Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Comparações de modelos estatísticos aplicados em dados zootécnicos de fazendas leiteiras

Erasnilson Vieira Camilo

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2019

Erasnilson Vieira Camilo
Bacharel em Estatística

Comparações de modelos estatísticos aplicados em dados zootécnicos de fazendas leiteiras

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:
Prof.^{a} Dr.^{a} **SÔNIA MARIA DE STEFANO PIEDADE**

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Camilo, Erasnilson Vieira

Comparações de modelos estatísticos aplicados em dados zootécnicos de fazendas leiteiras / Erasnilson Vieira Camilo. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2019 .

39 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Poisson 2. Superdispersão 3. Subdispersão 4. GAMLSS 5. COMPoisson . I. Título.

DEDICATÓRIA

Aos meus pais, Erasmo (in memoriam) e Margarida.
Ao meu irmão Erasnildo.

AGRADECIMENTOS

Agradeço a Deus por sua infinita bondade e misericórdia na minha vida.

A minha mãe Margarida pelo amor incondicional desde a minha existência, apoio e companheirismo mesmo na distância.

Ao meu pai Erasmo, que mesmo em memória sempre foi uma referência de caráter.

Ao meu irmão Erasnildo, meu amigo fiel que sempre esteve ao meu lado, me aconselhando e apoiando os meus sonhos.

A todos os professores e funcionários do departamento de Ciências Exatas da ESALQ/USP pelas contribuições dadas em todo o meu processo de conhecimento.

A minha orientadora Sônia Maria De Stefano Piedade, pelo carinho e disposição em me ajudar nos momentos mais difíceis da minha trajetória acadêmica.

Aos amigos que conquistei ao longo desses anos, Douglas, Rick, Simone, Daniel, Talita e Welder.

Agradeço especialmente a Andreza Jardelino pela parceria e colaboração, juntamente com Adriane de Andrade, professora da Universidade Federal de Uberlândia, e Ricardo Rosique Lara, pelo conjunto de dados cedido para elaboração da pesquisa.

Ao CNPq pelo auxílio financeiro concedido no período do doutorado.

A todos que contribuíram de forma direta ou indireta para a realização desse trabalho, o meu muito obrigado.

SUMÁRIO

Resumo 6
Abstract 7
1 Introdução 9
Referências 10
2 Modelagem Estatística: Uma abordagem para dados inflacionados de zeros aplicados a dados de rebanhos bovinos 13
Resumo 13
2.1 Introdução 13
2.2 Material e Métodos 14
2.2.1 Dados Simulados 16
2.2.2 Dados Reais 16
2.3 Resultados e Discussões 18
2.4 Considerações 22
Referências 23
3 Aplicação de Modelos de Contagens no Estudo de Inseminações Artificiais em Bovinos . 25
Resumo 25
3.1 Introdução 25
3.2 Material e Métodos 26
3.2.1 Dados Reais 29
3.2.2 Modelagem 30
3.3 Resultados e Discussões 30
3.4 Conclusão 36
Referências 37
4 Considerações finais 39

RESUMO

Comparações de modelos estatísticos aplicados em dados zootécnicos de fazendas leiteiras

Ao utilizar dados de contagens, frequentemente, tem-se o fenômeno da superdispersão. Entretanto, a variabilidade menor do que a esperada pelo modelo especificado, não é tão comum, sendo este conhecido como subdispersão. Nesse trabalho, foi dado ênfase ao estudo de técnicas que possibilitem a modelagem de observações com características de dispersões. A aplicabilidade empregada para proporcionar tal finalidade, foi por meio de pesquisas relacionadas a biotecnologias reprodutivas. Assim, no segundo capítulo foi proposto o estudo da distribuição COMPoisson, que considera tanto a variabilidade extra quanto o excesso de zeros, via dados simulados e dados reais. A variável resposta utilizada foi resultante do número de lactações, por apresentar valores inflacionados de zeros. Em síntese, a aplicação de dados reprodutivos bovinos colaboraram para a validação do modelo proposto, possibilitando uma alternativa de ajuste para pesquisadores que tenham observações com estrutura semelhante. No capítulo 3, a variável de interesse foi o número de prenhez, e para o ajuste foram considerados diferentes modelos aditivos generalizados de escala e forma, possibilitando alternativas adequadas para analisar dados de inseminação artificial. Os modelos apresentaram consistência quanto aos resultados mas o mais indicado foi a distribuição Delport, visto que essa apresentou resultados satisfatórios, principalmente na verificação da qualidade do ajuste. Além disso, em virtude dos resultados encontrados pode-se dizer que houve efeito de fazendas, estágios de vida da vaca, e estação do ano na condição de prenhez.

Palavras-chave: Poisson, Superdispersão, Subdispersão, GAMLSS, COMPoisson

ABSTRACT

Comparisons of statistical models applied to dairy farm zootechnical data

When using counting data, frequently, has the phenomenon of overdispersion. However, the variability smaller than expected by the specified model is not so common, which is known as underdispersion. In this work, the emphasis was given to study whose methods make possible the modeling of observations with these characteristics of dispersions. The applicability employed to provide this purpose was through research related to reproductive biotechnologies. Thus, in the Chapter 2, the study of the COMPoisson distribution was proposed, which considers both extra variability and excess zeros, use simulated data and real data. The response variable used is the result of the number of lactations, since they have inflated values of zeros. In summary, the application of bovine reproductive data contributed to the validation of the proposed model, allowing an alternative of adjustment for researchers who have observations in the presented structure. In Chapter 3, the number of pregnancies was obtained as the response variable, in which the different models of generalized scale and shape were considered for this adjustment, allowing adequate alternatives to analyze artificial insemination data. However, although the models presented consistency in the results, it can be said that the Delport distribution was the most indicated, since this one presented satisfactory results, mainly in the verification of the adjustment quality. In addition, due to the results found it can be said that there was an effect of the different farms studied, stages of life of the cow, and season of the year in the condition of pregnancy.

Keywords: Poisson, Overdispersion, Underdispersion, GAMLSS, COMPoisson

1 INTRODUÇÃO

Na pesquisa reprodutiva em bovinos, frequentemente são coletados dados de contagens. Com isso, há um crescente estudo acerca das técnicas de reprodução animal, visto que essa prática proporciona melhorias quanto à forma de seleção animal (Vaz et al., 2010; Roche, 2006).

É recorrente a busca sobre o conhecimento por metodologias que viabilizem análises de dados com essa característica, uma vez que deve ser considerada a natureza das observações na modelagem. Os modelos de regressão expressivamente utilizados nas análises estatísticas consistem em relacionar a média da variável resposta, uma variável aleatória, com as demais covariáveis explicativas, para evidenciar as possíveis influencias sofridas dentre elas (Hoffmann, 2006). Destaca-se a distribuição Gaussiana como o uso predominante entre as distribuições assumidas para a variável de interesse.

Todavia, tentar ajustar dados de contagens por meio do modelo normal resultará em estimativas dos erros-padrão inconsistentes, ocasionando predições negativas para o número de eventos (King, 1989). Até então, diferentes formas de transformações da variável resposta foram aplicadas com finalidade de atender os pressupostos do modelo Gaussiano (de Paula, 2013). No entanto, diante das dificuldades encontradas devido à estrutura das observações, Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MLGs), os quais apresentam uma classe de modelos mais flexíveis para a distribuição da variável de interesse que estará associada a um preditor linear por meio de uma função de ligação.

