce variedade Pera

Para este experimento, foram empregados como material vegetal nove genótipos de laranja doce, variedade Pera, enxertados sob quatro porta-enxertos diferentes (Tabela 3.1). Os genótipos foram plantados em espaçamento 2,5m x 6,0m, e receberam o manejo do cancro cítrico. Quando as plantas estavam com dois anos de idade aproximadamente, tiveram início as avaliações trimestrais subsequentes visando determinar a incidência de cancro cítrico nas folhas. Foram selecionadas 10 plantas por genótipo, sendo amostrados quatro ramos de cada planta e realizada a contagem total de folhas e de folhas doentes.

Tabela 3.1: Genótipos de laranja doce, variedade Pera e porta-enxertos.

Genótipos	Porta-enxerto
Ipigua-IAC	Limão cravo
IAC	Tangerina sunki
Bianchi	Tangerina cleópatra
IAC 2000	Laranja caipira
Olímpia	
EEL	
G59	
G58	
Arapongas	

O conjunto de dados obtidos nas avaliações trimestrais, foi analisado por Gonçalves-Zuliani (2014) em sua tese de doutorado, a qual considerou as épocas de avaliações como um fator, não levando em conta a correlação existente entre as avaliações nas diferentes épocas. Além disso, a incidência de cancro cítrico não foi detectada em grande parte das observações resultando em uma variável resposta cujos valores variaram no intervalo [0, 1). Dada a quantidade de zeros, a autora optou pela análise não-paramétrica, seguida de um teste de comparações múltiplas.

A metodologia não-paramétrica é formada por diversas técnicas, sendo que a análise realizada na tese de Gonçalves-Zuliani (2014) foi baseada em *Rankings*. Nesse caso, os dados coletados são substituídos por postos, podendo aplicar cálculos usuais e chegar a um teste, ou seja, através da adição de médias declaradas, qualquer uma das metodo-

logias paramétricas de comparação múltiplas de dados podem ser utilizadas (HOBBS, 2009). Posteriormente foi aplicado o teste de Tukey ($p \leq 0{,}05$), utilizando o *software* SAS – *Statistical Analysis System* (SAS *Institute*, Cary, CN, EUA).

Como principais resultados das análises, Gonçalves-Zuliani (2014) destacou que os genótipos enxertados sobre Laranja caipira mostraram menor incidência de folhas doentes, com destaques para os genótipos: Arapongas, G58, Olímpia e EEL. Por outro lado, o porta-enxerto Limão cravo mostrou ser bastante sensível ao patógeno.

Neste capítulo, propõe-se uma análise alternativa, que consiste em modelar a proporção de incidência de cancro cítrico, a partir das covariáveis genótipo e porta-enxerto, utilizado somente a segunda avaliação trimestral.

Inicialmente foi realizada uma análise exploratória, com objetivo de verificar o comportamento da variável incidência em relação a cada genótipo e porta-enxerto, para isso, foram retirados todos os zeros do conjunto de dados. Na Figura (3.1), observa-se que na variável genótipo a categoria Arapongas é possivelmente a mais vulnerável a incidência de cancro cítrico. Já para a Figura (3.2), nota-se que a categoria Laranja Caipira é provavelmente a menos vulnerável ao cancro cítrico.

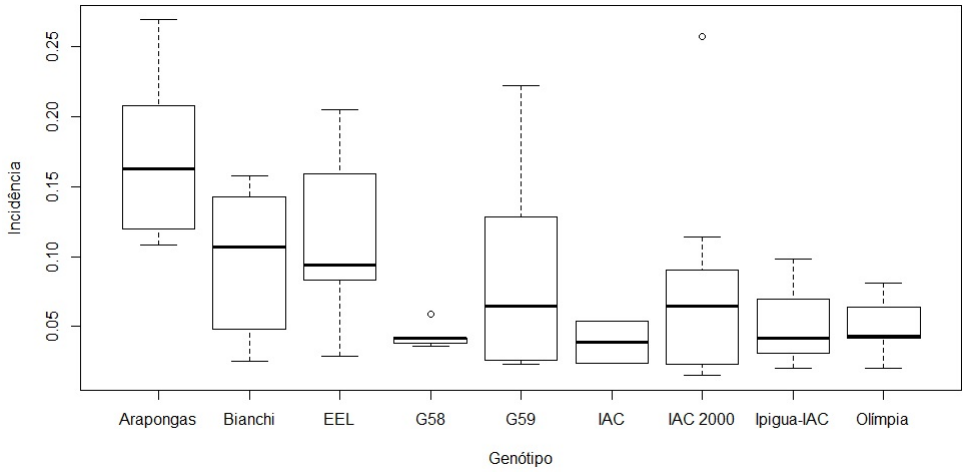

Figura 3.1: Box-plot da variável incidência de fungos em folhas laranjas doces variedade Pera por genótipo.

Para observar a frequência de zeros foi construído um histograma, Figura (3.3), na qual se observa que a distribuição da variável incidência é assimétrica. Além disso, a altura da linha vertical no histograma corresponde a quantidade de zeros na amostra,

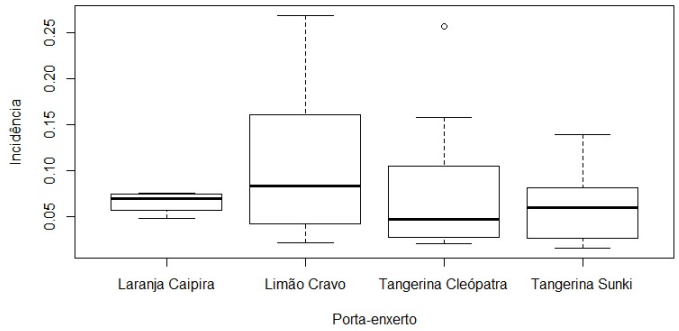

Figura 3.2: Box-plot da variável incidência de fungos em folhas laranjas doces variedade Pera por Porta-enxerto

aproximadamente 85,28% dos dados.

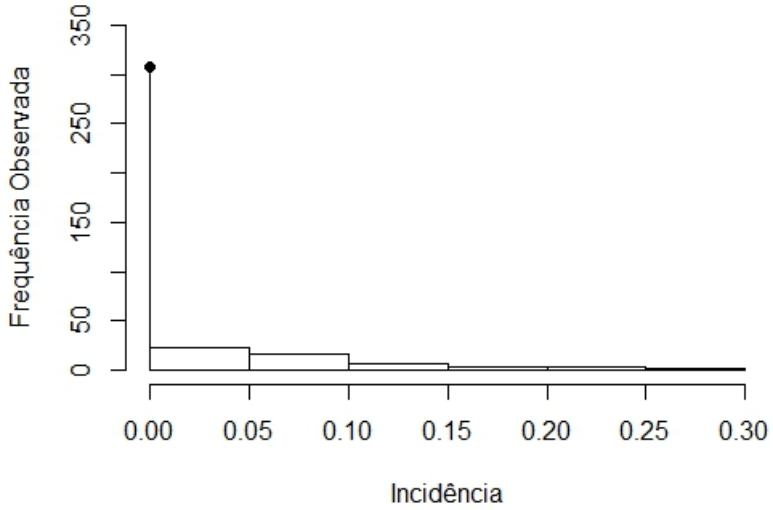

Figura 3.3: Frequência da incidência de fungos em folhas laranjas doces variedade Pera

