Universidade Federal de Juiz de Fora
Departamento de Estatística
Curso de Estatística

Vanessa Felix do Nascimento Sergio

Utilização das Distribuições Inflacionadas de Zeros no Monitoramento da Qualidade do Leite

Juiz de Fora
2012

Vanessa Felix do Nascimento Sergio

Utilização das Distribuições Inflacionadas de Zeros no Monitoramento da Qualidade do Leite

Monografia apresentada ao Curso de Estatística da UFJF, como requisito para a obtenção do grau de Bacharel em Estatística.

Orientador: Clécio da Silva Ferreira
Doutor em Estatística pela Universidade de São Paulo

Juiz de Fora
2012

Sergio, Vanessa
Utilização das Distribuições Inflacionadas de Zeros no Monitoramento da Qualidade do Leite
/ Vanessa Sergio – 2012
43 .p

CDU N/A

Vanessa Felix do Nascimento Sergio

Utilização das Distribuições Inflacionadas de Zeros no Monitoramento da Qualidade do Leite

Monografia apresentada ao Curso de Estatística da UFJF, como requisito para a obtenção do grau de Bacharel em Estatística.

Aprovada em 26 de Outubro de 2012.

BANCA EXAMINADORA

_____

Clécio da Silva Ferreira

DSc/IME – USP

_____

Profª Drª Angela Mello Coelho

DSc/ESALQ – USP

_____

Alfredo Chaoubah

DSc/PUC – Rio

Agradecimentos

A Deus, por me fazer forte para seguir em frente apesar das inúmeras dificuldades.

Aos meus pais, pela compreensão, paciência e incentivo que me ajudaram a chegar ao fim de mais uma caminhada.

A minha colega Leiliane por me ajudar nos momentos de dificuldades durante a faculdade e pelas conversas.

A Jéssica pela grande amizade que construímos desde o inicio do curso, pelos momentos de alegria e conselhos que sempre me ajudaram.

Aos meus colegas de turma que foram, de certa forma, companheiros durante todos esses anos de graduação.

Ao professor Alfredo pela confiança depositada em mim em vários momentos, que me ajudaram muito a me desenvolver profissionalmente.

Ao meu orientador professor Clécio pela supervisão, orientação e por acreditar no potencial deste trabalho.

Ao Cristiano e a EMBRAPA pela confiança depositada, e pelo incentivo na produção deste inicio de pesquisa.

Resumo

A contagem de células somáticas (CCS) no leite é o indicador mais usado em programas de controle e prevenção da mastite em todo o mundo. A mastite é uma doença endêmica que causa danos ao animal e a qualidade do leite.

O objetivo desse trabalho é identificar o comportamento da variável CCS, encontrando uma distribuição de probabilidade que se ajuste bem aos dados e observar seu comportamento em relação às covariáveis idade e dias de lactação. A utilização das distribuições inflacionadas de zeros se deve a uma característica da variável de interesse, o excesso de zeros, pois o valor zero em casos de contaminação é uma observação muito importante, pois pode representar o melhor animal em termos genéticos.

Com o auxilio do pacote GAMLSS do software estatístico R, pode-se comparar a qualidade dos ajustes de duas distribuições, a Poisson Inflada de Zeros – ZIP e a Binomial Negativa Inflada de Zeros – ZINB. Observamos que a ZINB apresentou melhor ajuste aos dados.

Palavras-Chave: Excessos de zeros, contagem de células somáticas (CCS), ZIP, ZINB.

Abstract

The somatic cell count (SCC) in milk provides a more programs used in the control and prevention of mastitis worldwide. Mastitis is an endemic disease which causes damage to the animal and milk quality.

The objective of this work is to identify the behavior of the variable CCS, finding a probability distribution that fits the data well and observe their behavior towards the covariates age and days in milk. The use of zero inflated distributions should be variable of interest characteristic, the excess of zero, because the value zero in contamination cases is a very important observation because it can represent the best animal genetically.

With the aid of the package GAMLSS from statistical software R, we can compare the fit quality of two distributions, the Zero Inflated Poisson - ZIP and Zero Inflated Negative Binomial - ZINB. We observed that the ZINB had better fit to the data.

Keywords: Excess zeros, somatic cell count (SCC), ZIP, ZINB.

Sumário

Capítulo 1

Introdução

A indústria de leite e derivados no país está passando de importadora para exportadora do produto. Em 2000, o país importou 373 milhões de dólares em leite e derivados. Em 2003, importou 112 milhões de dólares em leite e derivados, isto é, houve redução de 70% nos valores de importação (Anuário, 2004).

No Brasil, a análise da CCS foi introduzida em 1991 pelo Programa de Análise de Rebanhos Leiteiros do Paraná (PARLPR), da Associação Paranaense de Criadores de Bovinos da Raça Holandesa (APCBRH).

As células somáticas são células brancas ou leucócitos presentes no leite em resposta a danos do tecido, resultado de um processo natural de reposição de células velhas dentro da glândula mamária, ou ainda, de lesão física (Harmon, 1998).

A contagem de células somáticas (CCS) no leite é o indicador mais usado em programas de controle e prevenção da mastite em todo o mundo. Considerando que a mastite é uma doença endêmica em rebanhos leiteiros, a avaliação periódica da saúde do úbere baseada em análises de laboratório para CCS e/ou a identificação dos agentes infecciosos permite maior sucesso no controle e na prevenção da doença. Portanto, com o suporte laboratorial, pode-se definir limites aceitáveis de contaminação do leite e identificar animais com infecções. Com o objetivo de diminuir a ocorrência da mastite no rebanho, esses exames podem auxiliar na tomada de decisões, como antecipar o tratamento à secagem, fazer o descarte dos animais-problema e realizar treinamento dos ordenhadores quanto à utilização e manutenção do equipamento de ordenha (Coentrão et al., 2008).

A quantidade de células somáticas é influenciada por diversos fatores como gravidade da infecção, tipo de microrganismo causador, duração, idade do animal, época do ano, estado nutricional potencial genético, e lesões residuais de infecções anteriores (Schult, 1977).

Quanto maior a quantidade de CCS encontrada no leite menor é o valor de mercado deste produto, chegando até a perda total da produção. Por isso é crescente as exigências das indústrias por uma matéria-prima de melhor qualidade e preocupação dos consumidores pela aquisição de produtos seguros. Contudo, há um risco permanente desse produto veicular microrganismos patogênicos ou sofrer fraudes durante o processamento. Em ambas as circunstâncias, o produto passa a ser prejudicial para a saúde do consumidor. Por isso, a

qualidade do leite constitui um critério importante no processamento de leite e derivados (Evangelista, 2008).

Nas indústrias de laticínios podemos observar como principais práticas, a diferenciação no pagamento ao produtor e o aumento nas exigências de qualidade do leite. Parâmetros físico-químicos, microbiológicos e higiênico-sanitários são utilizados pelas indústrias para verificação da qualidade do leite, como por exemplo, a contagem de células somáticas, a contagem bacteriana e a pesquisa de resíduos de antimicrobianos, que estão sendo cada vez mais exigidos como parâmetros de qualidade (Guerreiro *et al*., 2005).

A escolha da família de distribuições para Dados de Contagem com Excesso de Zeros se deve a uma característica da variável de interesse, o excesso de zeros, pois o valor zero em casos de contaminação é uma observação muito importante, pois pode representar o melhor animal em termos genéticos e por isso não podemos retirar ou perder essas observações. Porém existem algumas restrições estatísticas quanto a este tipo de dados, como por exemplo, não podemos usar a transformação logarítmica ou a de Box-Cox.

O objetivo desse trabalho é identificar o comportamento da variável CCS, encontrando uma distribuição de probabilidade que se ajuste bem aos dados, utilizando inicialmente uma família de distribuições para dados inflacionados de zeros, e observar seu comportamento em relação às covariáveis idade e dias de lactação. Para isso, o pacote GAMLSS do software estatístico R (R Development Core Team, 2011) será utilizado em todas as análises estatísticas.

Capítulo 2

Metodologia

Através do banco de dados da EMBRAPA Gado de Leite de animais da raça GIR que pertenceram a um único criador. Destes, somente a 1ª lactação de cada animal foi considerada. Os animais tiveram pelo menos 3 controles leiteiros durante a lactação, e entre os que tiveram lactação duradoura (≥ 200 dias) pelo menos 2 controles foram feitos acima do 200º dia da lactação. Temos ao final, 9763 observações (controles leiteiros) coletadas entre os anos de 1996 e 2011, em 1785 animais, de 200 criadores diferentes, durante um período de 5 até 305 dias de lactação (dias após o parto), e em vacas com idades entre 1,8 e 5,5 anos. As datas de coletas não foram definidas previamente, por isso não temos intervalos igualmente espaçados de coleta, porém a maioria das amostragens é feita em média a cada 30 dias durante o período de lactação da vaca, que é de cerca de 300 dias. A variável de interesse é a quantidade de CCS presente na amostra do leite, cujo comportamento pode ser observado na Figura 1.

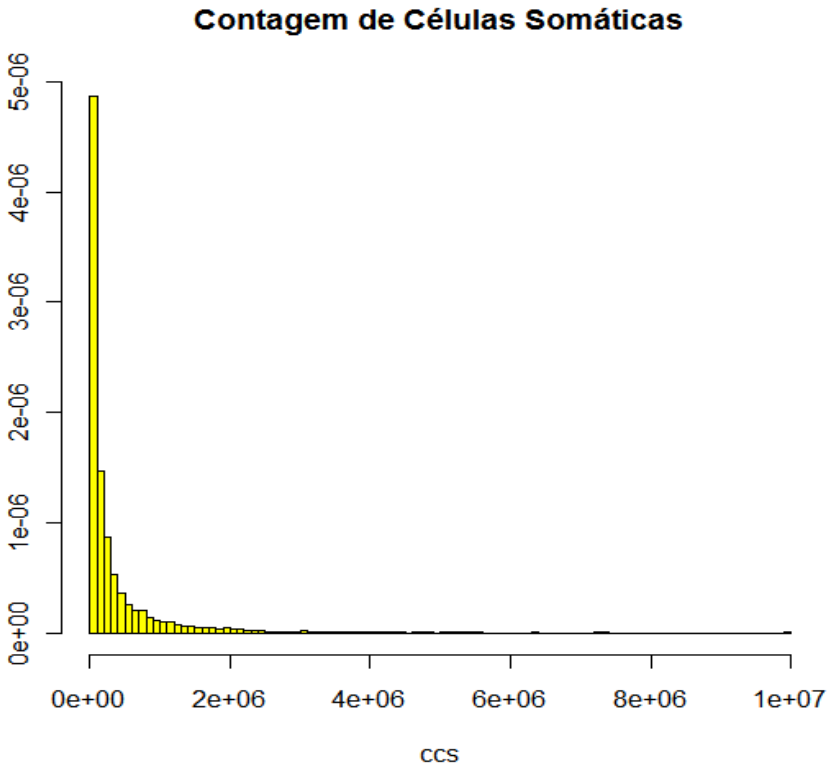


Figura 1: Histograma da variável CCS

A contagem da CCS é realizada de forma *óptico-eletrônica*, por citometria de fluxo, num aparelho capaz de detectar contagens entre 1.000 e 10.000.000 cél.s/mL., criando assim um limite inferior e superior da contagem. Essa limitação do aparelho cria uma dificuldade na modelagem, pois os zeros encontrados não são zeros reais, são valores que estão abaixo do nível inferior e que são desconsiderados, dada a capacidade de mensuração do aparelho. Que também pode criar alguns erros de mensuração para valores baixos da contagem como em 1000, 2000,..., 15000. Como podemos ver na figura A1 em anexo.

A microscopia direta é o método de referência para a determinação da contagem de células somáticas em leite cru. Uma alíquota de leite (0,01 mL) é distribuída homogeneamente em uma área delimitada (1 cm²) na superfície de uma lâmina com o auxílio de uma pipeta automática calibrada. Após a secagem, as lâminas devem ser coradas com uma solução de azul de metileno 0,6% (corante de Newman-Lampert modificado por Lebowitz-Weber) e, em seguida, as células coradas são contadas por meio da observação em um microscópio óptico (Marshall, 1992). O número de células contadas na área delimitada é multiplicado pelo fator de trabalho do microscópio e expresso em número de células por mililitro (International..., 1991c).