Na hipótese em que é assumido o modelo de Poisson, associado a modelagem de dados de contagens, tem-se a particularidade da equidispersão, isto é, $E(Y) = Var(Y) = \lambda$. Contudo, na prática essa relação não ocorre, sendo recorrente situações em que há uma variabilidade muito maior ou muito menor que a especificada pelo modelo, conhecida como superdispersão e subdispersão, respectivamente (Hinde e Demétrio, 1998b).

Há casos em que a superdispersão apresentada na modelagem de dados pode ser interpretada erroneamente, pois quando o modelo for mal especificado, isto é, na omissão de variáveis no preditor linear, pode-se interpretar as significâncias dos efeitos estimados erroneamente. Em geral, análises estatísticas consideram apenas um parâmetro de dispersão para captar tais variações, porém no mesmo ajuste, é possível atribuir diferentes fontes de dispersões à diferentes fatores. Sobre o tema, é notável que, a depender da variável a ser estudado, há de fato um comportamento de dispersão distinto. Dessa forma, uma alternativa possível dos MLGs seria a distribuição Conway – Maxwell – Poisson que além de acomodar diferentes estruturas de variações nas observações, levam em consideração o excesso de zeros presentes nos dados (Conway e Maxwell, 1962). Esse ajuste é possível devido à generalização da distribuição Poisson com a adição de um parâmetro denotado por ν, que por sua vez torna a razão de probabilidades sucessivas de forma não linear, o que acomodará o fenômeno extra Poisson (Shmueli et al., 2005b).

Em uma nova perspectiva, uma extensão dos MLGs foi proposta pelos autores Hastie e Tibshirani (1990), os quais incorporam uma função de suavização no preditor linear, caracterizando os Modelos Aditivos Generalizados (GAMs). Mais tarde, seriam desenvolvidos os modelos aditivos generalizados de locação e escala (GAMLSS), que trata-se de uma regressão semi-paramétrica, ou seja, assume-se uma mistura entre a distribuição paramétrica e uma função de suavização das variáveis explicativas, sendo essa uma combinação dos MLGs com os GAMs

(Stasinopoulos et al., 2005).

Assim, a busca por técnicas inovadoras quanto à análise de dados é de grande importância também na área das ciências animais. Dessa forma, o presente estudo tem por objetivo propor modelagens de dados com essa estrutura dos apresentados nesse trabalho. No segundo capítulo, serão apresentados dados simulados e a utilização de dados reais, que considerou como variável resposta o número de lactações, isto é, a quantidade de vezes que a fêmea bovina manteve e concebeu a gestação (BERGAMASCHI et al., 2010). Essa variável se torna interessante por apresentar uma quantidade excessiva de zeros. Para tal modelagem, foi empregada a distribuição COMPoisson. No terceiro capítulo, propõe-se uma comparação entre diversas distribuições da classe GAMLSS, objetivando a solução de problemas e modelagem estatística na área da zootecnia, especificamente em estudos de avaliações com respostas obtidas por meio das técnicas de biotecnologias reprodutivas. Considerou-se como variável resposta o número de vacas prenhes, as quais foram submetidas ao protocolo de superovulação. Os modelos contemplados para as comparações foram Poisson, Double Poisson, Poisson-normal-inversa, Delaporte e Sichel. Por fim, no quarto capítulo são apresentadas as considerações finais obtidas nesse estudo e perspectivas futuras no que tange as possíveis pesquisas.

Referências

BERGAMASCHI, M. A. C. M., MACHADO, R., e BARBOSA, R. T. (2010). Eficiência reprodutiva das vacas leiteiras. *Embrapa Pecuária Sudeste-Circular Técnica (INFOTECA-E)*.

Conway, R. W. e Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136.

de Paula, G. A. (2013). Modelos de Regressão com apoio computacional. Technical report, IME - USP, São Paulo.

Hastie, T. e Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, pages 1005–1016.

Hinde, J. e Demétrio, C. G. (1998b). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.

Hoffmann, R. (2006). *Análise de Regressão - Uma Introdução a Econometria*. Hucitec, S o Paulo, 4 edition.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science*, pages 762–784.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.

Roche, J. F. (2006). The effect of nutritional management of the dairy cow on reproductive efficiency. *Animal reproduction science*, 96(3-4):282–296.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., e Boatwright, P. (2005b). A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Stasinopoulos, M., Rigby, B., Akantziliotou, C., e Voudouris, V. (2005). Generalized additive models for location scale and shape. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54:507–554.

Vaz, R. Z., Lobato, J. F. P., e Restle, J. (2010). Productivity and efficiency of cow herds submitted to two weaning ages. *Revista Brasileira de Zootecnia*, 39(8):1849–1856.

2 MODELAGEM ESTATÍSTICA: UMA ABORDAGEM PARA DADOS INFLACIONADOS DE ZEROS APLICADOS A DADOS DE REBANHOS BOVINOS

Resumo

Na análise de dados de contagem pode ocorrer a presença de superdispersão ou subdispersão nos dados, uma vez que a distribuição de Poisson tem a característica de apresentar esperança e variância iguais. Quando a característica de equidispersão não for satisfeita, a distribuição se tornará inapropriada para essa modelagem. Uma possível alternativa para o ajuste de modelos de regressão se dá por meio do emprego da distribuição COMPoisson, que considera tanto a variabilidade extra quanto o excesso de zeros. Nesse sentido, foram realizadas análises via dados simulados, os quais foram construídos a partir da distribuição COMPoisson com diferentes tipos de dispersões e presença de zeros. Para fins de validação, utilizou-se observações reais referentes ao controle zootécnico em que os rebanhos bovinos foram submetidos à biotecnologia da inseminação artificial. A variável resposta utilizada é resultante do número de lactações, por apresentar valores inflacionados de zeros. Em síntese, a aplicação de dados reprodutivos bovinos colaboraram para a validação do modelo proposto.
Palavras-chave: Poisson; COMPoisson; Superdispersão; Subdispersão; Biotecnologia Reprodutiva.

2.1 Introdução

As análises de dados devem considerar a natureza da variável de interesse, sejam elas contínuas ou discretas. Nos modelos estatísticos, a escolha do Modelo Linear Generalizado (MLG) (Nelder e Wedderburn, 1972a) é devida à possibilidade de acomodar estruturas mais flexíveis para a variável de interesse que outras abordagens não conseguem. No entanto, há casos em que a variabilidade dos dados é maior ou menor do que a especificada pelo modelo, sendo esses fenômenos conhecidos por superdispersão ou subdispersão, respectivamente.

O ajuste de modelos que não consideram essa variação extra podem ocasionar uma superestimação, ou subestimação dos erros-padrão, levando a interpretações incorretas quanto as significâncias das estimativas dos parâmetros. Dessa forma uma alternativa possível dos MLGs clássicos seria a distribuição Conway – Maxwell – Poisson (Conway e Maxwell, 1962), que além de acomodar essas estruturas, levam em consideração o excesso de zeros presentes nos dados.