Analisando a Figura (3.3), observa-se que incidência de cancro cítrico pode ser descrita por uma distribuição Beta inflacionada de zeros. Assim sendo, um modelo RBIZ foi ajustado, em que $Y_i \sim BIZ(\alpha, \mu, \phi)$ conforme Equação (3.2), e as covariáveis porta-enxertos e genótipos foram incluídas no modelo da seguinte forma:

$$
\begin{aligned}
probit(\alpha_t) = \quad & \gamma_0 + \gamma_1 \times \text{Bianchi} + \gamma_2 \times \text{EEL} + \gamma_3 \times \text{IAC} + \gamma_4 \times \text{IAC 2000} \\
& + \gamma_5 \times \text{Ipigua-IAC} + \gamma_6 \times \text{G58} + \gamma_7 \times \text{G59} + \gamma_8 \times \text{Olímpia} \\
& + \gamma_9 \times \text{Limão Cravo} + \gamma_{10} \times \text{Tangerina Cleópatra} \\
& + \gamma_{11} \times \text{Tangerina Sunki}
\end{aligned}
$$

e

$$
\begin{aligned}
log(\mu_t) = \quad & \beta_0 + \beta_1 \times \text{Bianchi} + \beta_2 \times \text{EEL} + \beta_3 \times \text{IAC} + \beta_4 \times \text{IAC 2000} \\
& + \beta_5 \times \text{Ipigua-IAC} + \beta_6 \times \text{G58} + \beta_7 \times \text{G59} + \beta_8 \times \text{Olímpia} \\
& + \beta_9 \times \text{Limão Cravo} + \beta_{10} \times \text{Tangerina Cleópatra} \\
& + \beta_{11} \times \text{Tangerina Sunki.}
\end{aligned}
$$

Estas equações são denominadas de submodelos e representam os componentes discreto e o contínuo, respectivamente, no modelo descrito na Equação (3.2). O genótipo Arapongas e Porta-enxerto Laranja caipira foram utilizadas como referência no modelo de RBIZ.

Para o ajuste destes modelos foi utilizado o pacote *gamlss* do programa R, o qual maximiza a função de log-verossimilhança de forma iterativa utilizando o algoritmo RS, sendo este, uma generalização do algoritmo usado por Rigby e Stasinopoulos ((MARTÍNEZ, 2008) *apud* (RIGBY; STASINOPOULOS, 2005)). A Tabela (3.2) apresenta as estimativas do modelo RBIZ e seus respectivos erros-padrão.

Para avaliar a adequabilidade do modelo o pseudo R^2 de McFadeen foi calculado e o teste da razão de verossimilhança foi aplicado.

O valor da estimativa do pseudo R^2 de McFadeen é $\widehat{\rho}^2 = 0{,}385$ indicando que o ajuste do modelo é adequado. O teste da razão de verossimilhança indicou que há evidências amostrais para a rejeição da hipótese nula ($H_0 : (\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \gamma_7 = \gamma_8 = \gamma_9 = \gamma_{10} = \gamma_{11} = 0); (\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} =$

0)), ao nível de significância usual de 5%, com a estatística teste RV=51,003 e valor $p =$ 0,0024. Logo, conclui-se que pelo menos um dos parâmetros é significativo no modelo.

Tabela 3.2: Estimativas e erros-padrão do modelo de regressão Beta inflacionado de zeros, para a incidência de cancro cítrico em folhas de laranja doce.

Coeficiente $\widehat{\gamma}$	Estimativa	Erro Padrão	Valor p	Coeficiente $\widehat{\beta}$	Estimativa	Erro Padrão	Valor p
Intercepto	1,8974	0,3583	0,0000	Intercepto	-1,8892	0,3518	0,0000
Bianchi	-0,1841	0,3653	0,6146	Bianchi	-0,5312	0,2781	0,0570
EEL	-0,0147	0,3794	0,9692	EEL	-0,4288	0,2728	0,1170
IAC	0,4663	0,4350	0,2845	IAC	-1,1147	0,5225	0,0336
IAC 2000	-0,3787	0,3555	0,2875	IAC 2000	-0,8456	0,2741	0,0022
Ipigua-IAC	0,2716	0,4072	0,5053	Ipigua-IAC	-0,9805	0,4024	0,0153
G58	-0,0107	0,3799	0,9775	G58	-1,0994	0,3272	0,0009
G59	-0,5576	0,3491	0,1112	G59	-0,7041	0,2511	0,0053
Olímpia	-0,4349	0,3553	0,2219	Olímpia	-1,0704	0,2691	0,0001
Limão Cravo	-1,1102	0,2852	0,0001	Limão Cravo	0,2123	0,3181	0,5050
Tang. Cleópatra	-0,6418	0,2974	0,0316	Tang. Cleópatra	0,0313	0,3405	0,9268
Tang. Sunki	-0,7515	0,2934	0,0109	Tang. Sunki	-0,0818	0,3304	0,8046
$\widehat{\phi}$	3,4870	0,1985	0,0000				

Pela Tabela (3.2), observa-se que para os dois submodelos alguns parâmetros não são significativos. Assim sendo, para a seleção dos modelos foi utilizado o critério de informação de Akaike (AIC), com intuito de selecionar o modelo mais adequado ao dados. Para isso, foi utilizado o método automático *stepGAIC* no pacote *gamlss* do R. O modelo mais parcimonioso tem os seguintes submodelos:

$$\begin{aligned} probit(\alpha_t) = \quad & \gamma_0 + \gamma_9 \times \text{Limão Cravo} + \gamma_{10} \times \text{Tangerina Cleópatra} \\ & + \gamma_{11} \times \text{Tangerina Sunki} \end{aligned}$$

e

$$\begin{aligned} log(\mu_t) = \quad & \beta_0 + \beta_1 \times \text{Bianchi} + \beta_2 \times \text{EEL} + \beta_3 \times \text{IAC} + \beta_4 \times \text{IAC 2000} \\ & + \beta_5 \times \text{Ipigua-IAC} + \beta_6 \times \text{G58} + \beta_7 \times \text{G59} + \beta_8 \times \text{Olímpia}, \end{aligned}$$

com valor de AIC = 124,657. O pseudo R^2 de McFadden obtido para este modelo reduzido foi estimado em $\widehat{\rho}^2 = 0{,}271$, sugerindo um bom ajuste. O teste da razão de verossimilhança, indica que não existe evidência de se rejeitar a hipótese nula (H_0 : $(\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \gamma_7 = \gamma_8 = 0); (\beta_9 = \beta_{10} = \beta_{11} = 0)$), ou seja, estes parâmetros não são significativos para o modelo, ao nível de significância usual de 5%,

com a estatística teste de $RV = 15{,}112$ e valor $p = 0{,}3004$.

Na Tabela (3.3) observa-se as estimativas com seus respectivos erros padrões, para o modelo parcimonioso. Para o submodelo $probit(\alpha)$, nota-se que as estimativas dos parâmetros de regressão do componente discreto são significativas para a variável porta-enxerto. Já para o submodelo $log(\mu)$, os parâmetros de regressão do componente contínuo são significantes para os genótipos de laranja doce variedade Pera.