Figura 2: Equipamento eletrônico Somacount 300 da Bentley Instruments Incorporated® para contagem de células somáticas em amostras de leite.

O leite é uma emulsão estável de glóbulo de gordura e uma suspensão coloidal de micelas de caseína. A lactose, as proteínas do soro, a maior parte dos minerais e vitaminas hidrossolúveis encontram-se dissolvidos na água formando uma solução (Monardes, 1998).

A concentração de CCS acima de 280.000 cel/ml já indica a ocorrência de mastite subclínica, que já influencia na qualidade do leite e não pode ser diagnosticada visualmente.

Países como a União Européia, Nova Zelândia e Austrália adotam como limite máximo legal para a CCS do leite para o consumo humano o valor de 400.000 cél.s/mL enquanto o Canadá fixou esse limite em 500.000 cél.s/mL e os Estados Unidos o valor de 750.000 cél.s/mL. Recentemente a legislação brasileira sobre a produção de leite foi alterada pelo Ministério da Agricultura e Pecuária e adotado como limite máximo legal para a CCS do leite para o consumo humano o valor de 400.000 cél.s/ml.. O antes denominado "Programa Nacional de Melhoria da Qualidade do Leite", que tinha como objetivo implementar várias mudanças na legislação brasileira no que se refere à qualidade do leite, passou por consulta pública e após algumas alterações se transformou na Instrução Normativa nº 51. As principais mudanças que esta nova Instrução Normativa trará são: a adoção de parâmetros de qualidade como a contagem de padrão em placas, a contagem de células somáticas, a ausência de resíduos de antibióticos, entre outros. A adaptação dos produtores a esta nova lei será feita de forma gradual a atingir os níveis finais de requerimento em um prazo de 7 anos após a entrada em vigor desta legislação (Santos, 2004).

2.1. Relação CCS x Idade

A idade da vaca pode ser um fator de influência na quantidade de CCS observada.

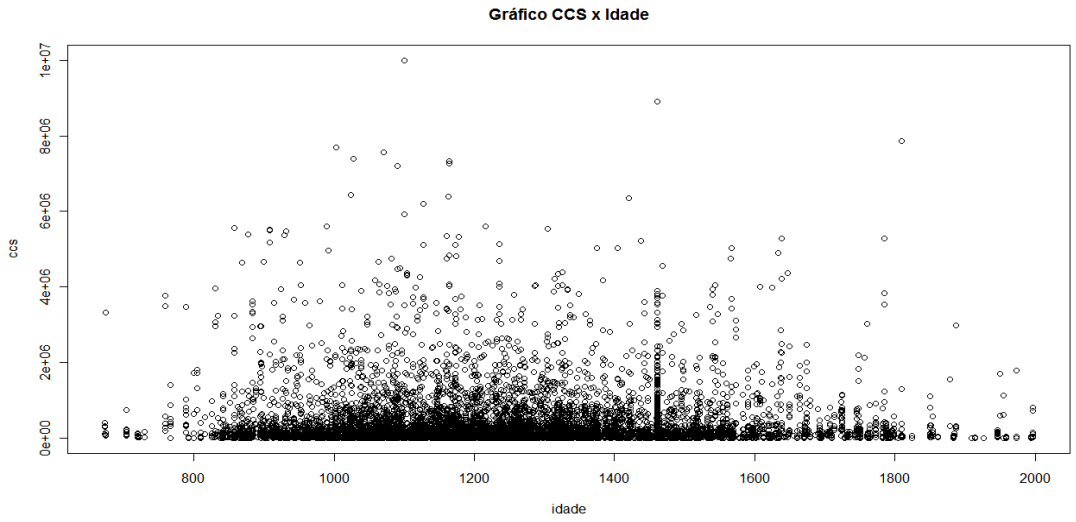


Figura 3:Valor absoluto de CCS x Idade

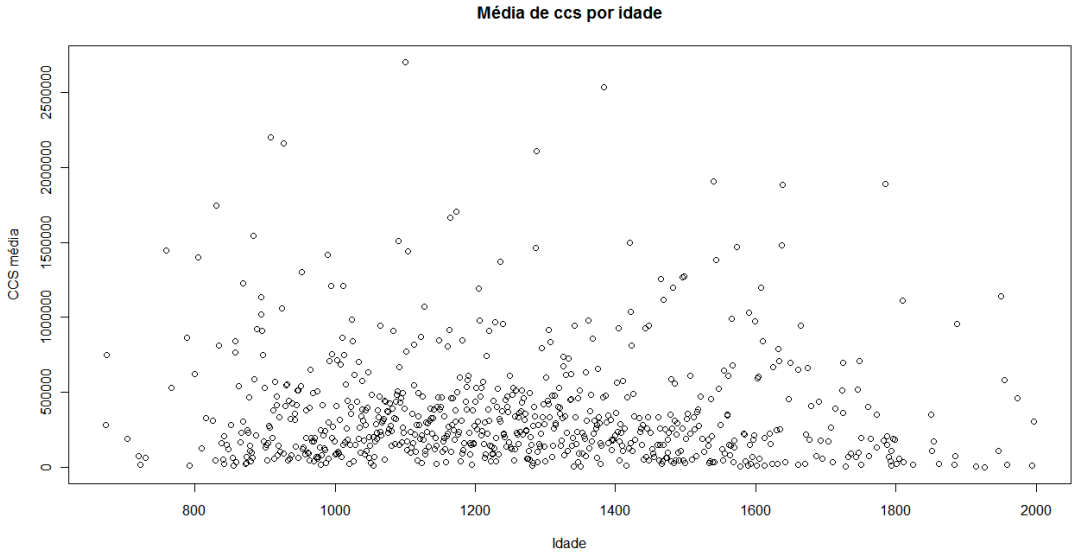


Figura 4: Média de CCS para uma idade i=673, 677, ...,1997

Fazendo um modelo de regressão linear simples para as médias de CCS por idade, temos um modelo $MédiaCCS = 460353{,}84 - 80 * idade(em\ dias)$ que possui uma relação linear negativa, o valor de CCS diminui ao longo da idade. Mas o modelo possui um coeficiente de correlação $R^2 = 0{,}34\%$, que é muito baixo. Ou seja, aproximadamente 0,34% da diminuição de CCS estão relacionadas com o aumento da idade.

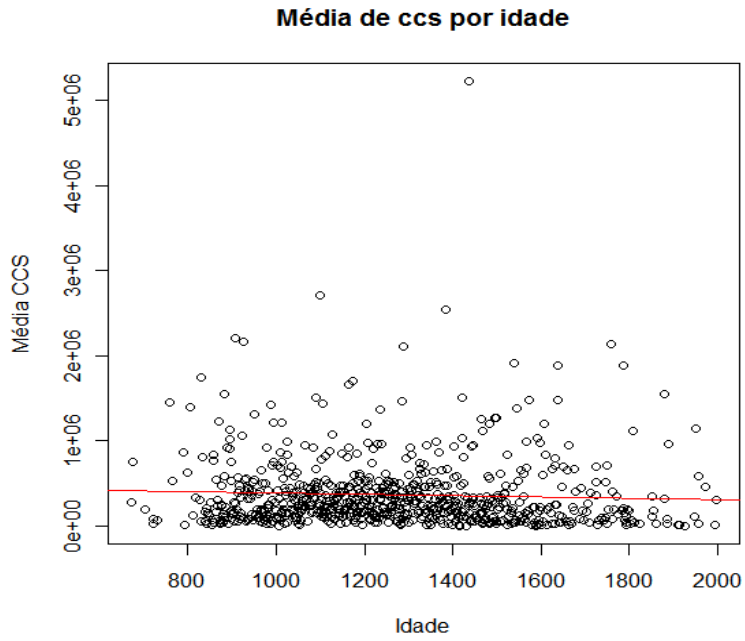


Figura 5: Gráfico da média por idade, com a reta da regressão linear simples.

Pela forte assimetria observada a média pode não ser uma boa medida de tendência central, talvez seja melhor observar um gráfico com as medianas dos dias.

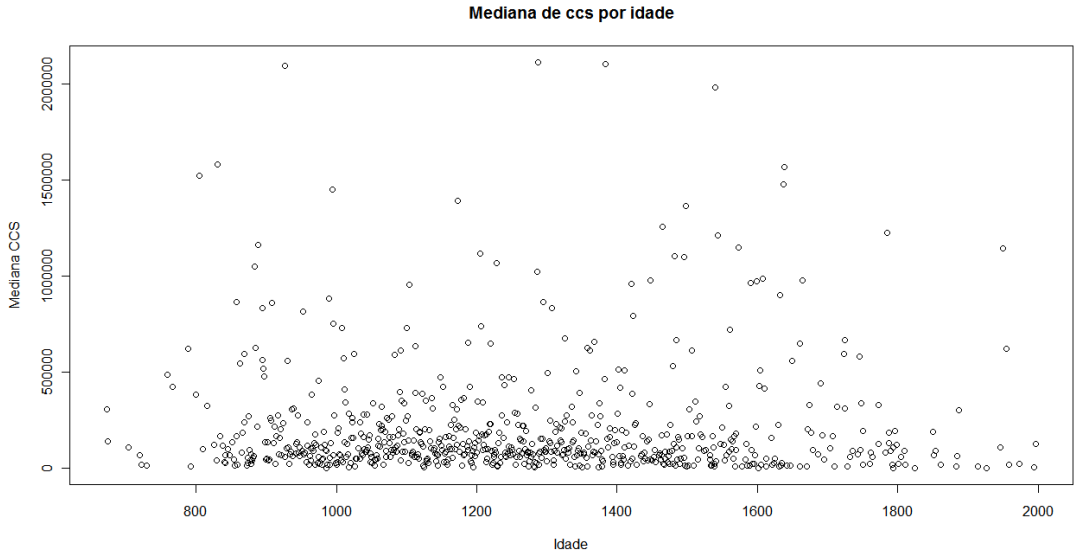


Figura 6: Mediana de CCS para uma idade i=673, 677, ...,1997

Fazendo um modelo de regressão linear simples para as medianas de CCS por idade, temos um modelo $MedianaCCS = 195400 + 9{,}512 * idade(em\ dias)$ que possui uma relação linear positiva, o valor de CCS aumenta ao longo da idade. Mas o modelo possui um coeficiente de correlação $R^2 = 0{,}007737\%$, que é muito baixo. Ou seja, aproximadamente 0,007737% do aumento de CCS estão relacionadas com o aumento da idade.

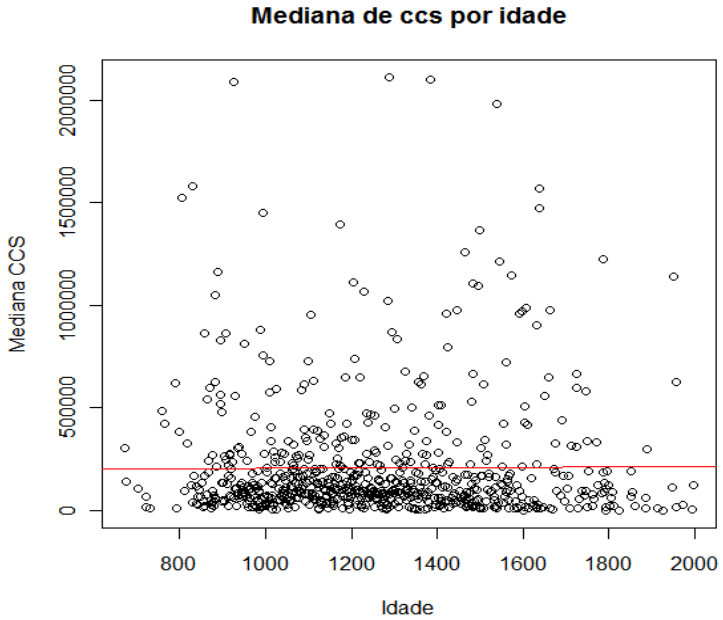


Figura 7: Gráfico da mediana por idade, com a reta da regressão linear simples.

Observando os gráficos podemos perceber uma que há uma leve correlação, mas nada muito expressivo, o que possivelmente é explicado pelos dados já que estamos trabalhando com animais jovens e de 1º lactação.

2.2. Relação CCS x Dias de lactação

Os dias de lactação pode ser um fator de muita importância na quantidade de CCS.