Na prática, estudos relacionados a eficiência reprodutiva em bovinos são de grande interesse nas ciências animais (Vaz et al., 2010; Roche, 2006). Entretanto, a forma de coleta dessas variáveis ocorre, frequentemente, por meio de mensurações de dados quantitativos na forma de contagens, e qualitativos para aqueles cuja ordenação não seja possível, a exemplo das observações de diagnóstico de prenhez (prenhas/vazias). Assim, torna-se importante a utilização de ferramentas que possibilitem a validação de modelos. Dessa forma, o presente estudo tem o objetivo de propor modelagens de dados com essas características e comparar os modelos Poisson e COMPoisson, viabilizando, nesse caso, uma alternativa aplicável para as análises da área.

Com tal finalidade foram feitas simulações da distribuição COMPoisson e o utilização de dados reais, com a variável resposta número de lactações, isto é a quantidade de vezes que a fêmea bovina (vaca) apresentou uma gestação completa (prenhez completa) (BERGAMASCHI et al., 2010).

2.2 Material e Métodos

Modelo ComPoisson (COMP)

A distribuição de COMPoisson (Conway e Maxwell, 1962) é de uma generalização da distribuição de Poisson que pertence a família exponencial, denotada por

$$P(Y = y_i) = \frac{{\lambda_i}^{y_i}}{(y_i!)^{\nu}} \frac{1}{Z(\lambda_i, \nu)} \quad i = 1, ..., n \tag{2.1}$$

e os momentos descritos como

$$E(Y^{r+1}) = \begin{cases} \lambda E(Y+1)^{1-\upsilon}, \ r = 0 \\ \lambda \frac{d}{d\lambda} E(Y^r) + E(Y)E(Y^r), \ r > 0 \end{cases}, \tag{2.2}$$

sendo

$$E(Y) = \lambda \frac{d\{log[Z(\lambda, \upsilon)]\}}{d\lambda} \approx \lambda^{1/\upsilon} - \frac{\upsilon - 1}{2\upsilon},$$

em que

$$Z(\lambda_i, \upsilon) = \sum_{j=0}^{\infty} \frac{{\lambda_i}^j}{(j!)^{\upsilon}} \ \ \lambda_i > 0, \ \upsilon > 0 \ \ e \ \ \ln(\lambda_i) = {X_i}'\beta. \tag{2.3}$$

Assim, o parâmetro ν poderá assumir diferentes distribuições associadas a modelagem de dados de contagem, ou seja, para $\nu = 0$, tem-se a distribuição geométrica, para $\nu = 1$, a distribuição de Poisson, e $\nu \to \infty$, caracteriza-se a distribuição de Bernoulli. Uma caraterística dessa distribuição consiste em apresentar valores positivos para ν, dificultando na escolha do método de otimização. Desse modo, uma possível solução seria considerar $\phi = log(\nu)$, uma vez que, se $0 < \nu < \infty$ logo $-\infty < \phi < \infty$. Desse modo, para $\phi > 1$ e $\phi < 1$ tem-se a subdispersão e superdispersão, respectivamente, enquanto que para $\phi = 1$ caracteriza-se a equidispersão.

Shmueli et al. (2005) mostram que é possível descrever a distribuição COMPoisson na família exponencial, de tal forma que exista um vetor de estatística suficiente associado a um vetor de parâmetros desconhecidos, isto é, $T(\mathbf{Y}) = \left[\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} \log(Y_i!)\right]$.

A estimação dos parâmetros do modelo COMPoisson está associada a maximização da função de verossimilhança, descrita por

$$\log L(\beta, \upsilon) = \sum_{i=1}^{n} y_i \log \lambda_i - e^{\phi} \sum_{i=1}^{n} \log y_i! - \sum_{i=1}^{n} \log Z(\lambda_i, \upsilon) \tag{2.4}$$

A Equação (2.4) não pode ser resolvida de forma analítica, sendo necessário a utilização de métodos computacionais iterativos, dos quais o método de Newton-Rahpson é um dos mais abordados na literatura. O processo iterativo consiste em maximizar a equação de verossimilhança de forma que, a cada passo da iteração, um novo vetor de estimativas será atualizado até que se obtenha a convergência por meio de um critério pré-estabelecido.

Desse modo, a avaliação da expressão dada em (2.4) apresenta a constante de normalização, Equação (2.3), sendo calculada para cada observação, potencializando neste caso a lentidão no processo de estimação.

Modelo COMPoisson Inflacionado de Zeros (ZICMP)

O excesso de zeros em conjunto de dados implica, na maioria das vezes, em uma redução do valor da média diferindo bastante do valor da variância, podendo assim ser caracterizado pelo fenômeno da superdispersão dos dados, porém, em algumas situações, o excesso de zero também pode indicar um fenômeno de subdispersão, ou seja, em situações em que a variabilidade dos dados é inferior a média. Nesse contexto, a distribuição COMPoisson inflacionada de zeros abordada em Sellers e Raim (2016) é uma opção para captar diferentes tipos de variações como também modelar conjuntos de dados com excesso de zeros.

No estudo de simulação proposto por Sellers e Raim (2016), os autores mostraram a capacidade com que a distribuição COMPoisson tem para modelar com precisão os conjuntos de dados inflacionados de zeros, uma vez que os autores utilizaram dados simulados para a exemplificação dos modelos de regressão associados a diversas distribuições de contagens, sendo estes casos particulares da distribuição COMPoisson.

A distribuição ZICMP é evolução da distribuição COMPoisson, sendo representada pela variável aleatória Y_i, $i = 1, 2, 3, ..., n$, com a seguinte estrutura

$$Y_i \sim \begin{cases} 0, & \text{se} \quad p_i \\ \text{COM-Poisson}(\lambda_i, \nu), & \text{se} \quad (1-p_i)\,, \end{cases}$$

em que

$$\begin{aligned} P(Y_i = y_i) &= \left[p_i + (1-p_i)\left(\frac{1}{Z(\lambda_i,\nu_i)}\right)\right]^{u_i}\left[\frac{(1-p_i){\lambda_i}^{y_i}}{(y_i!)^{\nu_i} Z(\lambda_i,\nu_i)}\right]^{1-u_i} \\ &= \left[\frac{p_i(Z(\lambda_i,\nu_i)-1)+1}{Z(\lambda_i,\nu_i)}\right]^{u_i}\left[\frac{(1-p_i){\lambda_i}^{y_i}}{(y_i!)^{\nu_i} Z(\lambda_i,\nu_i)}\right]^{1-u_i}. \end{aligned} \tag{2.5}$$

Considera-se ainda uma variável binária denotada por u_i e com estrutura

$$u_i = \begin{cases} 1, & \text{se } y_i = 0 \\ 0, & \text{se } y_i \neq 0. \end{cases}$$