Tabela 3.3: Estimativas e erros-padrão do modelo de regressão Beta inflacionado de zeros.

Coeficiente $\widehat{\gamma}$	Estimativa	Erro Padrão	*p-valor*
Intercepto	1,7013	0,2315	0,0000
Limão Cravo	-1,0442	0,2721	0,0001
Tangerina Cleópatra	-0,5905	0,2851	0,0391
Tangerina Sunki	-0,6884	0,2814	0,0149
Coeficiente $\widehat{\beta}$	Estimativa	Erro Padrão	*p-valor*
Intercepto	-1,7350	0,1670	0,0000
Bianchi	-0,6137	0,2671	0,0222
EEL	-0,4831	0,2718	0,0763
IAC	-1,2617	0,5030	0,0126
IAC 2000	-0,9423	0,2682	0,0005
Ipigua-IAC	-1,1169	0,4031	0,0059
G58	-1,1377	0,3323	0,0007
G59	-0,8093	0,2434	0,0010
Olímpia	-1,0802	0,2684	0,0001
$\widehat{\phi}$	3,4360	0,1986	0,0000

Para a análise dos resíduos foram construídos os gráficos apresentados na Figura (3.4), onde em (a) pode-se observar que os resíduos estão aleatoriamente espalhados em torno de zero, em (b) observa-se que os pontos estão na faixa de -3 a 3 e, em (c) e (d) nota-se que a função de distribuição dos resíduos se aproxima da normal.

Para verificar possíveis pontos aberrantes no modelo RBIZ foram construídos os gráficos dos resíduos r_{pt}^{D} e r_{pt}^{C}, definidos em (3.15) e (3.16), respectivamente. Para o componente discreto ($probit(\alpha)$) deve-se observar a Figura (3.5), gráficos (a) e (b) e, para o componente contínuo, gráficos (c) e (d). Em (a) nota-se que as observações 291, 305, 332 e 340 ultrapassam a faixa de -3 a 3 e são consideradas valores atípicos. Em (c) e (d) observa-se que os pontos estão na faixa de -3 a 3 e não apresentam nenhum ponto

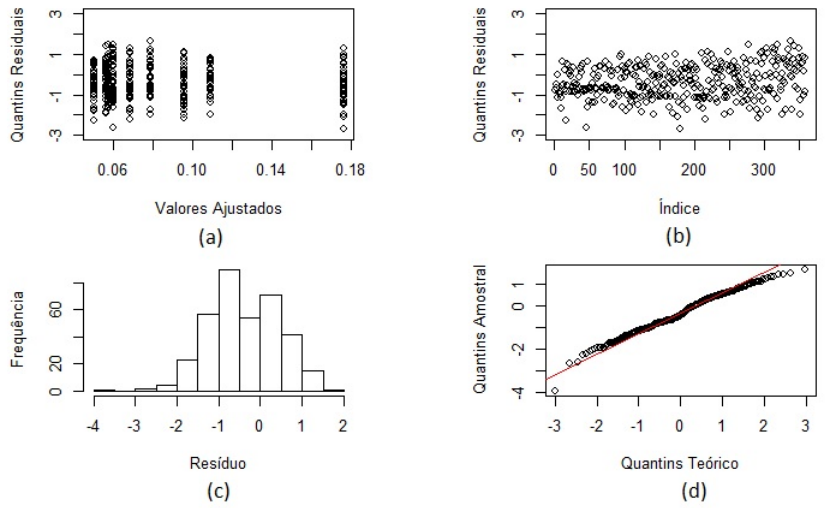

Figura 3.4: Gráficos de Resíduos. (a) Valores Ajustados *versus* Quantil Residual; (b) Índice *versus* Quantil Residual; (c) Histograma dos Resíduos; (d) Q-Qplot.

atípico.

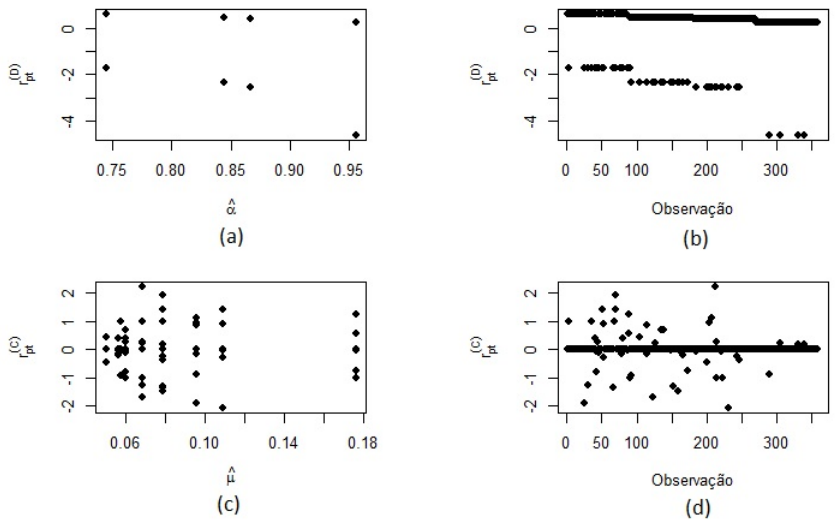

Figura 3.5: Gráficos de Resíduos. (a) e (b) Componente Discreto; (c) e (d) Componente Contínuo.

Com a exclusão das observações (291, 305, 332 e 340), notou-se que as estimativas não diferenciaram muito das obtidas com todas as observações.

3.3.2 Conclusão

De forma geral, o modelo de regressão Beta inflacionado de zeros mostrou-se adequado para descrever a incidência de fungos em folhas de laranja doce, variedade Pera. O componente discreto do modelo explicou o comportamento dos porta-enxertos na incidência do cancro cítrico. Observou-se que o porta enxerto Laranja Caipira foi o que mais contribuiu para o não aparecimento do cancro cítrico, seguido dos porta-enxertos: Tangerina Cleópatra, Tangerina Sunki e Limão Cravo, em que pode ser observado também no box-plot apresentado na Figura (3.2). Para a observação de incidência, explicada pelo componente contínuo do modelo, verificou-se que o genótipo Arapongas é o que mais contribui para a redução de cancro cítrico, seguido dos genótipos IAC, G58 e, os que mais contribuem para o aparecimento da incidência são os genótipos EEL e Bianchi. Neste estudo, em que somente uma das cinco avaliações foi utilizada, chegou-se a resultados semelhantes aos encontrados por Gonçalves-Zuliani (2014), onde a mesma destacou que os genótipos enxertados sobre Laranja caipira apresentaram as menores incidências de cancro cítrico, com destaques para os genótipos Arapongas, G58 e Olímpia. Sendo assim, pode-se dizer que esta metodologia mostrou-se mais adequada para a modelagem de dados de proporção com excesso de zeros, pois possibilitou detectar os genótipos e porta enxertos mais vulneráveis ao cancro cítrico com um número menor de avaliações das plantas e atendendo os pressupostos necessários a modelagem fornecendo, resultados mais confiáveis. Isto representa uma economia de tempo e recursos.