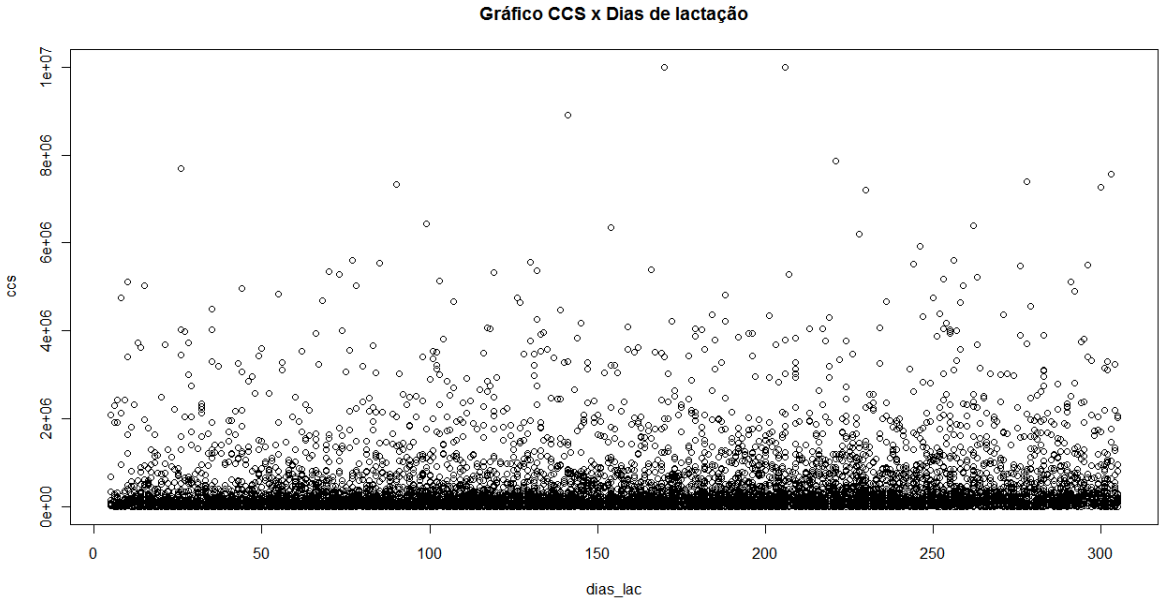


Figura 8:Valor absoluto de CCS x Dias de lactação

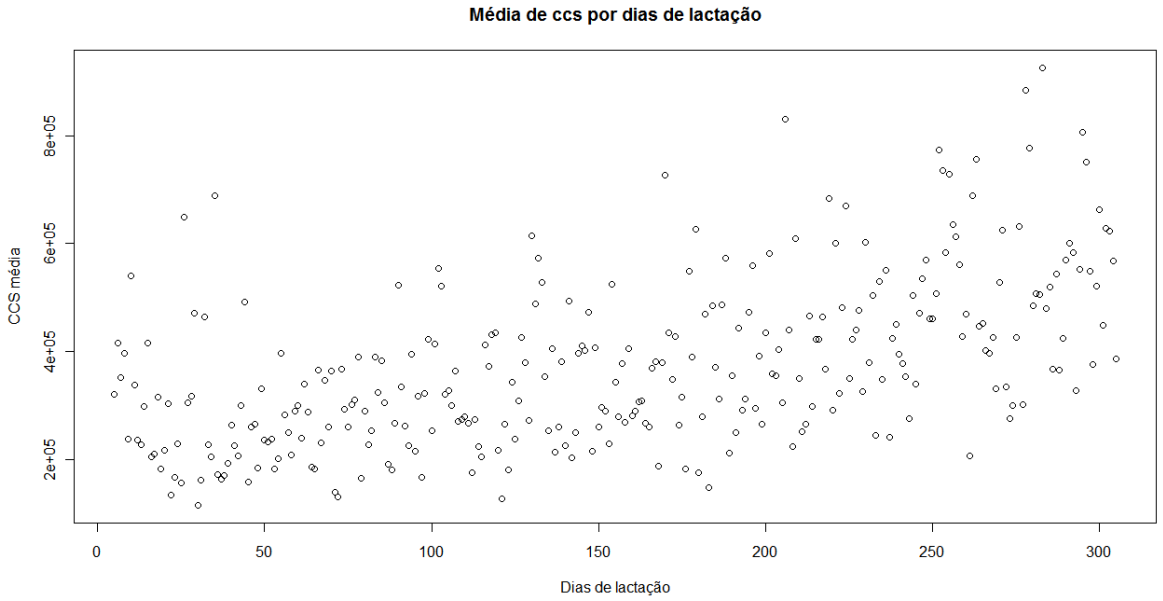


Figura 9: Média de CCS por dia de lactação i=5,6...,305

O gráfico da figura 8 com os valores absolutos de CCS não parece haver uma relação muito expressiva entre as covariáveis. Mas no gráfico da figura 9 podemos ver uma relação positiva entre a CCS e a covariável dia de lactação, ou seja, quanto maior o tempo de lactação maior são as chances de se observar alta quantidade de CCS.

Fazendo um modelo de regressão linear simples para as médias de CCS por dias de lactação, temos um modelo $MédiaCCS = 274200 + 3{,}174 * dias^2$ que possui uma relação linear positiva, o valor de CCS aumenta ao longo dos dias. O modelo possui um coeficiente de correlação $R^2 = 33{,}05\%$ que não é muito alto, mas já é muito melhor que o anterior. Ou seja, aproximadamente 33,05% do aumento de CCS estão relacionadas com o aumento de dias de lactação.

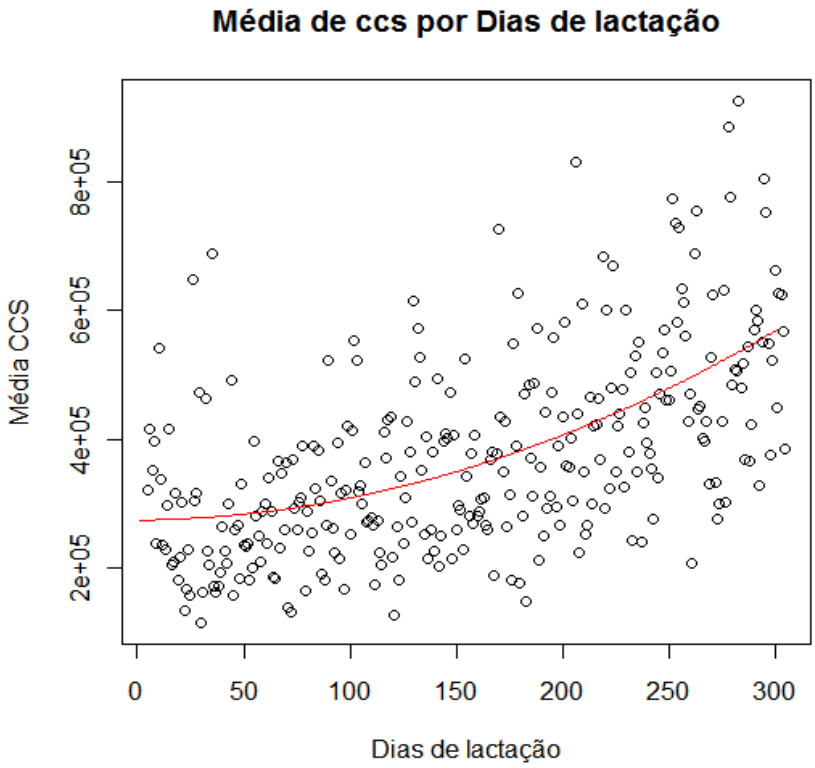


Figura 10: Gráfico da média por dias de lactação, com a reta da regressão linear simples.

Como a nossa variável de interesse apresenta forte assimetria, talvez a média não seja uma boa escolha como medida de tendência central.

Por isso talvez seja melhor observar um gráfico com a mediana dos dias. No gráfico da figura 11 abaixo, a relação entre as variáveis continua sendo positiva.

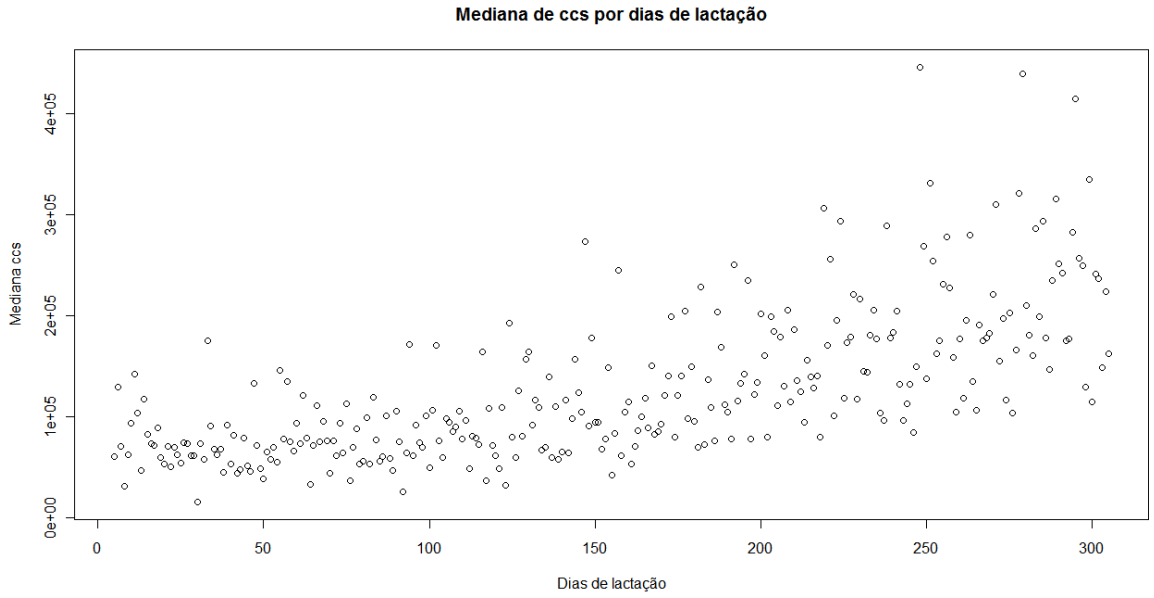


Figura 11: Mediana de CCS por dia de lactação i=5,6...,305

Fazendo um modelo de regressão linear simples para as medianas de CCS por dias de lactação, temos um modelo $MedianaCCS = 68740 + 1{,}872 * dias^2$ que possui uma relação linear positiva, o valor de CCS aumenta ao longo dos dias. O modelo possui um coeficiente de correlação $R^2 = 49{,}75\%$ que não é muito alto, mas já é muito melhor que o anterior. Ou seja, aproximadamente 49,75% do aumento de CCS estão relacionadas com o aumento de dias de lactação.

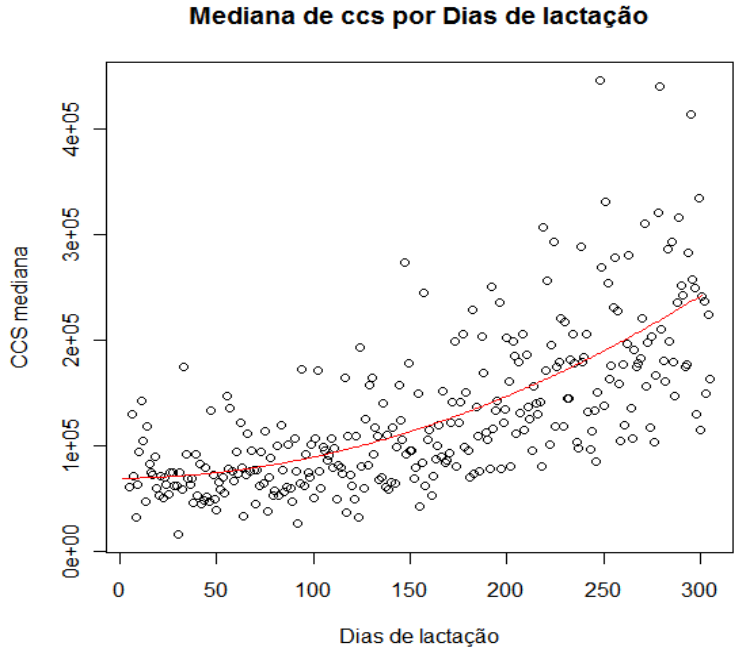


Figura 12: Gráfico da mediana por dias de lactação, com a reta da regressão linear simples.