A Equação (2.5) trata-se de uma mistura de modelos para dados de inflacionados de zeros; dessa forma para qualquer observação em $y_i \neq 0$ a sua probabilidade $P(Y = y_i)$ resume-se à segunda expressão da Equação (2.5), que é ponderada pela probabilidade de (1-p) associada a distribuição COMPoison. Assim, há uma subestimação da distribuição COM-Poisson quando esta apresentar uma quantidade de zeros elevada. Nesse sentido, tem-se a função de verossimilhança da distribuição ZICMP dada por

$$\begin{aligned} logL(\lambda,\nu,p|y) &= \sum_{i=1}^{n}\left\{u_i \log\left(p_i\frac{(1-p_i)}{Z(\lambda_i,\nu_i)}\right) + (1-u_i)\left[\log(1-p_i) + y_i\log(\lambda_i) - \nu_i\log(y_i!) - \log Z(\lambda_i,\nu_i)\right]\right\} \\ &= \sum_{i=1}^{n}\{u_i \log\left(p_i Z(\lambda_i,\nu_i) + (1-p_i)\right) + (1-u_i)\left[\log(1-p_i) + y_i\log(\lambda_i) - \nu_i\log(y_i!)\right] \\ &\quad - \log Z(\lambda_i,\nu_i)\} \end{aligned} \tag{2.6}$$

em que $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)^T$ e $\mathbf{p} = (p_1, p_2, ..., p_n)^T$ são modelados a partir das funções de ligação dos modelos lineares generalizados na forma $log\,(\boldsymbol{\lambda}) = \mathbf{X}\boldsymbol{\beta} \Rightarrow \boldsymbol{\lambda} = \exp(\mathbf{X}\boldsymbol{\beta})$ e $\text{logit}(\mathbf{p}) = \mathbf{W}\boldsymbol{\zeta} \Rightarrow \mathbf{p} = \text{logit}^{-1}(\mathbf{W}\boldsymbol{\zeta})$. Ademais, a dispersão será ajustada via função de ligação dada por $\log(\boldsymbol{\nu}) = \mathbf{G}\gamma \Rightarrow \boldsymbol{\nu} = \exp(\mathbf{G}\boldsymbol{\gamma})$, onde $\boldsymbol{\nu} = (\nu_1, \nu_2, ..., \nu_n)^T$. Nesse sentido, segue a função de máxima verossimilhança expressa por

$$\begin{aligned}\log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}|y) = \sum_{i=1}^{n} \{ & u_i \log\left(p_i \left[Z\left(\exp(\mathbf{X}_i\boldsymbol{\beta}), \exp(\mathbf{G}_i\boldsymbol{\gamma})\right) - 1\right] + 1\right) \\ & + (1-u_i)\left[-\log(1+\exp(\mathbf{W}_i\boldsymbol{\zeta})) + y_i\mathbf{X}_i\boldsymbol{\beta} - \exp(\mathbf{G}_i\boldsymbol{\gamma})\,\log(y_i!)\right] \\ & - \log Z\left(\exp(\mathbf{X}_i\boldsymbol{\beta}), \exp(\mathbf{G}_i\boldsymbol{\gamma})\right)\}\end{aligned}$$

sendo um caso particular de equidispersão da Equação (2.6) resultando no modelo ZIP dado por

$$\begin{aligned}logL(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}|y) = \sum_{i=1}^{n} \{ & u_i \log\left[\exp(\mathbf{W}_i\boldsymbol{\zeta}) + \exp(-\exp(\mathbf{X}_i\boldsymbol{\beta}))\right] - \log(1+\exp(\mathbf{W}_i\boldsymbol{\zeta})) \\ & + (1-u_i)\left[-y_i\mathbf{X}_i\boldsymbol{\beta} - \log(y_i!) - \exp(\mathbf{X}_i\boldsymbol{\beta})\right]\}.\end{aligned}$$

Uma outra abordagem apresentada por Sellers e Raim (2016) ilustra diversos casos de aplicação da distribuição ZICMP, associada a simulação estatística. Casos particulares desta são as distribuições Geométrica (ZIG) e Poisson inflacionada de zeros (ZIP). Além disso, tem-se a distribuição Bernoulli quando $\nu \longrightarrow \infty$, nesse caso, é possível utilizar a ZICMP para estimação do parâmetro constante $\mathbf{p}$ em que $Logit(p) = \mathbf{W}_i\zeta$, sendo $\mathbf{W}_i$ o vetor de uns.

2.2.1 Dados Simulados

Para a simulação dos dados foi considerado o Modelo COM-Poisson em que Y é a variável resposta gerada dessa distribuição sendo associada a variável explicativa discreta X, por um "Fator" que por sua vez possui dois níveis "A" e "B" de comprimento $n = 400$ cada, sendo possível associar uma mistura de dispersão no modelo de regressão, de forma que a matriz de delineamento do "Fator" foi utilizada para obter tal efeito.

2.2.2 Dados Reais

Os dados foram obtidos à partir de anotações zootécnicas de dados reprodutivos de cinco rebanhos pertencentes às fazendas da região de Passos, Minas Gerais, Brasil. Tais fazendas são caracterizadas por apresentarem manejos similares em relação à forma de alojamento com sistema de criação "*Free Stall*", em que os animais têm baias individuais para descanso, área de alimentação, sendo oferecida dieta de acordo com exigência nutricional para produção de leite, água de qualidade "*ad libitum*", sistema de controle ambiental para manutenção térmica, ideal para permitir o conforto do animal, com a presença de cortinas (controle de entrada de radiação solar), ventiladores e nebulizadores (controle de temperatura e umidade). Ademais, são realizadas limpezas periódicas nas instalações.

Dessa forma, os dados utilizados são resultantes de um estudo por amostragem conduzido em cinco fazendas denominadas como (1 - Brejo, 2 - IPE, 3 - Morro Branco, 4 - Petrópolis, 5 - Praia Vermelha). A pesquisa consistiu de 926 observações de vacas com padrão racial Holandês, em que foram avaliadas as variáveis número de lactação (0, 1, 2, 3, 4, 5, 6, 7), idade das

vacas até a última inseminação, compreendido no período de 2004 a 2017. Todos os animais foram submetidos à técnica de inseminação artificial, como biotecnologia reprodutiva padrão.

Sob a característica de subdispersão e com aproximadamente 30% dos valores representados por zeros no conjunto de dados, observa-se na Figura 2.1 as frequências das observações em relação ao número de lactações e o comportamento em cada uma das cinco fazendas.

Portanto, observou-se ainda que há diferenças entre as taxas de lactações das fazendas avaliadas, em que a fazenda Petrópolis apresentou o maior número de animais (40%) sem lactação, indicando que a fazenda realizou uma renovação do rebanho e apresenta-se com alto número de novilhas (animais jovens que ainda não estão em idade de parição) não necessariamente representando um problema reprodutivo no rebanho. As fazendas Morro Branco (27%) e Ipê (26%) apresentaram valores similares de ausência de lactação, sendo esses valores considerados ideais por Mion et al. (2012) de 30% para percentual de vacas secas em um rebanho bovino leiteiro.