3.3.3 Incidência de cancro cítrico em folhas de laranja doce variedade Natal

O objetivo deste experimento era verificar o efeito de diferentes formas de pulverização na cultura de citrus, para o controle da incidência de cancro cítrico em folhas de laranja doce, variedade Natal. Para isso, o experimento foi composto por nove manejos (tratamentos) combinados a partir dos insumos apresentados na Tabela (3.4), dispostos na forma de um delineamento experimental de blocos casualizados, com sete repetições para cada um dos manejos e, cada unidade experimental foi composta por seis plantas.

Para este experimento foram realizadas cinco avaliações em épocas distintas. Os valores observados para a variável resposta representam a frequência de cancro cítrico e variam no intervalo [0, 1). Os pesquisadores que executaram este experimento realizaram uma análise de variância, para determinar quais dos manejos apresentou melhor efeito na incidência de cancro cítrico, utilizando o teste de *Tukey*, sem considerar que o conjunto de dados estava inflacionado de zeros.

Tabela 3.4: Insumos utilizados para compor os tratamentos do manejo para combate ao cancro cítrico.

Insumo	Ativo	Dose (L p.c./ha)[1]	Dose (L i.a./ha)[2]	Nº de pulverizações
Testemunha	-	-	-	-
Comet	Piraclostrobina	0,300	0,075	2
Tutor	Hidróxido de Cobre	3,500	1,575	-
Kocide	Hidróxido de Cobre	1,500	1,500	2

A modelagem da proporção de incidência de cancro cítrico em folhas de laranjeiras, a partir da covariável tratamento, foi realizada utilizando apenas as observações da quinta avaliação. Inicialmente foi realizada uma análise descritiva para verificar o comportamento da variável incidência em relação a cada um dos tratamentos, para isso, foi retirado todos os zeros do conjunto de dados. Na Figura (3.6), observa-se que os tratamentos A, C e H são os que possivelmente os que mais contribuem para o não aparecimento do cancro cítrico. Já os tratamentos Testemunha, B, D, E e G são os que provavelmente menos contribuem para o não aparecimento do cancro cítrico, ou seja, a evidências de que estes tratamentos não são eficazes para o combate ao cancro cítrico.

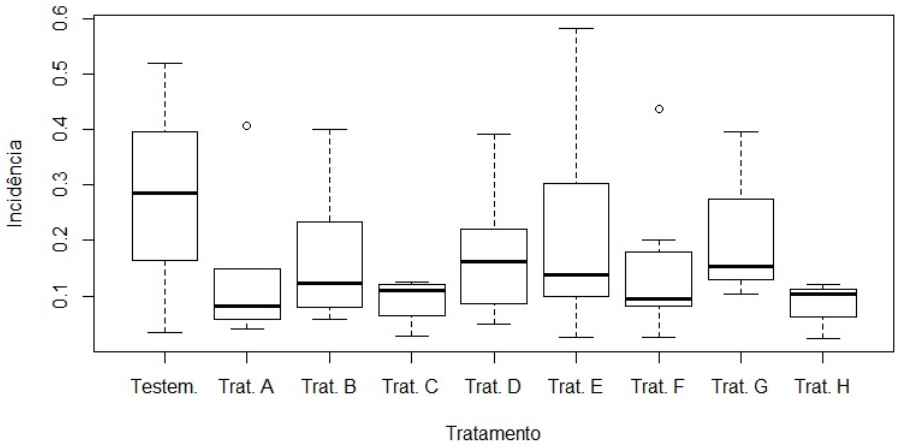

Figura 3.6: Frequência da incidência de fungos em folhas laranjas doces variedade Natal

O histograma foi construído para verificar a frequência de zeros no conjunto de dados, apresentado na Figura (3.7), em que se pode observar que a distribuição da variável incidência é assimétrica a esquerda. Além disso, a barra vertical com um ponto acima no histograma representa a quantidade de zeros na amostra, que corresponde a 76,19% dos dados.

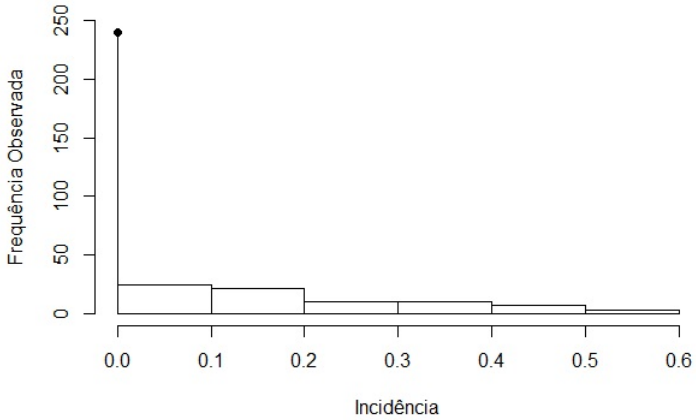

Figura 3.7: Frequência da incidência de fungos em folhas laranjas doces variedade Natal

Analisando a Figura (3.7), observa-se que a variável resposta, incidência de cancro cítrico pode ser descrita por uma distribuição Beta inflacionada de zeros, conforme a Equação (3.2). Sendo assim, o modelo RBIZ foi utilizado para este conjunto de dados e, a inclusão da covariável tratamento no modelo é dada da seguinte forma:

$$\begin{aligned} logit(\alpha_t) = \quad & \gamma_0 + \gamma_1 \times \text{Trat. A} + \gamma_2 \times \text{Trat. B} + \gamma_3 \times \text{Trat. C} \\ & + \gamma_4 \times \text{Trat. D} + \gamma_5 \times \text{Trat. E} + \gamma_6 \times \text{Trat. F} \\ & + \gamma_7 \times \text{Trat. G} + \gamma_8 \times \text{Trat. H} \end{aligned}$$

e

$$\begin{aligned} logit(\mu_t) = \quad & \beta_0 + \beta_1 \times \text{Trat. A} + \beta_2 \times \text{Trat. B} + \beta_3 \times \text{Trat. C} \\ & + \beta_4 \times \text{Trat. D} + \beta_5 \times \text{Trat. E} + \beta_6 \times \text{Trat. F} \\ & + \beta_7 \times \text{Trat. G} + \beta_8 \times \text{Trat. H.} \end{aligned}$$

Estes modelos denominados de submodelos representam os componentes discreto e o contínuo, respectivamente do modelo descrito na Equação (3.2).

A Tabela (3.5) apresenta as estimativas do modelo RBIZ e seus respectivos erros-padrão. Para o ajuste destes modelos foi utilizado o pacote *gamlss* do programa R, em que a função de log-verossimilhança do modelo é maximizada pelo algoritmo RS, sendo este, uma generalização do algoritmo usado por Rigby & Stasinopoulos ((MARTÍNEZ, 2008) *apud* (RIGBY; STASINOPOULOS, 2005)). Para avaliar a adequabilidade do modelo o pseudo R^2 de McFadden e o teste da razão de verossimilhança foram aplicados.

O valor da estimativa do pseudo R^2 de McFadden é $\widehat{\rho}^2 = 0{,}353$, indica que o ajuste do modelo é adequado. O teste da razão de verossimilhança, indicou que existe evidência de se rejeitar a hipótese nula, $(H_0 : (\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \gamma_7 = \gamma_8 = 0); (\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0))$, ao nível de significância usual de 5%, logo, pelo menos um dos parâmetros é significativo, com a estatística teste de RV = 85,708 e valor p = 1,893288e-10.