O ajuste dos dados a uma regressão linear de CCS com idade não apresentou um bom ajuste para a média e para mediana. Já para a idade vemos um ajuste bem melhor. O que pode ser comprovado pelos gráficos de envelope abaixo.

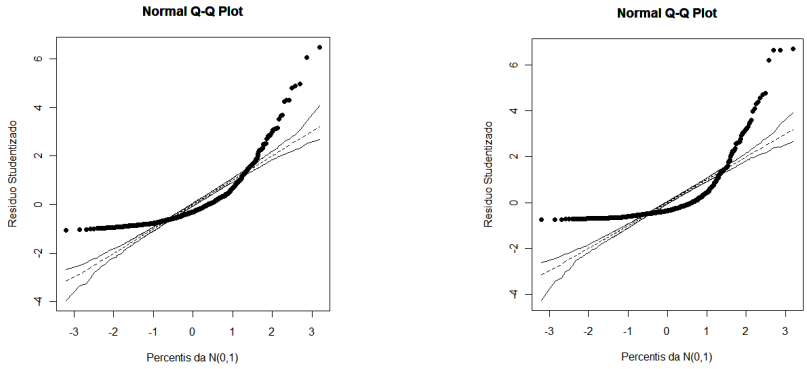


Figura 13: GráficoQ-Q Plot do ajuste de CCS com Idade para a média e para mediana, respectivamente.

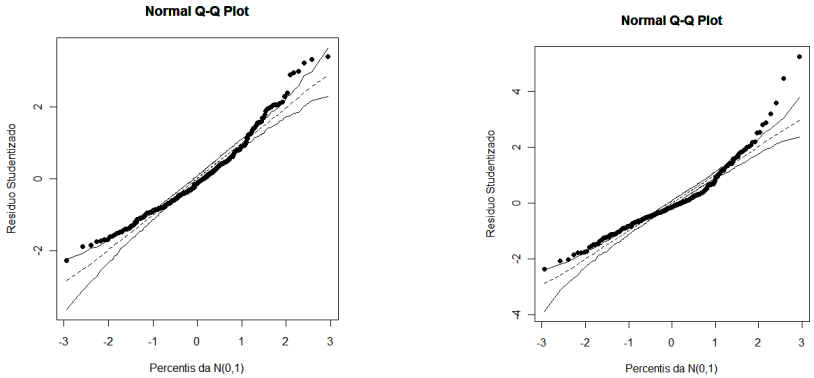


Figura 14: GráficoQ-Q Plot do ajuste de CCS com Dias de lactação para a média e para mediana, respectivamente.

Com o objetivo de tentar um melhor ajuste para os dados vamos tentar ajustes da família de distribuições inflacionadas de zeros.

Capítulo 3

Modelos para Dados de Contagem com excesso de Zeros

A família de distribuições para dados inflacionados de zeros é na verdade é uma combinação de distribuições já conhecidas. Usualmente estas distribuições são utilizadas para modelar dados resultantes do processamento de fabricação, de economia, entre outras aplicações.

O valor zero em casos de contaminação é uma observação muito importante, pois pode representar o melhor animal em termos genéticos e por isso não podemos retirar ou perder essas observações. A utilização de transformações na variável reposta é frequentemente utilizada para alcançar a normalidade, mas não é ideal a utilização em variáveis onde existam valores iguais à zero, pois além dos problemas usuais como dificuldade na interpretação dos resultados e viés de estimação, existem algumas restrições estatísticas quanto a este tipo de dado, como por exemplo, não podemos usar a transformação logarítmica ou a de Box-Cox.

Então uma solução possível é utilizar uma família de distribuições para dados inflacionados de zeros. Dada a natureza dos dados, vamos utilizar a distribuição Poisson Inflada de Zeros e Binomial Negativa Inflada de Zeros.

3.1 – Modelo Poisson Inflado de Zeros – ZIP

O modelo ZIP é um dos mais utilizado entre os modelos para dados de contagem, ele é utilizado quando observamos em uma distribuição discreta de Poisson com maior quantidade de observações iguais a zero que o modelo permite.

3.1.1 – Distribuição de Poisson

Definição 1: Uma variável aleatória Y segue uma distribuição Poisson se sua função de probabilidade (fp) é dada por

$$f(y|\lambda) = \frac{e^{-\lambda}\,\lambda^{y}}{y!} \quad , \quad y = 0,1,2 \ldots\ldots , \quad 0 \leq \lambda < \infty.$$

Notação: Y ~ Poisson(λ).

Com a $fgm\ M_x(t) = e^{\lambda(e^t-1)}$

Logo, temos que

$$E(Y) = \lambda \qquad \text{e} \qquad Var(Y) = \lambda$$

3.1.2 – Distribuição de Poisson Inflada de Zeros

Os dados observados estão distribuídos de forma discreta no conjunto dos inteiros positivos($\mathbb{Z}$), mas que inclui muitas observações no extremo igual a zero. Vamos assumir que a distribuição dos dados é uma mistura entre a distribuição Poisson discreta definida no intervalo [0,+∞) e a distribuição Bernoulli, a qual atribui probabilidades não negativas aos inteiros 0 e 1.

Como, segundo o modelo proposto, os dados são observados no conjunto dos inteiros positivos($\mathbb{Z}$), teremos então que a distribuição de probabilidade é uma mistura entre uma distribuição discreta e uma distribuição degenerada em 0. O modelo proposto faz parte da classe dos modelos inflacionados, onde a massa de probabilidades dos pontos iguais a zero excede o que é permitido pelo modelo Poisson.

Definição 2: Uma variável aleatória segue uma distribuição Poisson inflada de zeros se sua fp for dada por

$$P(Y = y) = \begin{cases} p + (1-p)e^{-\lambda} & , \quad y = 0 \\ (1-p)\dfrac{e^{-\lambda}\lambda^{y}}{y!} & , \quad y = 1, 2, 3, \ldots \end{cases}$$

onde $0 \leq p < 1$, $\lambda > 0$. O parâmetro *p* pode ser interpretado como a proporção de zeros e λ como a taxa média de ocorrência de eventos em uma unidade de tempo, também conhecido como parâmetro de intensidade.

Notação: Y ~ ZIP(λ,p)

O valor esperado e a variância da distribuição Poisson Inflada de Zeros são dados por:

$$E(Y) = (1-p)\lambda \quad \text{e} \quad Var(Y) = \lambda(1-p)(1-p\lambda)$$

Exemplo 1: Função de probabilidade gerada de uma ZIP

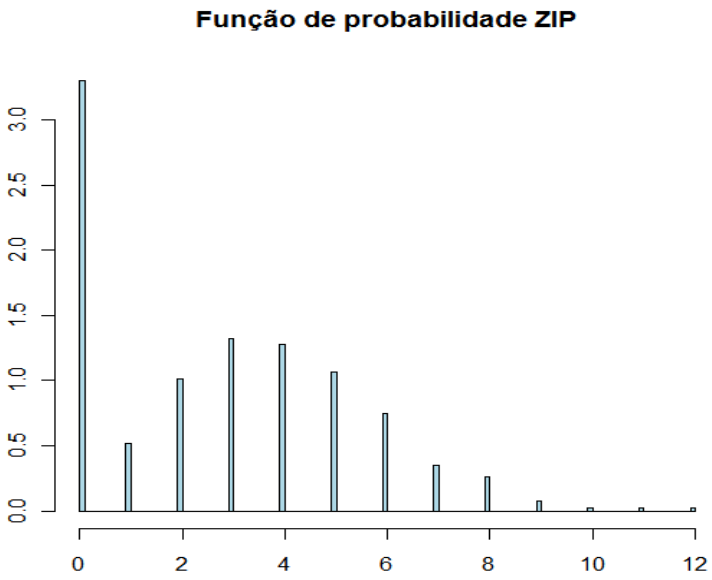


Figura 15: Função de probabilidade ZIP com média 4 e proporção de zeros 0,3

3.2 – Modelo Binomial Negativo Inflado de Zeros – ZINB

O modelo ZINB é utilizado quando observamos em uma distribuição discreta de uma Binomial Negativa com maior quantidade de observações iguais a zero que o modelo permite e com superdispersão, como no estudo feito por Yau *et al.* (2003).

3.2.1 – Distribuição Binomial Negativa

Definição 3: Uma variável aleatória Y segue uma distribuição Binomial Negativa se sua função de probabilidade (fp) é dada por

$$f(y|\mu;\sigma) = \frac{\Gamma\left(\mathrm{y}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(\mathrm{y}+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^{y}\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}, \quad y = 0,1,2\ldots, \qquad \sigma > 0, \qquad \mu > 0,$$

onde, μ é o parâmetro de média e σ é o parâmetro de dispersão (Evans, 1953).

Notação: Y ~ BN(μ,σ)

O valor esperado e a variância da distribuição Binomial Negativa são dados por:

$$E(Y) = \mu \qquad \text{e} \qquad Var(Y) = (1+\sigma)\mu$$

3.2.2 – Distribuição Binomial Negativa Inflada de Zeros

Os dados observados estão distribuídos de forma discreta no conjunto dos inteiros positivos ($\mathbb{Z}^+$), mas que inclui muitas observações no extremo igual a zero. Vamos assumir que a distribuição dos dados é uma mistura entre a distribuição Binomial Negativa discreta definida em $\mathbb{Z}_*^+$ e a distribuição Bernoulli, a qual atribui probabilidades não negativas aos inteiros 0 e 1.

Como, segundo o modelo proposto, os dados são observados no conjunto dos inteiros não negativos ($\mathbb{Z}_*^+$), teremos então que a distribuição de probabilidade é uma mistura entre uma distribuição discreta e uma distribuição degenerada em 0. O modelo proposto faz parte da classe dos modelos inflacionados, onde a massa de probabilidades dos pontos iguais a zero excede o que é permitido pelo modelo Binomial Negativo.

Definição 4: Uma variável aleatória segue uma distribuição Binomial Negativa inflada de zeros se sua fp for dada por:

$$P(Y=y) = \begin{cases} p + (1-p)\left(\dfrac{1}{1+\sigma\mu}\right)^{1/\sigma}, & y = 0 \\ (1-p)\dfrac{\Gamma\left(\mathrm{y}+\dfrac{1}{\sigma}\right)}{\Gamma\left(\dfrac{1}{\sigma}\right)\Gamma(\mathrm{y}+1)}\left(\dfrac{\sigma\mu}{1+\sigma\mu}\right)^{y}\left(\dfrac{1}{1+\sigma\mu}\right)^{1/\sigma}, & y = 1, 2, \ldots \end{cases}$$

onde $0<p<1$ é a proporção de zeros, $\mu > 0$ é o parâmetro de média, e $\sigma > 0$ é o parâmetro de dispersão.

Notação: $Y \sim ZI\mathrm{BN}(\mu, \sigma, p)$

O valor esperado e a variância da distribuição Binomial Negativa Inflada de Zeros são dados por:

$$E(Y) = (1-p)\,\mu \quad \text{e} \quad Var(Y) = (1-p)(1+\mu\sigma+p\mu)\mu$$

Exemplo 2: Função de probabilidade gerada de uma ZINB

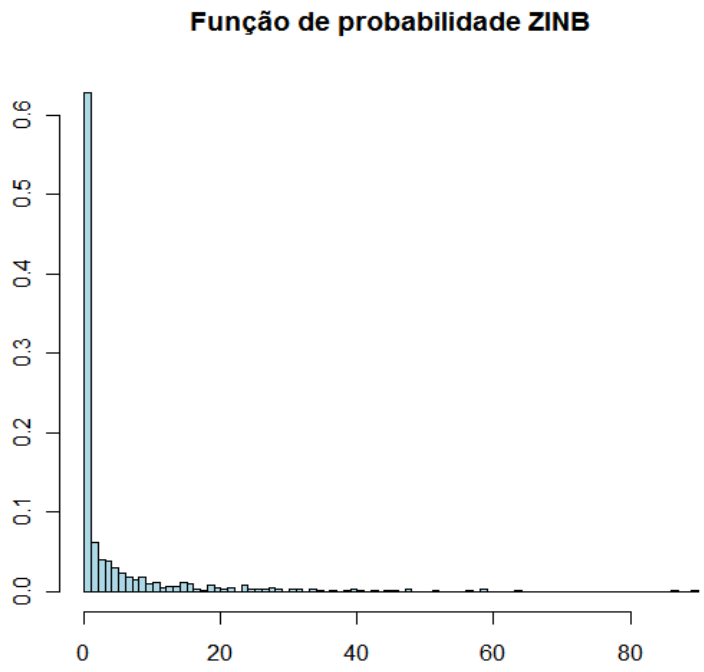


Figura 16: Função de probabilidade ZINB com média 5, dispersão 4 e proporção de zeros 0.1

3.3 – GAMLSS

Os Modelos aditivos generalizados para locação, escala e forma (Generalized Additive Models for Location, Scale and Shape – GAMLSS) foram introduzidos por Rigby e Stasinopoulos (2001, 2005) e Akantziliotou et al. (2002) como uma forma de superar algumas das limitações associadas aos Modelos Lineares Generalizados (GLM) e aos Modelos Aditivos Generalizados (GAM).