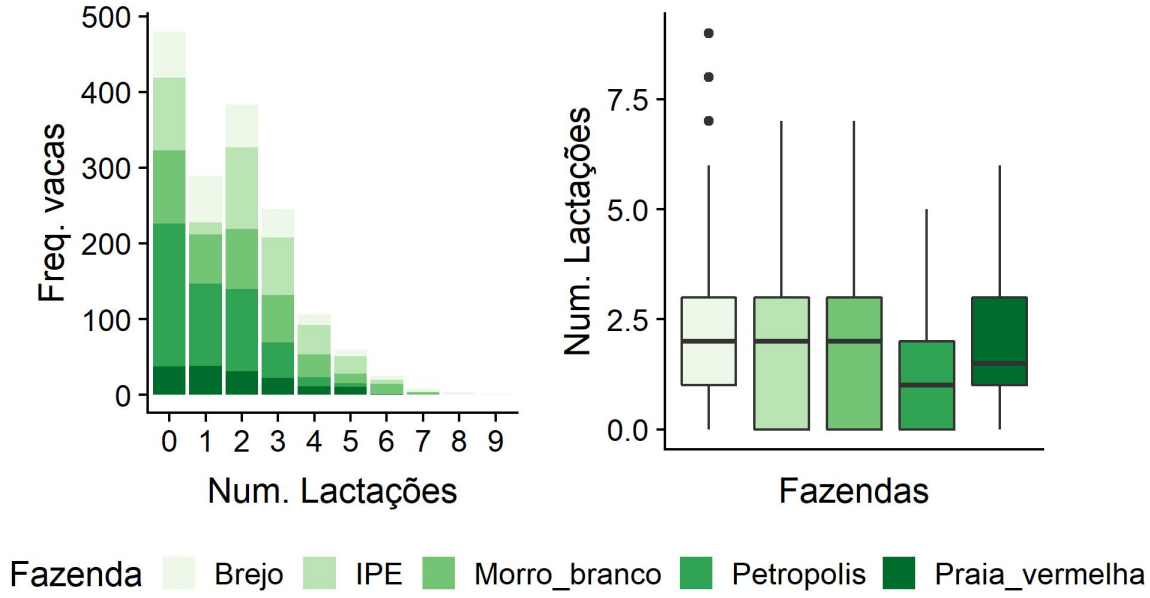


Figura 2.1. Gráfico de barras para o número de lactações *versus* o número de vacas, por fazenda; Box-Plot referente as fazendas *versus* o número de lactações, respectivamente

Os pontos discrepantes podem ser explicados uma vez que algumas fazendas realizam a manutenção de alguns animais com idades elevadas, próximo de 10 anos, porém, mas que apresentam média de produção de leite maior do que o rebanho padrão e assim justificando a sua permanência no plantel, enquanto estiverem prolíferas (vacas com capacidade de se reproduzirem).

Considerando a elevada taxa de zeros observados nos dados das fazendas, relacionada ao número de prenhez, este trabalho propôs a utilização do modelo COMPoisson inflacionado de zeros, cujas estimativas dos parâmetros serão obtidas por meio do Método da Máxima Verossimilhança (MMV).

Assim, o modelo é expresso em (2.7) por

$$\eta_i = \log(\lambda_{ij}) = \beta_0 + \beta_{1i} + \beta_2 Idade \tag{2.7}$$

em que β_0 é uma constante de efeito fixo do modelo, β_{1i} o efeito fixo da Fazenda i, β_2 o efeito fixo da idade, a dispersão e a influência de zeros serão modeladas por meio dos parâmetros $\text{logit}(\mathbf{p}) = \mathbf{W}\boldsymbol{\zeta}$ e $\log(\boldsymbol{\nu}) = \mathbf{G}\gamma$, respectivamente.

A proposta do modelo na Equação (2.7) será obtido por meio do pacote *COMPoisson-Reg* do *Software* R.

2.3 Resultados e Discussões

Nesse trabalho foi abordado uma modelagem mais flexível para análise de dados de contagens baseados na família de distribuição Poisson, a qual naturalmente apresenta o fenômeno de super ou subdispersão, uma vez que a $E(Y) = V(Y) = \mu$ (Nelder e Wedderburn, 1972b). Dessa forma, ao introduzir estudos de simulação, os resultados foram similares aos encontrados em Sellers e Shmueli (2010), no qual foi evidenciado que o modelo COMPoisson demonstrou uma maior flexibilidade quanto à presença de zeros.

Por meio de simulação estatística, foi possível ajustar um modelo COMPoisson inflacionado de zeros, como descrito na subseção 2.2.1, em que foi considerada a presença do fenômeno não atípico para a modelagem de dados de contagem, denominada misturas de dispersões que está associada ao ajuste do modelo linear generalizado, o qual refere-se à presença de mais de um tipo de dispersão, ou seja, a subdispersão e superdispersão nos dados, podendo ser separadas por níveis de um fator em uma modelagem estatística ou por meio de uma variável binária.

Na prática, o emprego de diferentes tipos de dispersões está associado geralmente de forma única a todo o conjunto de dados, podendo ocorrer uma falsa equidispersão nos mesmos. Dessa forma, Sellers e Shmueli (2010) propuseram uma forma de captar as diferentes variabilidades dentro da mesma amostra a partir da utilização de variáveis dammys, sendo esta adicionada no preditor linear e associada ao ajuste da dispersão. De forma similar, porém, modelando a parte inflacionada de zeros, Sellers e Raim (2016) apresentaram uma modelagem COMPoisson inflacionada de zeros para captar o efeito gerado pelos excessos de zeros.

A partir da Equação (2.5) foi possível identificar as diferentes funções de ligação empregadas para captar os efeitos de dispersão e excesso de zeros associados ao modelo COMPoisson. Na Figura 2.2, por exemplo, é apresentada uma síntese das estimativas dos referidos parâmetros utilizados como valores reais para o modelo simulado, uma vez que suas respectivas respostas associadas aos valores de $\gamma = -0.3$ e $\gamma = 0.3$, são ilustradas pelas linhas nas cores vermelho e marrom, respectivamente. Tais valores expressam a superdispersão e subdispersão utilizadas na simulação, enquanto que o valor associado a $\zeta = -1$ representa a probabilidade de excessos de zeros no modelo.

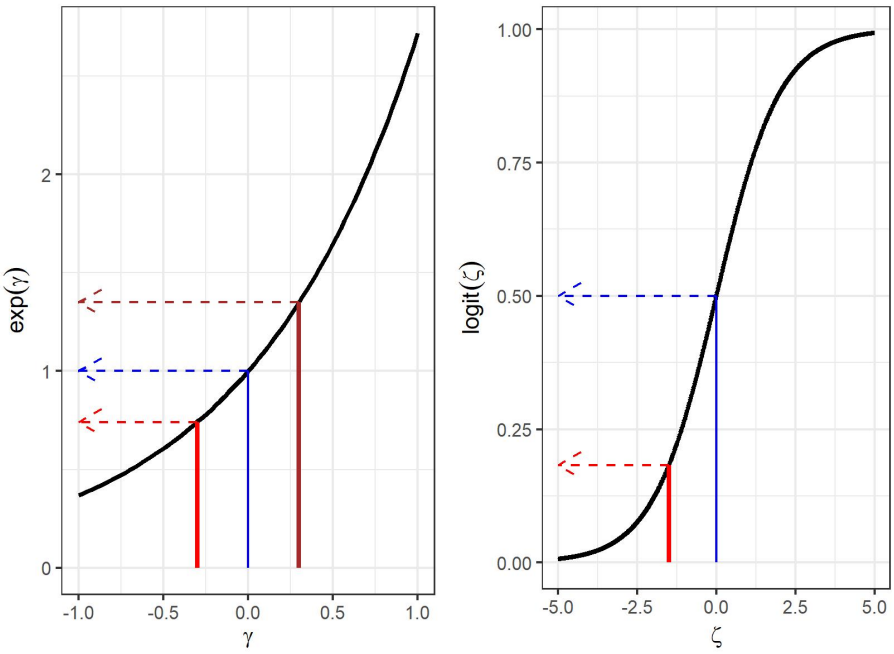


Figura 2.2. Função de ligação associada aos parâmetros de dispersão (γ) e inflacionado de zeros (ζ)

Para o ajuste do modelo de médias quando este apresenta excessos de zeros, resulta, de certa forma, em uma subestimação da média. Isso ocorre pelo fato que o valor de n é ampliado, enquanto que a soma de valores é reduzida. Nos modelos de regressão quando não há uma ponderação do excesso de zeros no modelo, como é o caso da distribuição de Poisson, nota-se uma subestimação da curva estimada distanciando levemente da distribuição ZICMP (Figura 2.3).