Para o componente discreto ou submodelo $(logit(\alpha))$, nota-se que as estimativas do parâmetros de regressão foram positivas, indicando que os tratamentos contribuem para o não aparecimento do cancro cítrico. Já para o componente contínuo ou submodelo $(logit(\mu))$, as estimativas do parâmetros de regressão foram negativas indicando, que houve uma redução na incidência de cancro cítrico em folhas de laranjeira (Tabela (3.5)).

Tabela 3.5: Estimativas e erros-padrão do modelo de regressão Beta inflacionado de zeros.

Coef. $\widehat{\gamma}$	Estimativa	Erro Padrão	valor p	Coef. $\widehat{\beta}$	Estimativa	Erro Padrão	valor p
Intercepto	-1,386	0,423	0,001	Intercepto	-0,971	0,129	0,000
Trat. A	2,962	0,616	0,000	Trat. A	-0,827	0,348	0,018
Trat. B	2,037	0,553	0,000	Trat. B	-0,567	0,252	0,025
Trat. C	3,434	0,679	0,000	Trat. C	-1,009	0,431	0,020
Trat. D	2,773	0,598	0,000	Trat. D	-0,542	0,310	0,082
Trat. E	3,178	0,642	0,000	Trat. E	-0,429	0,351	0,222
Trat. F	2,773	0,598	0,000	Trat. F	-0,720	0,320	0,025
Trat. G	3,753	0,737	0,000	Trat. G	-0,271	0,429	0,528
Trat. H	3,753	0,737	0,000	Trat. H	-1,149	0,504	0,023
$\widehat{\phi}$	2,230	0,161	0,000				

Para testar a validade do modelo foram construídos os gráfico de resíduos apresentados na Figura (3.8). No gráfico (a) pode-se observar que os resíduos estão aleatoriamente espalhados em torno de zero, em (b) observa-se que pontos estão na faixa de -3 a 3 e, em (c) e (d) nota-se que a função de distribuição dos resíduos se aproxima da normal.

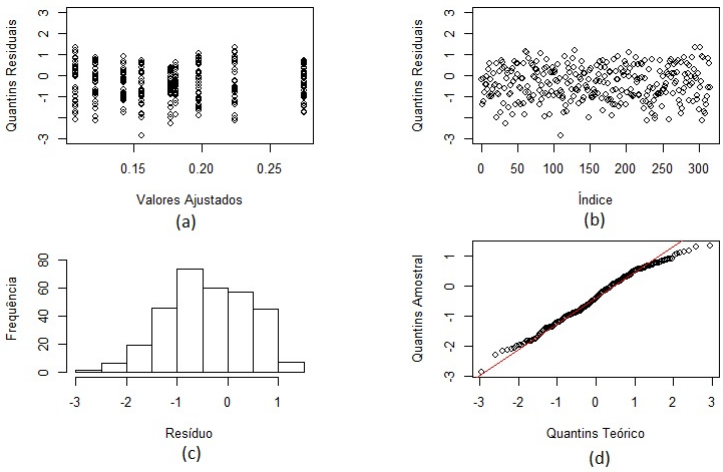

Figura 3.8: Gráficos de Resíduos. (a) Valores Ajustados *versus* Quantil Residual; (b) Índice *versus* Quantil Residual; (c) Histograma dos Resíduos; (d) Q-Qplot.

Para verificar possíveis pontos aberrantes foram construídos os gráficos apresentados na Figura (3.9). Nesta Figura, observa-se os gráficos (a) $r_{pt}^{(D)}$ (componente discreto) e (c) $r_{pt}^{(C)}$ (componente contínuo) *versus* os valores ajustados $\widehat{\alpha}_t$ e $\widehat{\mu}_t$, respectivamente. Note que, para o componente discreto as observações 55, 185, 199, 232 e 310 ultrapassam a faixa de -3 a 3 e são consideradas como valores atípicos, o mesmo é observado no gráfico (b). Para o componente contínuo, nota-se que os resíduos estão no intervalo -3 a 3 e não apresentam nenhum ponto atípico, o mesmo é observado no gráfico (d).

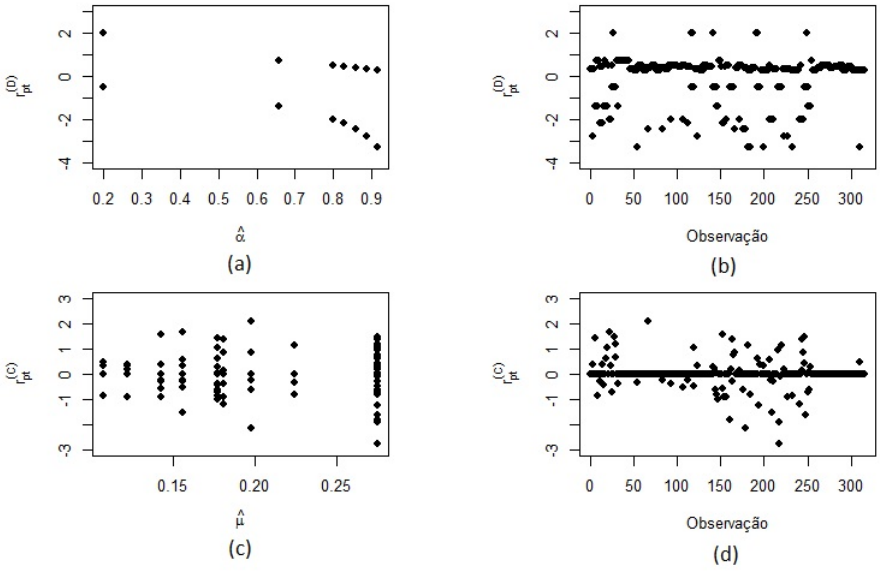

Figura 3.9: Gráficos de Resíduos. (a) e (b) Componente Discreto; (c) e (d) Componente Contínuo.

3.3.4 Conclusão

O modelo de regressão Beta inflacionado de zeros mostrou-se adequado para descrever a incidência de fungos em folhas de laranja doce, variedade Natal. Esta adequação pode ser observada a partir da análise dos resíduos, em que os erros encontram-se de forma aleatória em torno de zero, mostrando assim, a homogeneidade e a independência dos dados. Observou-se que para a ocorrência de zeros, todos os tratamentos contribuem para o não aparecimento do cancro cítrico nas folhas de laranjeira, com destaques para os tratamentos: G, H e C, que mais contribuíram. Já para a observação de incidência, verificou-se que os tratamento: A, C, F e H influenciam para a redução de cancro cítrico e, os tratamentos: G, E, D e B são os que menos contribuem para a redução da incidência. Sendo assim, pode-se dizer que esta metodologia é adequada para a modelagem de dados de proporção com excesso de zeros.