Em GAMLSS, a suposição de distribuição da família exponencial para a variável resposta (Y) é relaxada e substituída por uma família geral de distribuições, incluindo aquelas com forte assimetria e/ou curtose. A parte sistemática do modelo é expandida para permitir a modelagem, não apenas da média (ou locação), mas outros parâmetros da distribuição de Y como linear paramétrico e/ou aditivos não-paramétricos de funções de variáveis explicativas e/ou efeitos aleatórios. A estimação da máxima verossimilhança (penalizada) é usada para ajustar os modelos.

Há dois algoritmos para ajustar os modelos, os algoritmos de CG e RS, que são discutidas em detalhe em Rigby e Stasinopoulos (2005).

Capítulo 4

Resultados

Vamos utilizar os modelos citados anteriormente em dados reais de contagem de células somáticas em proporção de milhares, ou seja, dividido por 1000.

Para fazermos uma comparação com modelos inflados, vamos ajustar a CCS com uma distribuição de Poisson (λ)

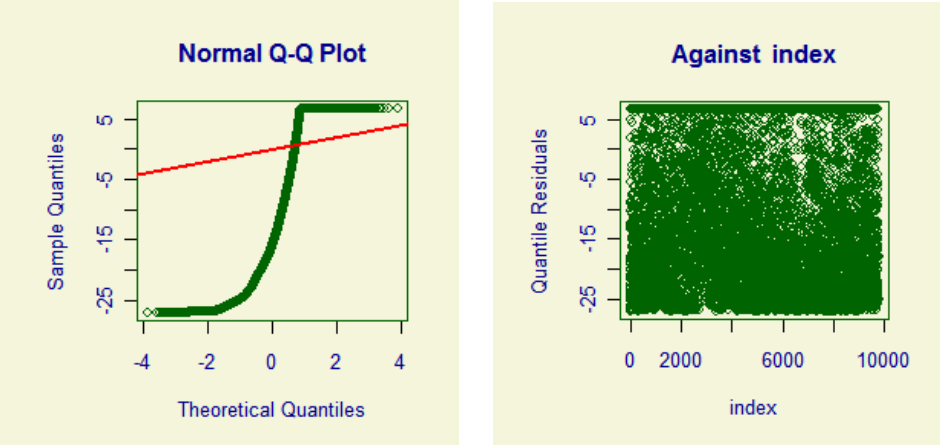


Figura 17: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma Poisson

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 7428078 é apresentado na tabela abaixo:

Parâmetro	Estimativa
λ	366,8672

Pela analise dos gráficos de resíduos vemos que o modelo não está bem ajustado aos dados.

Para fazermos uma comparação com modelos inflados, vamos ajustar a CCS com uma distribuição BN(μ,σ)

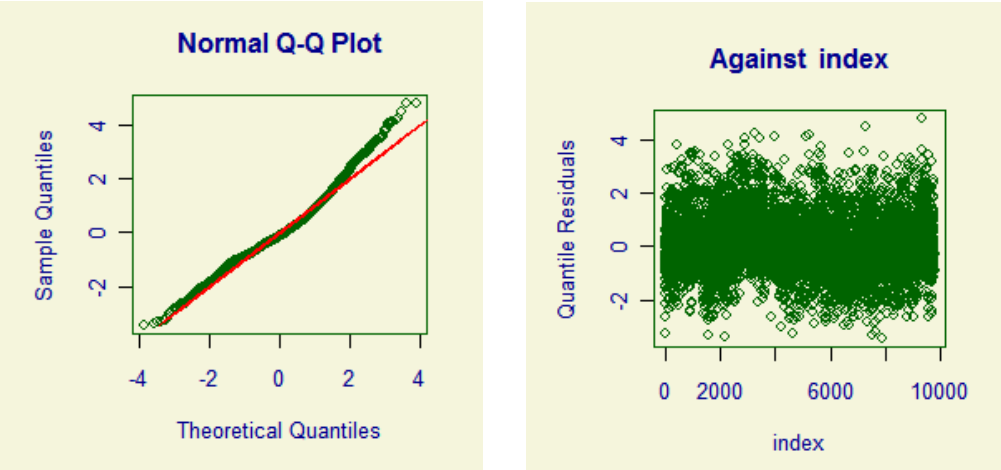


Figura 18: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma BN sem covariáveis

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 128921,1 é apresentado na tabela abaixo:

Parâmetro	Estimativa
μ	367,0213
σ	2,297542

Pelo gráfico de resíduos vemos uma melhora substancial comparando com o modelo anterior de Poisson.

4.1 – Aplicação a um modelo de Regressão de Poisson com Excesso de Zeros

Considere n observações (Y_1, Y_2, ..., Y_n) de uma variável aleatória Y~ZIP(λ, p).

Associado a i-ésima observação, suponha que tenhamos observações de k variáveis explicativas x_{i1} , $x_{i2}, \dots x_{ik}$. Assim, temos que $\mu_i = E(Y_i) = (1-p)\lambda_i$. Utilizamos uma função de ligação para modelar a proporção de zeros e outra para modelar o parâmetro de locação.

Segundo Montoya (2009), as funções de ligação utilizadas são $log(\lambda_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ (logarítmica) e $log\left(\frac{p}{1-p}\right) = \pi$ (logito).

A nossa variável resposta será a CCS; idade e dias de lactação serão as variáveis explicativas.

Modelando a CCS com uma distribuição de ZIP(λ, p)

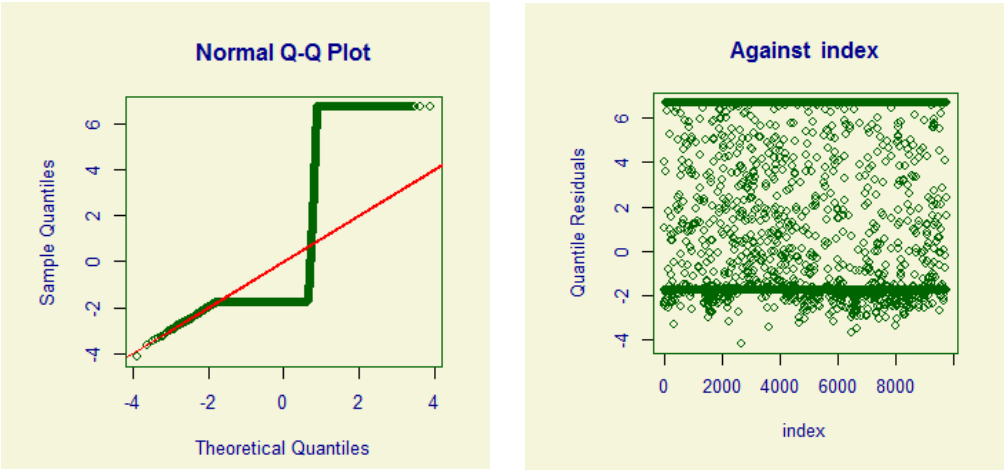


Figura 19: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZIP sem covariáveis

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 7136171 é apresentado na tabela abaixo:

Parâmetro	Estimativa
p	0,04034912
λ	382,6038

4.1.1 – CCS por Idade

Modelo: $log(\lambda_i) = \beta_0 + \beta_1 X_1$

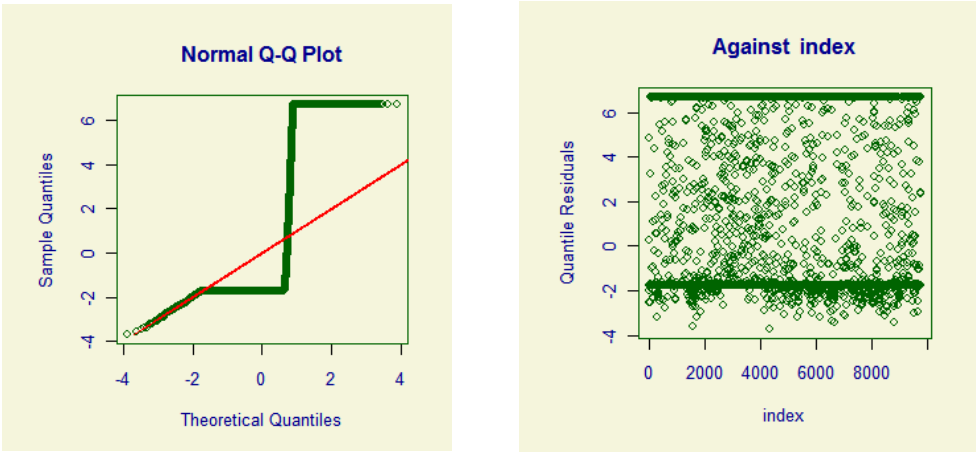


Figura 20: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZIP com a explicativa idade

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 7130415 é apresentado na tabela abaixo:

Parâmetro	Estimativa
p	0,04034912
β_0	6,1750274
β_1	-0,0001838

4.1.2 – CCS por Dias de lactação

Modelo: $log(\lambda_i) = \beta_0 + \beta_2 X_2$

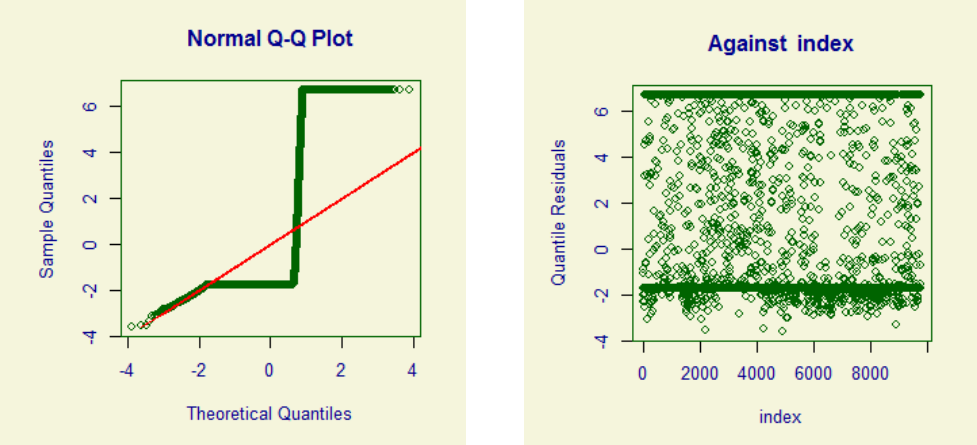


Figura 21: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZIP com a explicativa, dias de lactação

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC=6973640 é apresentado na tabela abaixo:

Parâmetro	Estimativa
p	0,04034912
β_0	5,525106
β_2	0,002626

4.1.3 – CCS por Idade e dias de lactação

Modelo: $log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

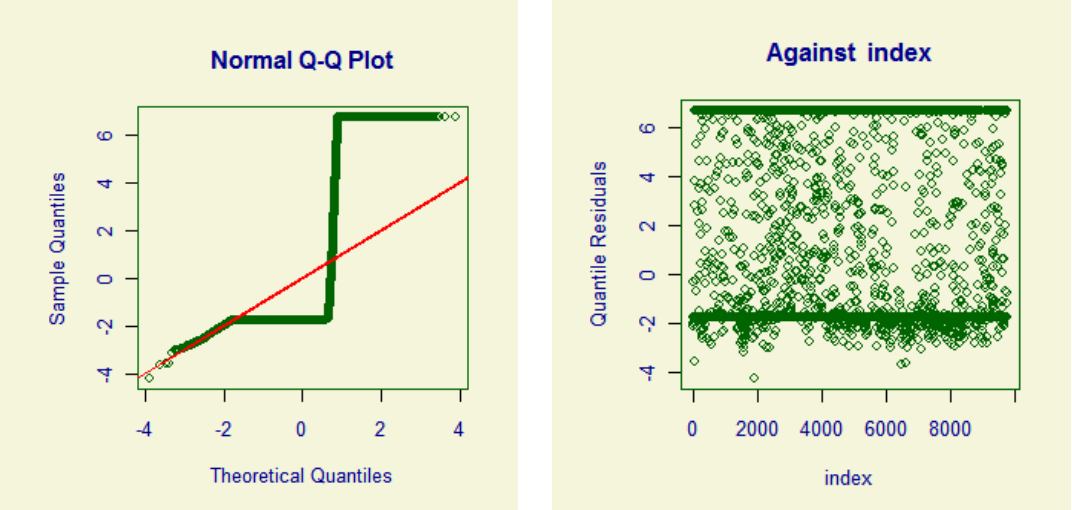


Figura 22: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZIP com as explicativas, idade e dias de lactação

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC=6967612 é apresentado na tabela abaixo:

Parâmetro	Estimativa
p	0,04204562
β_0	5,758
β_1	-0,00019
β_2	0,00263

Comparando o AIC dos modelos ajustados com a ZIP, o melhor modelo é o que utiliza as duas variáveis explicativas.