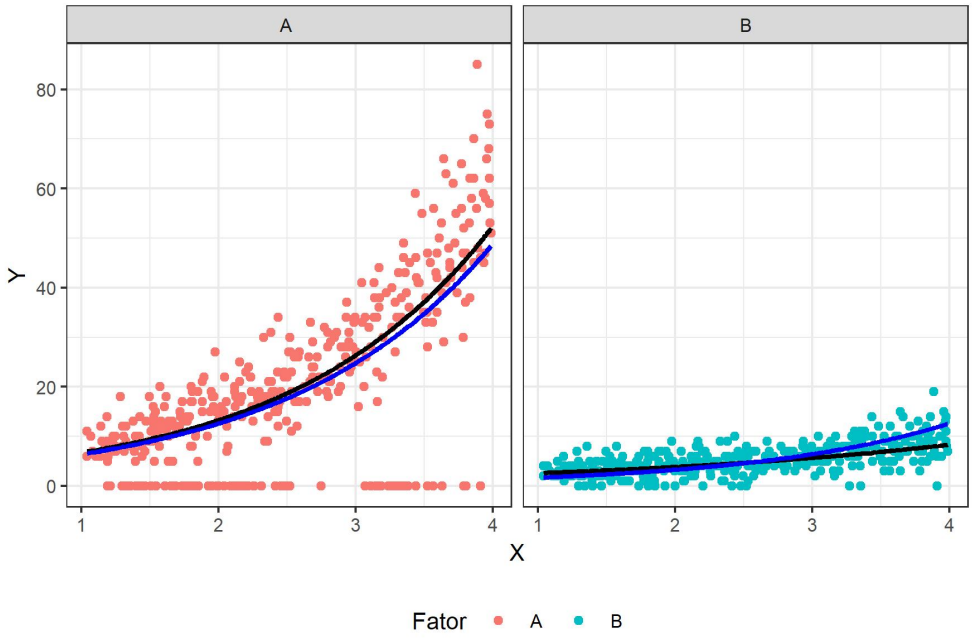


Figura 2.3. Ajuste do modelo Poisson(linha azul) e COMPoisson (linha preta)

Apesar do Modelo ZICMP possuir um melhor ajuste comparado ao Poisson, faz-se necessário apresentar algumas críticas; nesse caso referente às ponderações do modelo COMPoisson descrito para cada y_i da Equação (2.5). Ao assumir o valor de p_i elevado, essa ponderação refletirá em uma subestimação da distribuição COMPoisson, expressa na referida equação. Na Tabela 2.1 têm-se algumas estimativas que evidenciam essas diferenças nos modelos.

Tabela 2.1. Estimativas dos parâmetros para os Modelos Poisson e ZICMP

Modelo ZICMP	Estimativa	Erro Padrão	Lim.Inferior	Lim. Superior
β_0	0,9488	0,0685	0,8145	1,0831
β_1	0,4990	0,0296	0,4409	0,5570
γ_A	-0,3262	0,0575	-0,4389	-0,2134
γ_B	0,2666	0,0560	0,1569	0,3763
ζ	-2,0280	0,1109	-2,2453	-1,8106
Loglik	-2.097,2527	-	-	-
AIC	4.204,5054	-	-	-
BIC	4.227,9285	-	-	-
Modelo Poisson	**Estimativa**	**Erro padrão**	**lower**	**upper**
β_0	1,1778	0,0383	1,1027	1,2529
β_1	-1,3524	0,0244	-1,4002	-1,3046
β_2	0,6778	0,0125	0,6533	0,7023
LogLik	-3.470,1391	-	-	-
AIC	6.946,2782	-	-	-
BIC	6.960,3320	-	-	-

Observa-se na Figura 2.4 que as estimativas para o modelo ZICMP são significativas ao nível de 5% de significância, indicando, nesse caso, evidências de subdispersão e a inflação de zeros, isto é, o modelo é indicado visto que foi significativo o efeito desses fenômenos.

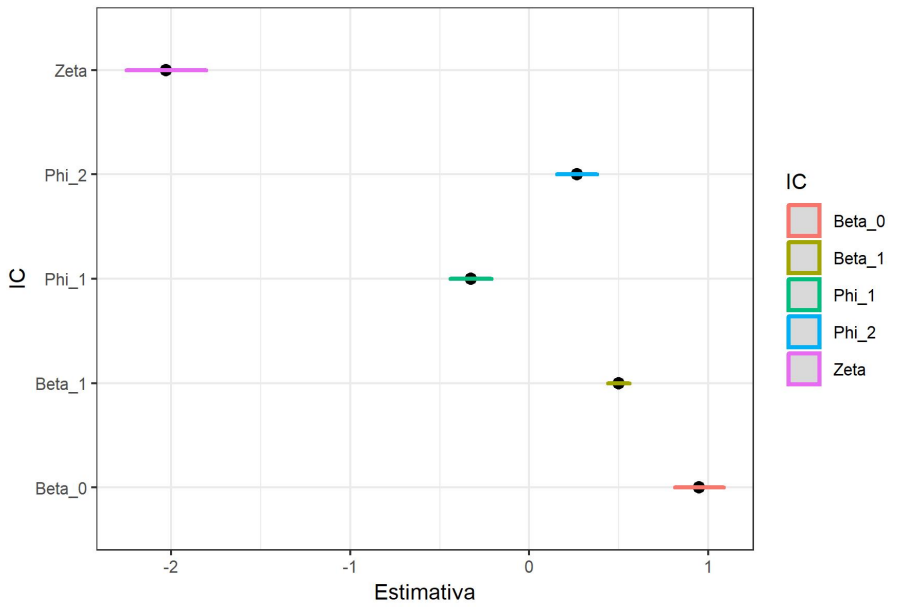


Figura 2.4. Intervalo de confiança com 5% de significância para as estimativas dos parâmetros do modelo ZICMP

Ao analisar dados reais foi verificada a presença de apenas um parâmetro de dispersão ϕ constante para todas as fazendas, porém, tal resultado não foi satisfatório, sendo necessário a especificação do parâmetro de dispersão para cada fazenda e a especificação de um preditor linear para a modelagem do excesso de zeros. Tais técnicas foram necessárias, pois os dados em análise apresentam um comportamento distinto de dispersão, associadas aos diferentes tipos de fazendas.