Capítulo 4

Considerações Finais

Neste estudo, apresentamos a distribuição Beta inflacionada e o modelo de regressão Beta inflacionado, conforme proposto por Ferrari e Cribari-Neto (2004) e por Martínez (2008). Esta metodologia é de grande relevância, pois muitas pesquisas resultam em dados que são mensurados em forma de proporções, frações e taxas e podem apresentar uma grande quantidade de zeros e/ou de uns. Nesta situação, a distribuição de mistura Beta inflacionada pode ser utilizada para representar a variável de interesse, em que a parte inflacionada vai ser descrita pela distribuição Bernoulli e a parte contínua vai ser descrita distribuição Beta. Os aspectos da construção dessa distribuição de mistura foi apresentado e exemplificado pela modelagem de um conjunto de dados observados. No modelo de regressão Beta inflacionado, supõe-se que a variável resposta segue a distribuição Beta inflacionada e seus parâmetros são modelados por preditores lineares, utilizando-se funções de ligação que são indicadas a cada caso. Alguns aspectos da estimação dos parâmetros pelo método de máxima verossimilhança foram discutidos e duas aplicações apresentadas.

Para a primeira aplicação utilizou-se de um conjunto de dados referente a incidência de cancro cítrico em folhas de laranja doce, variedade Pera, cujos genótipos foram enxertados em outras variedades de cítricos, o objetivo foi avaliar a influência dos genótipos e porta-enxertos na incidência do cancro cítrico. O modelo de regressão Beta inflacionado de zeros, mostrou-se adequado para descrever a incidência de fungos em folhas de laranja doce. O componente discreto do modelo explicou o comportamento dos porta-

enxertos e, o componente contínuo do modelo explicou o comportamento dos genótipo na incidência do cancro cítrico. Foi também possível determinar quais os porta-enxertos e genótipos mais influenciavam dentro de cada componente.

Para a segunda aplicação foi utilizado um conjunto de dados referente a incidência de cancro cítrico em folhas de laranja doce, variedade Natal, cujo objetivo foi de verificar o efeito de diferentes formas de pulverização na cultura de citrus, para o controle da doença. O modelo de regressão Beta inflacionado de zeros, mostrou-se adequado para descrever a incidência de fungos em folhas de laranja doce, variedade Natal. Observou-se que na ocorrência de zeros, todos os tratamentos contribuem para o não aparecimento do cancro cítrico nas folhas de laranjeira, com destaques para os tratamentos: G, H e C, que mais contribuíram. Já para a observação de incidência, verificou-se que os tratamento: A, C, F e H influenciam para a redução de cancro cítrico. Sendo assim, pode-se dizer que esta metodologia é adequada para a modelagem de dados de proporção com excesso de zeros.

Referências Bibliográficas

AITCHISON, J. On the distribution of a positive random variable having a discrete probability mass at the origin*. *Journal of the american statistical association*, Taylor & Francis Group, v. 50, n. 271, p. 901–908, 1955.

ANDRADE, A. C. G. d. *Efeitos da especificação incorreta da função de ligação no modelo de regressão beta*. Tese (Doutorado) — Universidade de São Paulo, 2007.

BURY, K. *Statistical distributions in engineering*. [S.l.]: Cambridge University Press, 1999.

COOK, D. O.; KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*, Elsevier, v. 15, n. 5, p. 860–867, 2008.

COX, D. R.; SNELL, J. E. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 248–275, 1968.

DOMENCICH, T.; MCFADDEN, D. Urban travel demand: a behavioural approach. *Amsterdam: North-Hollan Publishing Co*, 1975.

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

FEUERVERGER, A. On some methods of analysis for weather experiments. *Biometrika*, Biometrika Trust, v. 66, n. 3, p. 655–658, 1979.

FUMES, G.; CORRENTE, J. Zero-inflated models: applications in a food frequency questionnaire. *Rev. Bras. Biom*, v. 28, n. 1, p. 24–38, 2010.

GONÇALVES-ZULIANI, A. M. O. *Resistência de genótipos de laranja doce* ***Citrus sinensis*** *ao cancro cítrico e diversidade de* ***Xanthomonas Citri subsp.citri***. Tese (Doutorado) — Universidade de Estadual de Maringá, 2014.

GOURIEROUX, C.; MONFORT, A. *Statistics and econometric models*. [S.l.]: Cambridge University Press, 1995.

HALL, D. B. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, Wiley Online Library, v. 56, n. 4, p. 1030–1039, 2000.

HELLER, G.; STASINOPOULOS, M.; RIGBY, B. et al. The zero-adjusted inverse gaussian distribution as a model for insurance claims. In: *Proceedings of the 21th International Workshop on Statistical Modelling*. [S.l.: s.n.], 2006. v. 226233.

HOBBS, G. *Using SAS for Nonparametric Statistics*. 2009.

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

LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, Taylor & Francis, v. 34, n. 1, p. 1–14, 1992.

LEHMANN, E. L.; CASELLA, G. *Theory of point estimation*. [S.l.]: Springer Science & Business Media, 1998.

LEVINE, D. M.; STEPHAN, D. Teoria e aplicações usando o microsoft® excel em português. *Rio de Janeiro: LTC, c2000*, v. 1, 2003.

LOUVIERE, J. J.; HENSHER, D. A.; SWAIT, J. D. *Stated choice methods: analysis and applications*. [S.l.]: Cambridge University Press, 2000.

MARTÍNEZ, R. O. *Modelos de regressao beta inflacionados*. Tese (Doutorado) — Universidade de São Paulo, 2008.

MCCULLAGH, P.; NELDER, J. A. Generalized linear models. London England Chapman and Hall 1983., 1989.

MCFADDEN, D. Conditional logit analysis of qualitative choice behavior. Institute of Urban and Regional Development, University of California, 1973.

MEYER, P. L. Probabilidade: aplicações à estatística. In: *Probabilidade: aplicações à estatística*. [S.l.]: Livro Técnico, 2000.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. *Introduction to Statistical Theory*. [S.l.]: McGraw-Hill, New York, 1974.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. *Applied linear statistical models*. [S.l.]: Irwin Chicago, 1996.

OSPINA, R.; FERRARI, S. L. Inflated beta distributions. *Statistical Papers*, Springer, v. 51, n. 1, p. 111–126, 2010.

OSPINA, R.; FERRARI, S. L. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 6, p. 1609–1623, 2012.

PAOLINO, P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, SPM-PMSAPSA, v. 9, n. 4, p. 325–346, 2001.

PEREIRA, G. H. A. *Modelos de regressão beta inflacionados truncados*. Tese (Doutorado) — Universidade de São Paulo, 2012.

PEREIRA, T. L. *Regressão beta inflacionada: inferência e aplicações*. Tese (Doutorado) — Ph. D. Thesis, Universidade Federal de Pernambuco (UFPE), 2010.

PEREIRA, T. L.; SOUZA, T. C. de; CRIBARI-NETO, F. Modeling administrative efficiency scores of brazilian municipalities: Regional differences. *Ciencia & Natura*, v. 36, n. 3, p. 23–36, 2014.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005.

SOUZA, A. M.; JACOBI, L. F.; PEREIRA, J. E. Gráficos de controle de regressão usando o statistica. *Florianópolis: VisualBooks*, p. 16, 2005.

VASCONCELLOS, K. L.; CRIBARI-NETO, F. Improved maximum likelihood estimation in a new class of beta regression models. *Brazilian Journal of Probability and Statistics*, v. 19, n. 1, p. 13–31, 2005.

VIEIRA, A.; HINDE, J. P.; DEMÉTRIO, C. G. Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, Taylor & Francis, v. 27, n. 3, p. 373–389, 2000.