Modelo ZIP	AIC
$y_i = \beta_0$ *	7428078
$y_i = \beta_0$	7136171
$y_i = \beta_0 + \beta_1 X_1$	7130415
$y_i = \beta_0 + \beta_2 X_2$	6973640
$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	6967612

*Modelo Poisson

4.2 – Aplicação de um modelo de Regressão Binomial Negativa com Excesso de Zeros

Considere n observações (Y_1, Y_2, ..., Y_n) de uma variável aleatória Y~ZINB(μ, σ, p).

Associado a i-ésima observação, suponha que tenhamos observações de k variáveis explicativas x_{i1} , $x_{i2}, \dots x_{ik}$. Assim, temos que $E(Y_i) = (1-p)\mu_i$. Utilizamos uma função de ligação para modelar a proporção de zeros e outra para modelar os parâmetros de locação e escala.

Segundo Montoya (2009), as funções de ligação utilizadas são $log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ (logarítmica), $log(\sigma) = \gamma$ (logarítmica) e $log\left(\frac{p}{1-p}\right) = \pi$ (logito).

A nossa variável resposta será a CCS; idade e dias de lactação são as variáveis explicativas.

Ajustando a CCS com uma distribuição ZINB(μ,σ,p)

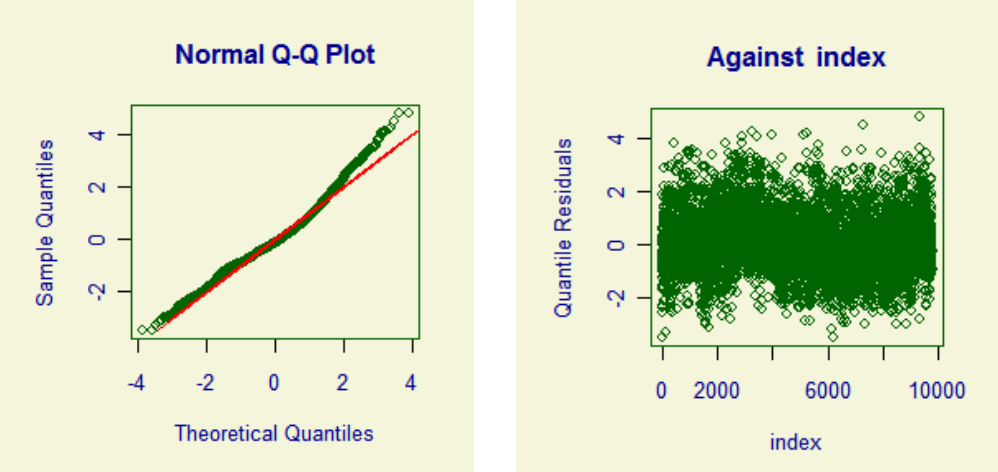


Figura 23: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZINB sem covariáveis

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 128923,1 é apresentado na tabela abaixo:

Parâmetros	Estimativa
μ	366,8672
σ	2,297450
p	0

4.2.1 – CCS por Idade

Modelo: $log(\mu_i) = \beta_0 + \beta_1 X_1$

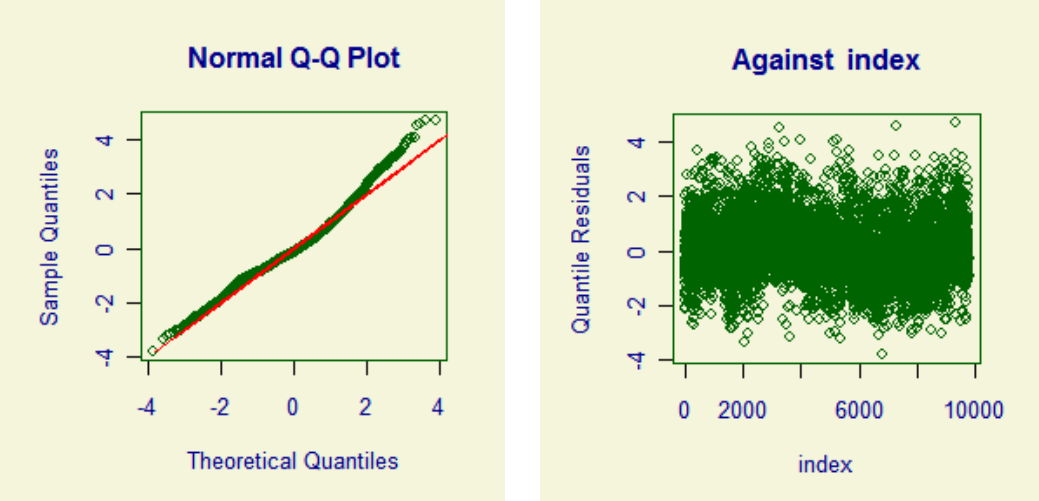


Figura 24: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZINB com a covariável idade

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 128915,3 é apresentado na tabela abaixo:

Parâmetros	Estimativa
σ	2,295843
p	0
β_0	6,1737775
β_1	-0,0002161

4.2.2 – CCS por Dias de lactação

Modelo: $log(\mu_i) = \beta_0 + \beta_2 X_2$

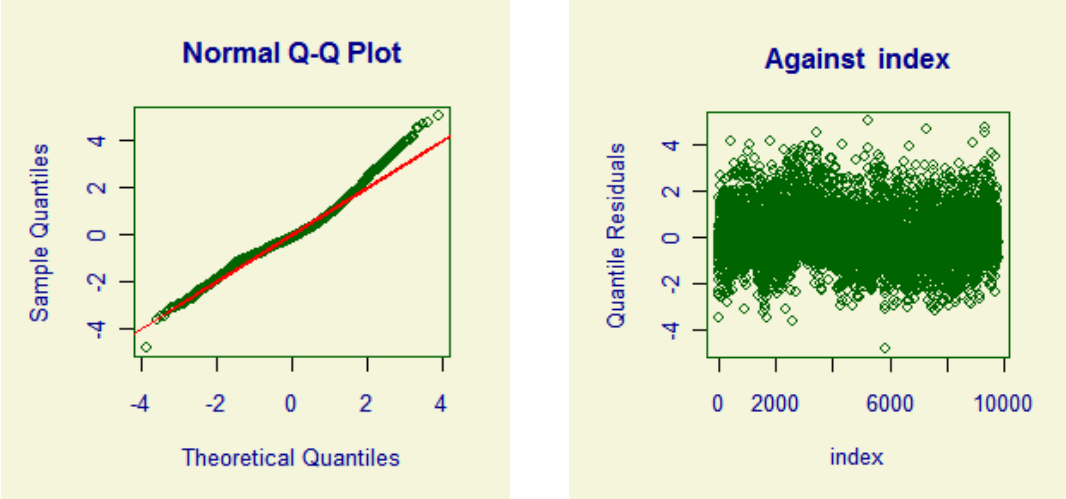


Figura 25: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZINB com a covariável dias de lactação

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 128723.7 é apresentado na tabela abaixo:

Parâmetros	Estimativa
σ	2,263699
p	0
β_0	5,484548
β_2	0,002624

4.2.3 – Idade e dias de lactação

Modelo: $log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

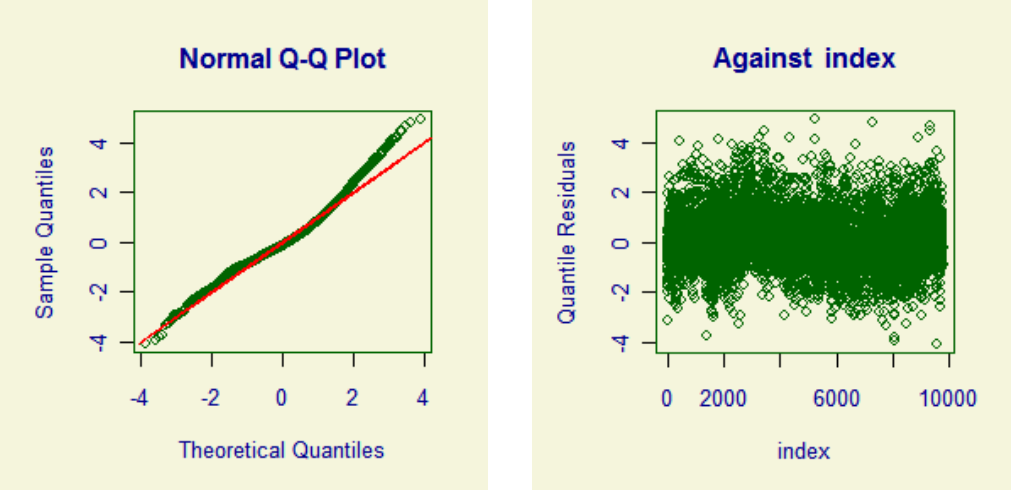


Figura 26: Q-Q plot e gráfico dos resíduos do ajuste dos dados a uma ZINB com as covariáveis, idade e dias de lactação

O resultado apresentado pelo gamlss como o melhor ajuste com um AIC= 128716.8 é apresentado na tabela abaixo:

Parâmetros	Estimativa
σ	2,262341
p	0
β_0	5,7395867
β_1	-0,0002049
β_2	0,0026194

Nesse ajuste da ZINB a proporção de zeros para os modelos é nula, isto significa que a variável CCS não excede a quantidade de zeros suportada pelo modelo NB. De acordo com o AIC o melhor modelo, o que possui menor AIC, é o ZINB com as duas explicativas.

Modelo ZINB	AIC
$y_i = \beta_0$ *	128921,1
$y_i = \beta_0$	128923,1
$y_i = \beta_0 + \beta_1 X_1$	128915,3
$y_i = \beta_0 + \beta_2 X_2$	128723,7
$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	128716,8

*Modelo Binomial Negativo

Capítulo 5

Discussão e Conclusão

A partir da análise descritiva e dos modelos de regressão linear podemos perceber que há uma relação muito interessante entre a quantidade de células somáticas encontradas e os dias de lactação, o que na prática é bem razoável. O tempo é de suma importância para o controle de infecções, já que no decorrer do tempo as lesões tendem a ser mais graves se não forem devidamente cuidadas. O modelo proposto visa ajustar dados positivos com excessos de 0s. Podemos ver que o modelo Binomial Negativo Inflacionado de Zeros abordado alcançou algumas melhoras, como a diminuição da variabilidade dos resíduos (Anexos 2 e 3). A aplicação a dados de contagem de células somáticas (CCS) apresentada com ZINB melhorou substancialmente a qualidade do ajuste, em relação ao modelo de contagem Poisson com excessos de zeros. Embora a distribuição ZINB tenha melhorado bastante o ajuste da CCS, a estimativa para *p* (proporção de zeros) foi não significativa. Assim, a fim de diminuir a complexidade das análises, poderíamos ajustar somente um modelo BN. Pela Figura 1, vemos que existe uma frequência elevada de valores positivos no extremo esquerdo dos dados o que impediu um perfeito ajuste pela distribuição ZIP.