As estimativas dos parâmetros são apresentadas na Tabela 2.2, na qual foram consideradas a presença de heterogeneidade de dispersões, que têm as seguintes estimativas: $\hat{\phi_1} = 8.16$ $\hat{\phi_2} = 9.44$, $\hat{\phi_3} = 7.74$, $\hat{\phi_4} = 7.52$, $\hat{\phi_5} = 6.58$. Nesse caso, observa-se também que as fazendas apresentaram a subdispersão nos dados para o ajuste do modelo COMPoisson, captando a variabilidade extra não explicada pela distribuição Poisson.

Nos valores descritivos da log-verossimilhança para o preditor linear (2.7), nota-se que o modelo Poisson indicou irregularidades na validação do ajuste, em discordância com o COM-Poisson que mostrou ser satisfatório, uma vez que o modelo se adaptou à dispersão dos dados. Para avaliação do parâmetro γ, o uso do efeito de dispersão por fazenda pode ser justificado por haver variações entre os locais, a exemplo de clima e técnicas de manejo. Além disso, nota-se que os valores do AIC para o modelo considerando o excesso de zeros foi menor quando comparado com o de Poisson, corroborando nesse caso, a escolha do modelo COMPoisson.

Tabela 2.2. Estimativas dos parâmetros para os Modelos COMPoisson e Poisson

Parâmetros	Modelo COMPoisson			Modelo Poisson		
	Estimativa/SE	LI	LS	Estimativa/SE	LI	LS
β_0	0,4882 (0,3163)	-0,1317	1,1081	-1,0003 (0,0669)	-1,1314	-0,8692
β_{12}	2,1579 (0,5017)	1,1746	3,1413	0,1045 (0,0580)	-0,0092	0,2182
β_{13}	0,148 (0,4311)	-0,6969	0,9928	0,1439 (0,0593)	0,0277	0,2601
β_{14}	-0,9278 (0,4855)	-1,8793	0,0237	-0,0187 (0,0650)	-0,1461	0,1087
β_{15}	-1,3299 (0,5493)	-2,4066	-0,2532	0,1801 (0,0773)	0,0286	0,3316
β_2	1,7561 (0,084)	1,5916	1,9207	0,3566 (0,0084)	0,3401	0,3731
γ_{11}	2,0993 (0,0545)	1,9925	2,2062	-	-	-
γ_{12}	2,246 (0,0521)	2,144	2,348	-	-	-
γ_{13}	2,0467 (0,0545)	1,9398	2,1536	-	-	-
γ_{14}	2,0188 (0,0646)	1,8922	2,1454	-	-	-
γ_{15}	1,8836 (0,0694)	1,7475	2,0197	-	-	-
ζ_0	11,5978 (2,2295)	7,228	15,9676	-	-	-
ζ_2	-5,8513 (1,123)	-8,0523	-3,6503	-	-	-
LogLik	-1.134,2499	-	-	-1.982,7594	-	-
AIC	2.294,4999	-	-	3.977,5187	-	-
BIC	2.364,4675	-	-	4.009,8114	-	-

Uma vez escolhido o modelo, como evidenciado na Equação (2.7), faz-se necessário mostrar a qualidade do ajuste como parte da validação. Assim, é possível observar na Figura 2.5, que o modelo COMPoisson apresenta um ajuste satisfatório, enquanto que o mesmo não pode ser observado para o ajuste do modelo Poisson, uma vez que apresentou uma inflação nos quantis teóricos. Tal característica pode ser explicada devido à variabilidade expressa pelos parâmetros de dispersões em cada fazenda e devido à penalidade do excesso de zeros presentes no modelo COMPoisson.

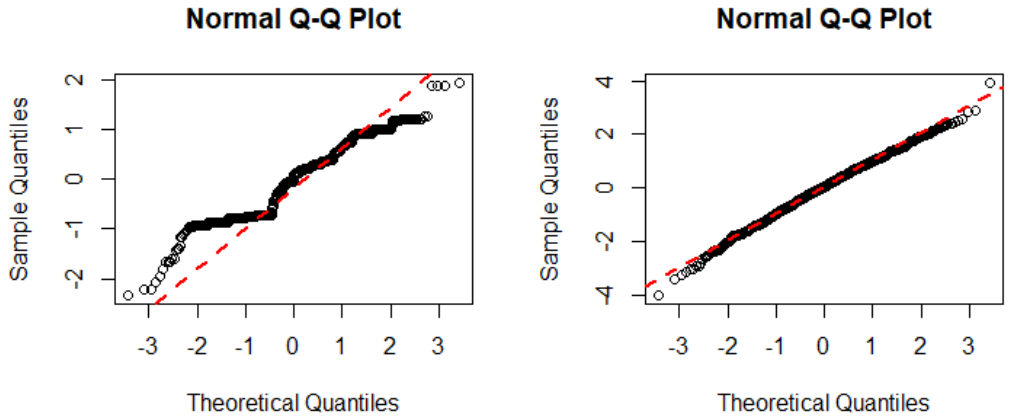


Figura 2.5. Análise de resíduos para os modelos Poisson (esquerda) e COMPoisson (direita)

Portanto, baseado no ajuste obtido por meio do modelo COMPoisson apresentado na Tabela 2.2, obteve-se as curvas dos comportamentos das fazendas em que todas apresentaram similaridades. Além disso, o modelo não apresentou divergências quanto à taxa de crescimento em relação ao número de lactações por idade das vacas; ainda, maiores valores foram encontrados tanto na fazenda Brejo quanto na fazenda IPE.

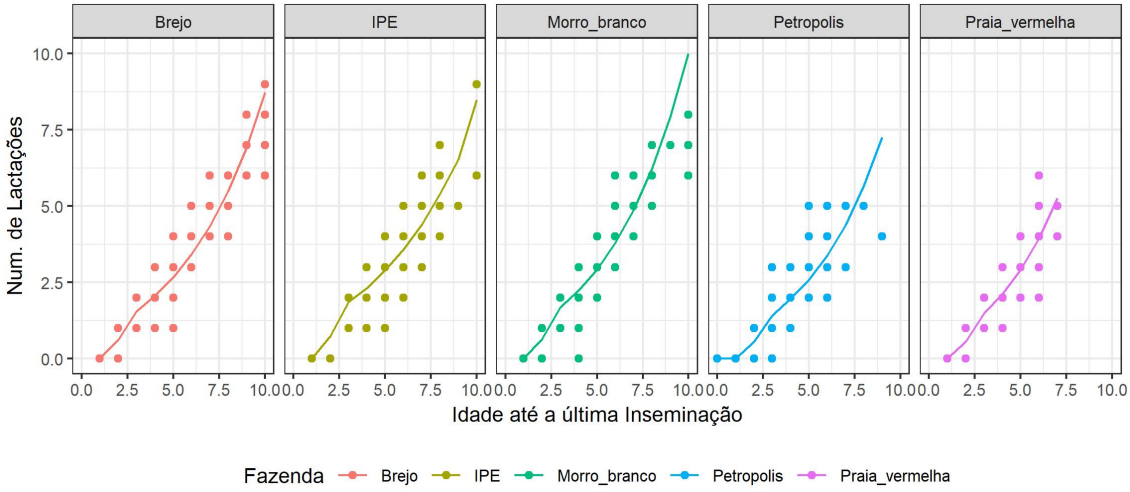


Figura 2.6. Gráfico do ajuste do modelo COMPoisson por fazenda

2.4 Considerações

Poderíamos discutir nesse trabalho várias possibilidades de extensões para a modelagem de dados de contagens, incluindo por exemplo, Poisson-Tweed, Poisson com penalidade por splines ou até mesmo métodos que podem incluir um efeito para cada parâmetro, como é o caso dos modelos GAMLSS. No entanto, preferimos discutir as distribuições Poisson e COMPoisson para mostrar a eficiência desses modelos envolvendo aplicações genéticas com dados apresentando excesso de zeros.