Apêndice A

Programação

```
##############################################################
#             Ajuste do modelo RBIZ
#         POMPEU, D. S.; GUEDES, T. A.
##############################################################

library(gamlss)                                                # Pacote

dados<- read.table("Av2.csv",sep=';',header=T)
head(dados)
attach(dados)

fit <- gamlss(Incidencia ~ (factor(Genot) + factor(PE)),
nu.formula= ~ factor(Genot) + factor(PE),
family=BEZI(mu.link = "log", nu.link = "probit"),
data=dados)                                        # Modelo ajustado


summary(fit)                                                   # Resultados


logLik(fit)                                                    # Logaritimo da função de verossimilhança


fit2 <- gamlss(Incidencia ~ (factor(Genot)),nu.formula= ~  factor(PE),
method=RS(10),  family=BEZI(mu.link = "log",
nu.link = "probit"), data=dados)                     # Modelo Reduzido
```

```
summary(fit2)                                                    # Resultados


logLik(fit2)                                                     # Logaritimo da função de verossimilhança

stepAIC(fit2)                                                    # Selecionando o modelo mais parcimonioso

# Gráfico para os resíduos

res1  <- residuals(fit2, what=c('z-score'), type=c('weighted'))
res2  <- residuals(fit2, what=c('nu'), type=c('weighted'))       # Resíduo de Alpha
res3  <- residuals(fit2, what=c('mu'), type=c('weighted'))       # Resíduo de Mu
pred1 <- fitted(fit2, what='nu', type=c('responce'))             # Valores preditos de Alpha
pred2 <- fitted(fit2, what='mu', type=c('responce'))             # Valores preditos de mu


result<- cbind(Incidencia, res1, res2,res3,pred1, pred2)
head(result)


# Gráficos para observar possiveis pontos aberrantes

par(mfrow=c(2,2))
par(mar=c(5,5,5,5))                                              # Margem do gráfico


plot(pred1,res2,pch=19,xlab= expression(paste(italic(hat(alpha)))),
ylab=expression(paste(ylab=r[pt]^(D))))                       # Parte Discreta

plot(res2,pch=19,xlab= 'Observação', ylab=expression(paste(ylab=r[pt]^(D))))



plot(pred2,res3,pch=19,xlab= expression(paste(italic(hat(mu)))),
ylab=expression(paste(ylab=r[pt]^(C))))                       # Parte contínua


plot(res3,pch=19,xlab= 'Observação', ylab=expression(paste(ylab=r[pt]^(C))))

# Gráficos de Homogênidade, Independência e Normalidade


plot(fitted(fit2),fit2$residuals,ylim=c(-3,3), ylab='Quantins Residuais',xlab="Valores Ajustados")
plot(fit2$residuals, ylim=c(-3,3), ylab='Quantins Residuais',xlab="Índice")
```

```
x<- residuals(fit2)
hist(x, ylab='Frequência',xlab="Resíduo",ylim=c(0,90), main='')
qqnorm(x, main="", ylab="Quantins Amostral", xlab='Quantins Teórico')
qqline(x,col="red")
```

Apêndice B

Neste apêndice, apresentamos cálculos detalhados desenvolvidos para obtenção das equações de estimação descritas na primeira parte deste trabalho.

B.1 Esperança e variância

A esperança de uma variável aleatória da variável aleatória com distribuição Beta dada por

$$\begin{aligned}
E[X^k] =& \frac{1}{B(a,b)} \int_0^1 x^{k+a-1}(1-x)^{b-1} dx \\
=& \frac{B(k+a,b)}{B(a,b)} = \frac{\Gamma(k+a)\Gamma(b)}{\Gamma(k+a+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\
=& \frac{\Gamma(k+a)\Gamma(a+b)}{\Gamma(a)\Gamma(k+a+b)} \\
=& \frac{a}{a+b}
\end{aligned}$$

Sabendo que a variância de X é obtida por

$$Var(X) = E[X^2] - (E[X])^2$$

então

$$
\begin{aligned}
Var(X) =& \frac{\Gamma(a+2)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+2)} - (\frac{a}{a+b})^2 \\
=& \frac{(a+1)a}{(a+b+1)(a+b)} - (\frac{a}{a+b})^2 \\
=& \frac{ab}{(a+b+1)(a+b)^2}
\end{aligned}
$$

B.2 Regressão Beta

As contas detalhadas para obtenção da equação descrita em 2.3, sabemos que $\mu = \frac{a}{a+b}$ e $\phi = a+b$, colocando a em evidência temos então

$$\mu = \frac{a}{\phi},$$

logo $a = \mu\phi$

O mesmo será feito para sabermos quem é b, para isso temos que

$$
\begin{aligned}
\mu =& \frac{a}{a+b} \\
=& \frac{\mu\phi}{\mu\phi + b} \\
\mu(\mu\phi + b) =& \mu\phi \\
\mu^2\phi + \mu b =& \mu\phi \\
\mu b =& \mu\phi - \mu^2\phi \\
\mu b =& \mu(\phi - \mu\phi) \\
b =& \phi(1-\mu)
\end{aligned}
$$

B.3 Esperança e variância da distribuição beta inflacionada no ponto zero (BIZ)

A esperança da distribuição BIZ é dada por

$$
\begin{aligned}
E(Y) =& \sum_{y=c} yP(y=c) + \int_0^1 y(1-\alpha)f(y,\mu,\phi)dy \\
=& c\alpha + (1-\alpha)\int_0^1 yf(y,\mu,\phi)dy \\
=& c\alpha + (1-\alpha)\mu
\end{aligned}
$$

Sendo que a $\int_0^1 yf(y,\mu,\phi)dy = E(Y) = \mu$.

A variância da distribuição BIZ é obtida apartir da esxpressão

$$
Var(Y) = E(y^2) - [E(y)]^2
$$

logo

$$
\begin{aligned}
E(y^2) =& \sum_{y=c} y^2P(y=c) + \int_0^1 y^2(1-\alpha)f(y,\mu,\phi)dy \\
=& c^2\alpha + (1-\alpha)\int_0^1 y^2f(y,\mu,\phi)dy \\
=& E(y^2) = c^2\alpha + (1-\alpha)[\frac{V(\mu)}{\phi+1} + \mu^2]
\end{aligned}
$$

Sendo que $\int_0^1 y^2f(y,\mu,\phi)dy = E(Y^2) = [\frac{V(\mu)}{\phi+1} + \mu^2]$.