Referências

COENTRÃO, C.M., SOUZA, G.N., BRITO, J.R.F., PAIVA, M.A.V. E LILENBAUM, W. (2008). *Fatores de Risco para mastite subclínica em vacas leiteiras. Arq. Bras. Med. Vet. Zootec.*, **60**(2), 283-288.

ANUÁRIO da Agropecuária Brasileira. 11.ed. São Paulo: OESP Gráfica, 2004.

SCHULTZ, L.H. (1977). *Somatic cells in milk physiological aspects and relationship to amount and composition of milk. J. Food Prot.*, v.40, p.125-131.

EVANGELISTA, D.T. (2008). *Comparação entre métodos de referência e eletrônico por citometria de fluxo na contagem bacteriana total (CBT) e de células somáticas (CCS) em leite submetido a diferentes*. Dissertação de Mestrado – Universidade Federal de Minas Gerais, Escola de Veterinária.

GUERREIRO, P.K.; MACHADO, M.R.F.; BRAGA, G.C.; *et al.*(2005). *Qualidade microbiológica de leite em função de técnicas profiláticas no manejo de produção*. *Ciência e Agrotecnologia*, v.29, n.1, p.216-222.

HARMON, R.J. (1998). *Fatores que afetam as contagens de células somáticas*. In: SIMPÓSIO INTERNACIONAL SOBRE QUALIDADE DE LEITE, Curitiba. p.7-15.

MONARDES, H. Programa de pagamento de leite por qualidade em Quebéc, Canadá. In: SIMPÓSIO NTERNACIONAL SOBRE QUALIDADE DO LEITE, 1., 1998, Curitiba. *Anais*... Curitiba: biblioteca da UFPR, 1998, p.40-43.

HARDING, F. *Milk quality*. New York: Blackie Academic & Professional, 1995. 165 p.

MARSHALL, R.T. *Standard methods for the examination of dairy products.* Baltimore: American public Health Association, 1992. 546 p.

INTERNATIONAL Dairy Federation. Methods for estimating colony forming units. *IDF Standard 56.* Brussels: International Dairy Federation, 1991a. 5 p.

Santos, M.V. e Fonseca, L.F.L. (2004). *Uso da contagem de células somáticas para o monitoramento da qualidade do leite. Curso Online: Monitoramento da Qualidade do Leite. Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo*.

Rigby, R. A. and Stasinopoulos, D. M. (2001). *The GAMLSS project: a flexible approach to statistical modelling. In: Klein, B. and Korsholm, L. (eds.), New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, pp. 249–256. Odense, Denmark.

Rigby, R. A. and Stasinopoulos, D. M. (2005). *Generalized additive models for location, scale and shape, (with discussion).* Appl. Statist., 54: 507–554.

Akantziliotou, K. Rigby, R. A. and Stasinopoulos, D. M. (2002). *The R implementation of Generalized Additive Models for Location, Scale and Shape.* In: Stasinopoulos, M. and Touloumi, G. (eds.), Statistical modelling in Society: Proceedings of the 17th International Workshop on statistical modelling, pp. 75–83. Chania, Greece.

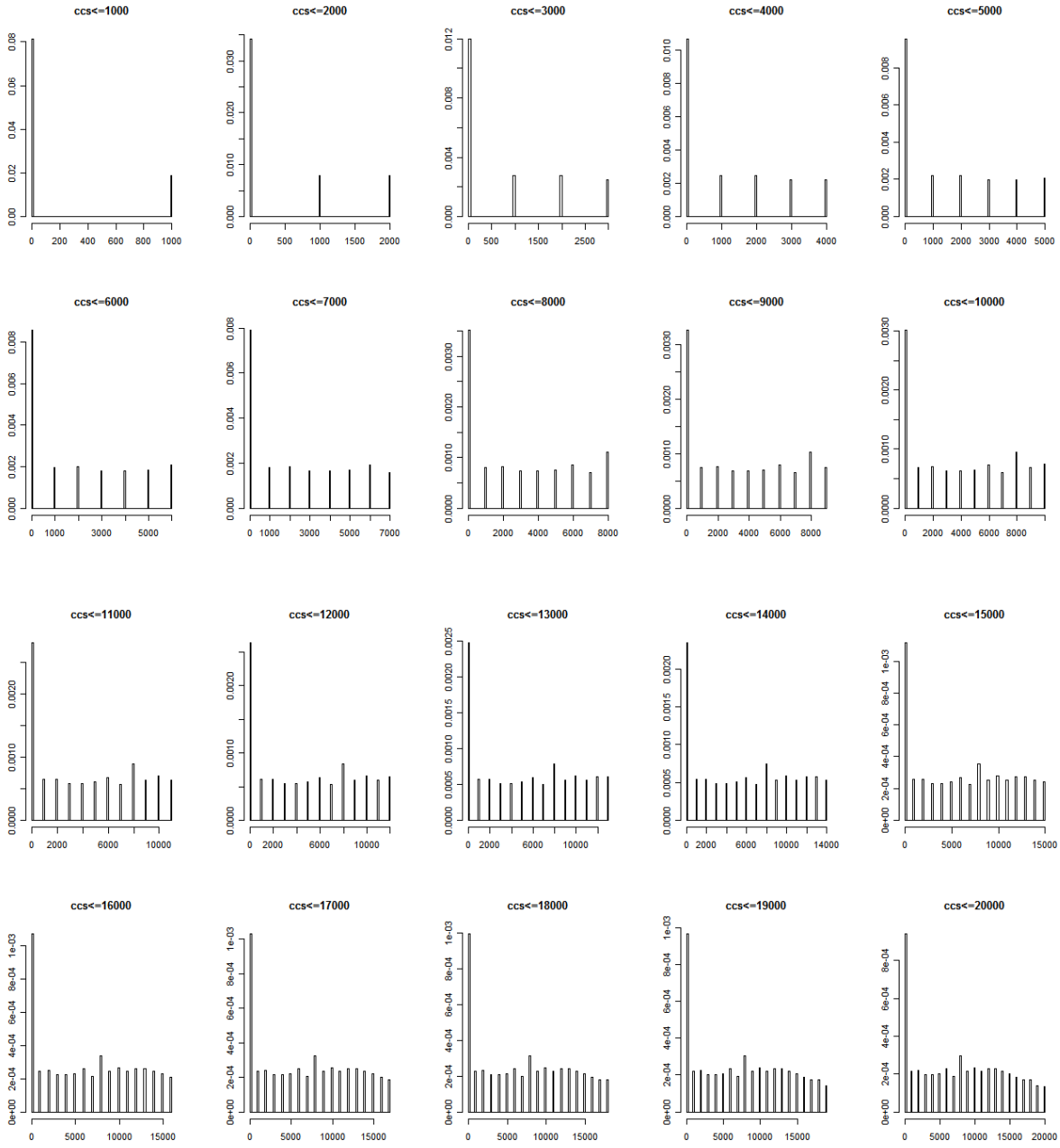
Yau, K., Wang, K. & Lee, A. (2003). *Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros*. *Biometrical Journal*,45, 4, 437-452.

Montoya, A.G.M. (2009). *Inferência e diagnóstico em modelos para dados de contagem com excesso de zeros.* Dissertação de Mestrado. Universidade Estadual de Campinas.

Evans, D. A. (1953). *Experimental evidence concerning contagious distributions in ecology*. Biometrika, 40: 186–211.

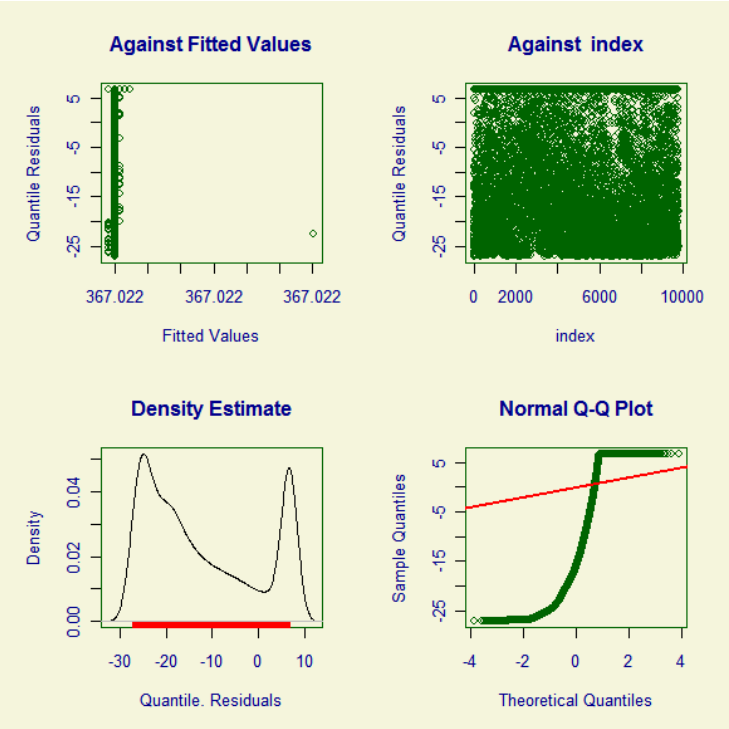
Anexos

A1. Histograma da CCS para valores baixos de contagem.