$$
\begin{aligned}
Var(Y) =& c^2\alpha + (1-\alpha)[\frac{V(\mu)}{\phi+1} + \mu^2] - [c\alpha + (1-\alpha)\mu]^2 \\
=& c^2\alpha + \frac{V(\mu)}{\phi+1} - \alpha(\frac{V(\mu)}{\phi+1}) + \mu^2 - \alpha\mu^2 - [(c\alpha + (1-\alpha)\mu)(c\alpha + (1-\alpha)\mu)] \\
=& \frac{V(\mu)}{\phi+1}(1-\alpha) + c^2\alpha + \mu^2 - \alpha\mu^2 - [c^2\alpha^2 + \alpha c\mu - \alpha^2 c\mu \\
& + \mu^2 + \mu c\alpha - \alpha\mu^2 - \alpha^2 c\mu - \alpha\mu^2 + \alpha^2\mu^2] \\
=& \frac{V(\mu)}{\phi+1}(1-\alpha) + c^2\alpha + \mu^2 - \alpha\mu^2 - [-2\alpha^2 c\mu + 2\alpha c\mu - 2\alpha\mu^2 + \alpha^2 c^2 + \mu^2 + \alpha^2\mu^2] \\
=& \frac{V(\mu)}{\phi+1}(1-\alpha) + c^2\alpha + \mu^2 - \alpha\mu^2 + 2\alpha^2 c\mu - 2\alpha c\mu + 2\alpha\mu^2 - \alpha^2 c^2\alpha - \mu^2 - \alpha^2\mu^2 \\
=& \frac{V(\mu)}{\phi+1}(1-\alpha) + c^2 + \alpha\mu^2 + 2\alpha^2 c\mu - 2\alpha c\mu - \alpha^2 c^2 - \alpha^2\mu^2 \\
=& \frac{V(\mu)}{\phi+1}(1-\alpha) + \alpha(1-\alpha)(c-\mu)^2
\end{aligned}
$$

B.4 Matriz de informação de Fisher para a distribuição beta inflacionada

Para a obtenção da matriz de informação de Fisher é necessário calcular a derivada de segunda ordem do logaritmo da função de verossimilhança 2.12. A matriz é escrita.

$$
J(\widehat{\theta}) = \begin{bmatrix} \frac{\partial^2 l_1}{\partial\alpha^2} & 0 & 0 \\ 0 & \frac{\partial^2 l_2}{\partial\mu^2} & \frac{\partial^2 l_2}{\partial^2(\mu\phi)^2} \\ 0 & \frac{\partial^2 l_2}{\partial^2(\mu\phi)^2} & \frac{\partial^2 l_2}{\partial\phi^2} \end{bmatrix}
$$

onde, a esperança da derivada de segunda ordem para α é

$$
\mathrm{E}[-\frac{\partial^2 l_1(\theta)}{\partial_\alpha}] = \mathrm{E}[-\frac{\mathrm{T}_1}{\alpha^2} - \frac{(n-\mathrm{T}_1)}{(1-\alpha)^2}] = \mathrm{E}[\frac{\mathrm{T}_1 + 2\alpha\mathrm{T}_1 - n\alpha^2}{\alpha^2(1-\alpha)^2}] = \frac{\mathrm{E}[\mathrm{T}_1] + 2\alpha\mathrm{E}[\mathrm{T}_1] - n\alpha^2}{\alpha^2(1-\alpha)^2}
$$

Sabendo que, $\mathrm{T}_1 = \sum_{t=1}^{n} \mathrm{I}_{\{c\}(y_t)} \sim \mathrm{B}(n,\alpha)$, logo, $\mathrm{E}[\mathrm{T}_1] = n\alpha$.

Com isso, $\mathrm{E}[-\dfrac{\partial^2 l_1(\theta)}{\partial_\alpha}] = \dfrac{\mathrm{E}[\mathrm{T}_1] + 2\alpha\mathrm{E}[\mathrm{T}_1] - n\alpha^2}{\alpha^2(1-\alpha)^2}$

$$= \frac{n\alpha + 2\alpha(n\alpha) - n\alpha^2}{\alpha^2(1-\alpha)^2} = \frac{n\alpha + 2n\alpha^2 - n\alpha^2}{\alpha^2(1-\alpha)^2}$$

$$= \frac{n\alpha(1-\alpha)}{\alpha^2(1-\alpha)^2} = \frac{n}{\alpha(1-\alpha)}$$

logo, $\dfrac{\partial^2 l_1(\theta)}{\partial_\alpha} = \dfrac{n}{\alpha(1-\alpha)}$

A esperança da derivada de segunda ordem para μ é.

$$\mathrm{E}[-\frac{\partial^2 l_2(\theta)}{\partial_\mu}] = \mathrm{E}[\phi(n-\mathrm{T}_1)[\psi'((1-\mu)\phi)(-\phi) - \psi'(\mu\phi)\phi]$$

$$= \mathrm{E}[\phi^2(n-\mathrm{T}_1)[-\psi'((1-\mu)\phi) - \psi'(\mu\phi)]]$$

$$= \phi^2[-\psi'((1-\mu)\psi) - \phi'(\mu\phi)]\mathrm{E}[(n-\mathrm{T}_1)]$$

$$= \phi^2[-\psi'((1-\mu)\phi) - \psi'(\mu\phi)](n-n\alpha)$$

$$= \phi^2[-\psi'((1-\mu)\phi) - \psi'(\mu\phi)](1-\alpha)n$$

.

logo, $\dfrac{\partial^2 l_2(\theta)}{\partial_\mu} = n\phi^2(1-\alpha)[\psi'((1-\mu)\phi) + \psi'(\mu\phi)]$

A esperança da derivada de segunda ordem para ϕ é escrita.

$$\mathrm{E}[\frac{\partial^2 l_2(\theta)}{\partial_\phi}] = \mathrm{E}[(n-\mathrm{T}_1)\{\psi'(\phi) - \mu^2\psi'(\mu\phi) + (1-\mu)^2\psi'((1-\mu)\phi)\}]$$

$$= \mathrm{E}(n-\mathrm{T}_1)\{\psi'(\phi) - \mu^2\psi'(\mu\phi) + (1-\mu)^2\psi'((1-\mu)\phi)\}$$

$$= (n - n\alpha)\{\psi'(\phi) - \mu^2\psi'(\mu\phi) + (1-\mu)^2\psi'((1-\mu)\phi)\}$$

$$= n(1-\alpha)\{\psi'(\phi) - \mu^2\psi'(\mu\phi) + (1-\mu)^2\psi'((1-\mu)\phi)\}$$

logo, $\dfrac{\partial^2 l_2(\theta)}{\partial_\phi} = n(1-\alpha)\{-\psi'(\phi) + \mu^2\psi'(\mu\phi) - (1-\mu)^2\psi'((1-\mu)\phi)\}$

Para os parâmetros $\mu\phi$.

$$\frac{\partial^2 l_2(\theta)}{\partial_{\mu\phi}} = \frac{\partial^2 l_2(\theta)}{\partial_{\phi\mu}}$$

Sabe-se que as derivadas de μ e ϕ, respectivamente são

$$U_2 = \frac{\partial l_2(\mu\phi)}{\partial\mu} = \phi(n - \mathrm{T}_1)[\psi((1-\mu)\phi) - \psi(\mu\phi) + \mathrm{T}_2 + \mathrm{T}_3,$$

$$U_3 = \frac{\partial l_2(\mu\phi)}{\partial\phi} = (n - \mathrm{T}_1)[\psi(\phi) - \mu\psi(\mu\phi) - (1-\mu)\phi((1-\mu)\phi)] + n\mathrm{T}_2 - (1-\mu)\mathrm{T}_3.$$

então,

logo, $\dfrac{\partial^2 l_2(\theta)}{\partial_{\mu\phi}} = (1-\alpha)\phi\{\psi^{'}(\mu\phi)\mu - \psi^{'}((1-\mu)\phi)(1-\mu)\}.$