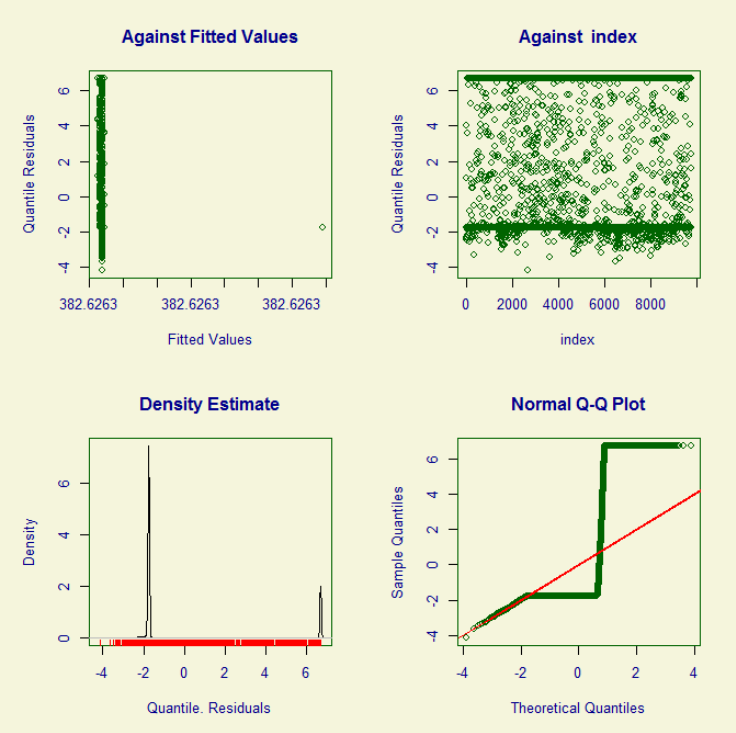


A2. Ajustes de Poisson

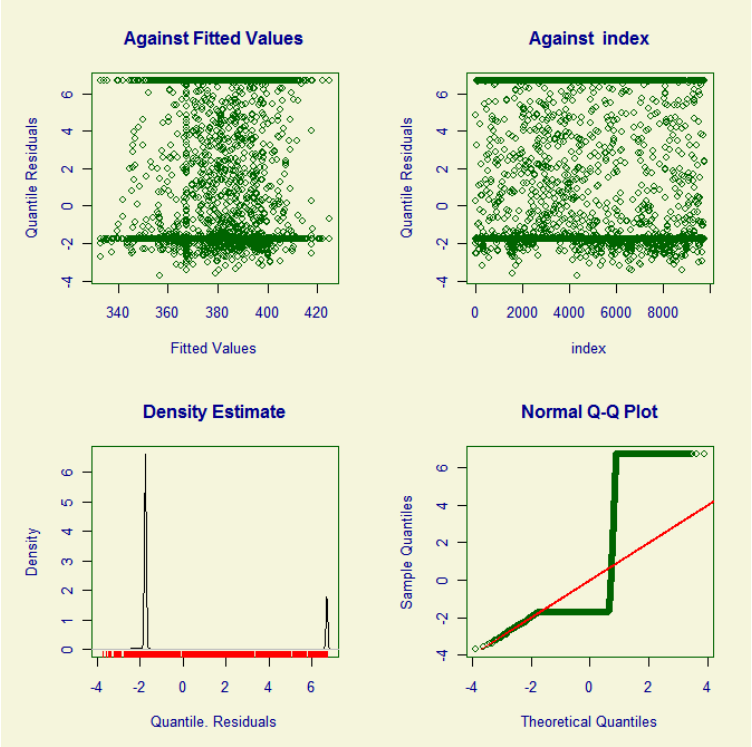
1. Resíduos de um ajuste Poisson



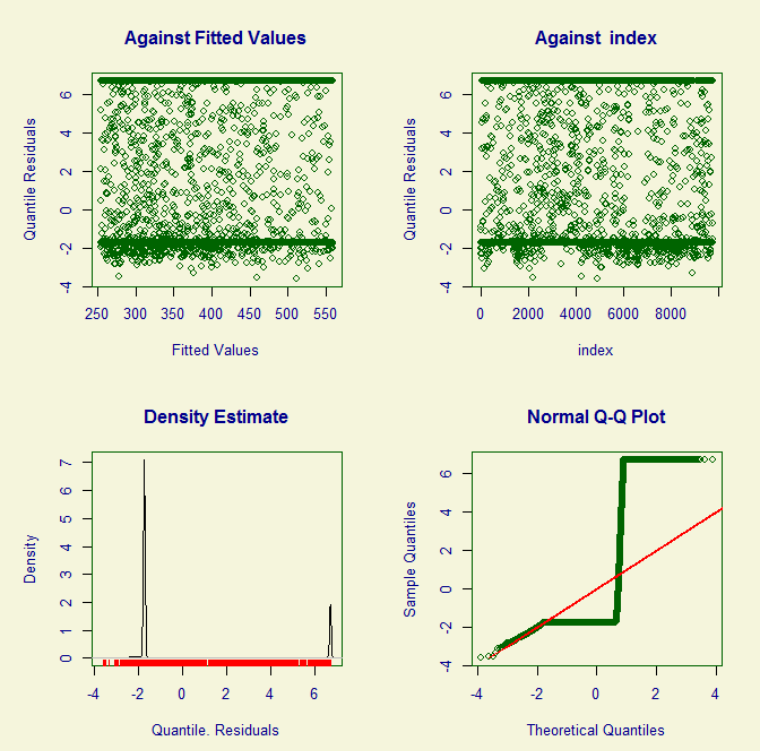
2. Resíduos de um ajuste ZIP



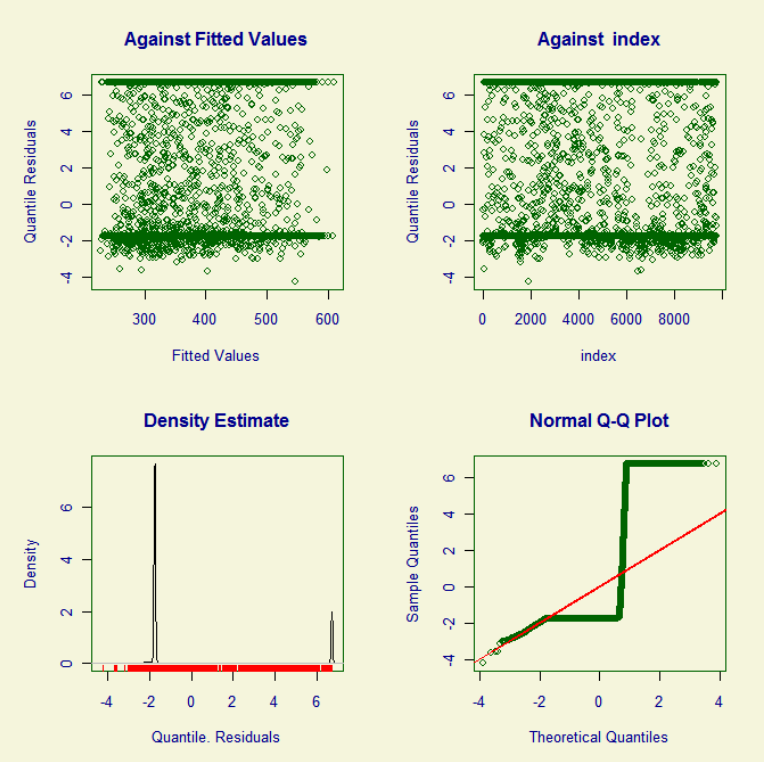
3. Resíduos de um ajuste ZIP com Idade



4. Resíduos de um ajuste ZIP com Dias de lactação

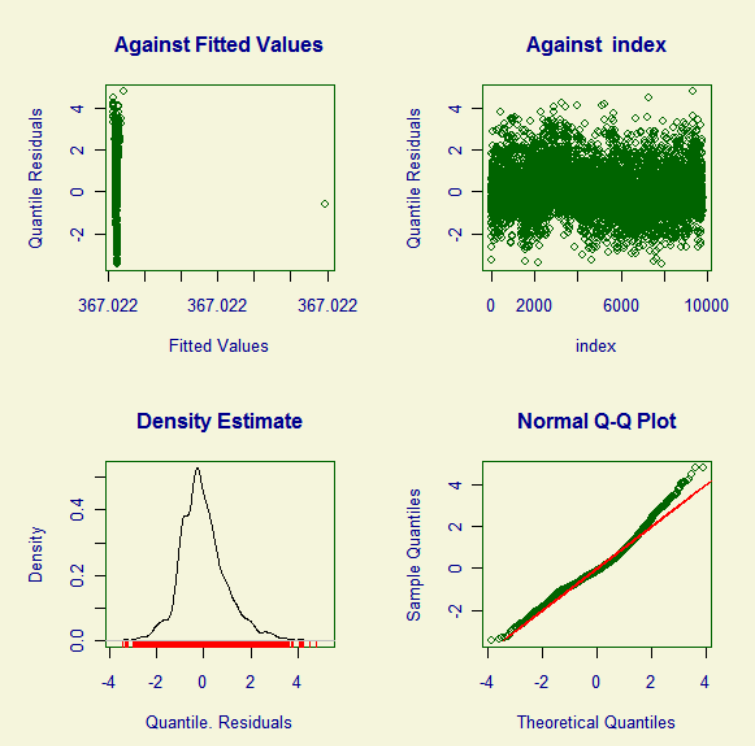


5. Resíduos de um ajuste ZIP com Idade e Dias de lactação

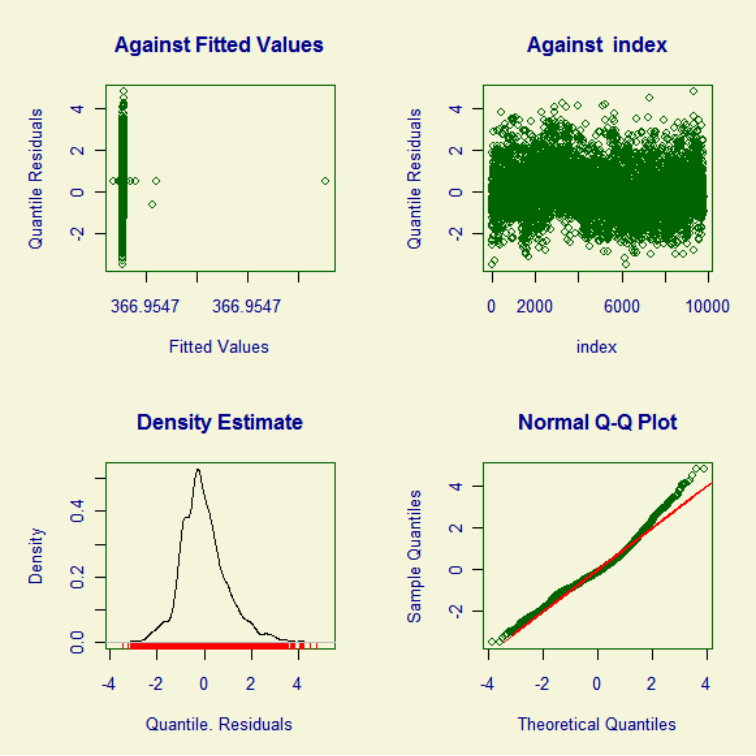


A3. Ajustes de Binomial Negativa

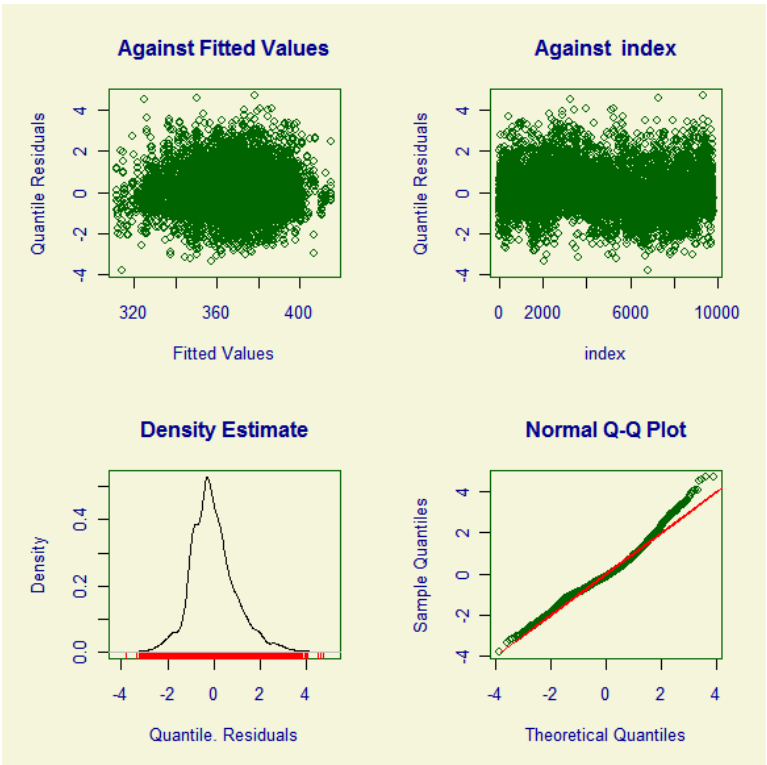
1. Resíduos de um ajuste BN



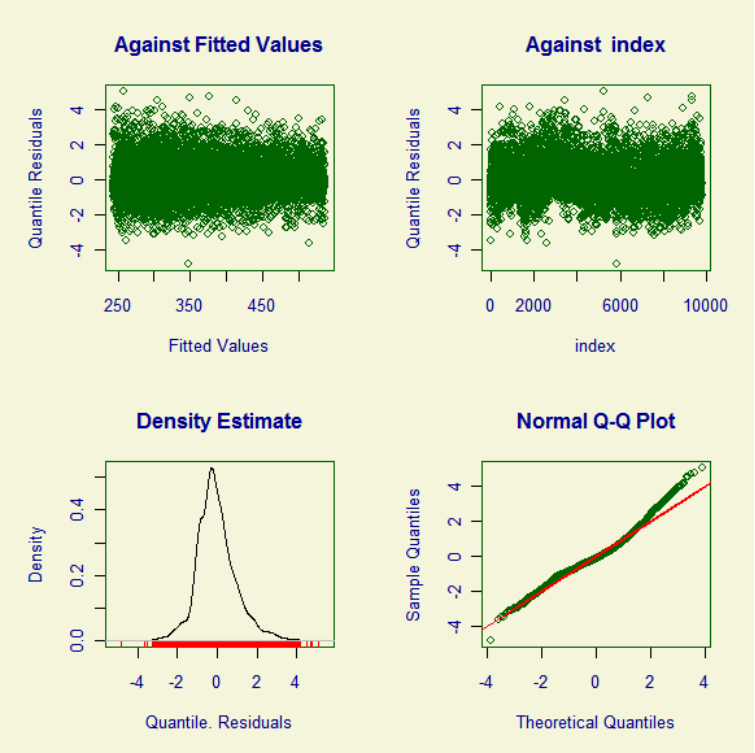
2. Resíduos de um ajuste ZINB



3. Resíduos de um ajuste ZINB com Idade



4. Resíduos de um ajuste ZINB com Dias de lactação



5. Resíduos de um ajuste ZINB com Idade e Dias de lactação.