Por fim, concluímos que a utilização de dados quantitativos obtidos em estudos da área de Ciências Animais que apresentem características de superdispersão e subdispersão apresentaram validação com o uso do modelo COMPoisson. O ajuste foi possível em função da distribuição ser capaz de captar os fenômenos de dispersão intrinsecamente associados a dados de contagem, incluindo a habilidade do modelo para dados com excesso de zeros.

Referências

BERGAMASCHI, M. A. C. M., MACHADO, R., e BARBOSA, R. T. (2010). Eficiência reprodutiva das vacas leiteiras. *Embrapa Pecuária Sudeste-Circular Técnica (INFOTECA-E)*.

Conway, R. W. e Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136.

Mion, T. D., Daroz, R. Q., Jorge, M. J. A., MORAIS, J. P. G. d., e Gameiro, A. H. (2012). Indicadores zootécnicos e econômicos para pequenas propriedades leiteiras que adotam os princípios do projeto balde cheio. *Informações Econômicas, SP*, 42(5).

Nelder, J. e Wedderburn, R. (1972a). Generalized Linear Models. *Journal of the Royal Statistical Society A*, 135:370–384.

Nelder, J. A. e Wedderburn, R. W. M. (1972b). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.

Roche, J. F. (2006). The effect of nutritional management of the dairy cow on reproductive efficiency. *Animal reproduction science*, 96(3-4):282–296.

Sellers, K. F. e Raim, A. (2016). A flexible zero-inflated model to address data dispersion. *Comput. Stat. Data Anal.*, 99(C):68–80.

Sellers, K. F. e Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, 4(2):943–961.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., e Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Vaz, R. Z., Lobato, J. F. P., e Restle, J. (2010). Productivity and efficiency of cow herds submitted to two weaning ages. *Revista Brasileira de Zootecnia*, 39(8):1849–1856.

3 APLICAÇÃO DE MODELOS DE CONTAGENS NO ESTUDO DE INSEMINAÇÕES ARTIFICIAIS EM BOVINOS

Resumo

A modelagem de dados de contagem é geralmente complexa devido à necessidade em especificar novas variáveis ao modelo, funções de ligações e captar a variação extra dos dados. Assim, a insuficiência dos modelos que não consideram tal característica pode ser solucionada com a aplicação de técnicas estatísticas como dos Modelos Lineares Generalizados Mistos, visto que essa metodologia viabiliza uma maior flexibilidade para a variável resposta, ao passo que conseguem captar estruturas de correlação entre indivíduos, por meio da superdispersão existente nos dados. Nesse trabalho utilizou-se uma aplicação em dados de inseminação artificial e como variável de interesse, o número de prenhez. Para tal ajuste, foi considerado o modelo aditivo generalizado de escala e forma Delport, uma vez que esse apresentou resultados satisfatórios dentre os modelos estudados. Além disso, em virtude dos resultados encontrados pode-se dizer que houve efeito das diferentes fazendas estudadas, estágios de vida da vaca e estação do ano na condição de prenhez.

Palavras-chave: Variação Extra Poisson; Modelo Misto; GAMLSS.

3.1 Introdução

Na busca de solucionar problemas práticos junto a modelagem de dados e à validação dos mesmos, a estatística está evoluindo para suprir as diversas necessidades dos pesquisadores por meio de novas técnicas e implementações computacionais mais robustas. Os Modelos Lineares Gaussianos, por exemplo, tiveram origem no século XIX com Galton, e ainda são uma das técnicas mais utilizadas para averiguação e ajuste de um modelo de regressão linear, na qual a variável de interesse é modelada em relação à média da distribuição em função das variáveis explicativas, além de considerar a variância constante. No entanto, dados medidos na forma de proporções ou contagens não conseguem ter um ajuste satisfatório por meio dos modelos de regressão Gaussianos, uma vez que uma das pressuposições para tal modelagem é que a variável resposta tenha distribuição normal (Draper e Smith, 1981; Hoffmann, 2006).

Propostos por Nelder e Wedderburn (1972), os Modelos Lineares Generalizados (MLGs) permitem relacionar a variável de interesse, seja ela contínua ou discreta, com uma distribuição pertencente a família exponencial canônica, viabilizando uma maior flexibilidade quanto ao ajuste do modelo, por meio da associação de um componente sistemático a um preditor linear, ligados por uma função de ligação.

Uma extensão dos MLGs foi proposta pelos autores Hastie e Tibshirani (1990), em que incorporam uma função de suavização ao seu preditor linear, caracterizando os Modelos Aditivos Generalizados (GAMs). Mais tarde, seriam desenvolvidos os modelos aditivos generalizados de locação e escala (GAMLSS), os quais tratam-se de uma regressão semi-paramétrica, ou seja, assume-se uma mistura entre a distribuição paramétrica e uma função de suavização das variáveis explicativas, sendo uma combinação dos MLGs com os GAMs (Stasinopoulos et al., 2005).

Nas próximas seções são abordados as principais distribuições para a análise de dados de contagem em substituição aos modelos gaussianos. Além disso, propõe-se a utilização dos modelos Delaporte (DEL), pertencente ao GAMLSS, como uma alternativa para os modelos lineares generalizados e suas extensões, com o objetivo de resolver problemas relacionados às pesquisas realizadas na zootecnia, especificamente em estudos de avaliações com respostas obtidas por meio das técnicas de biotecnologias reprodutivas, cuja variável resposta foi o número de vacas prenhes.

3.2 Material e Métodos

A modelagem estatística, associada a modelos de regressão, é geralmente empregada com base na distribuição gaussiana, mas somente a partir de trabalhos como os de Nelder e Wedderburn (1972) foi possível obter o ajuste de diferentes modelos com base em distribuições distintas, que fazem parte da família exponencial de distribuição. Tal característica pôde auxiliar os pesquisadores a definir qual distribuição mais conveniente ao ajuste do modelo, com base no tipo de variável resposta a ser empregada. Em dados de contagem, por exemplo, o uso de modelos gaussianos pode resultar em um mal ajuste apresentado na análise de resíduos. Assim, a utilização de outras distribuições, associadas à contagem são empregadas na literatura para resolver tais inconsistências presentes na modelagem clássica.

Com o objetivo de explicar o comportamento do número de prenhez em quatro fazendas do estado de Minas Gerais, Brasil optou-se em trabalhar com modelos generalizados aditivos de locação, forma e escala, também conhecidos como GAMMLSS. Tais modelos podem ser associados às famílias de distribuições que possuem um ou mais parâmetros que controlam efeitos de assimetria e de dispersão dos dados, fato este presente na modelagem de dados de contagem.

Segundo Stasinopoulos et al. (2017), um modelo GAMMLSS pode ser representado na forma matricial por

$$
\begin{aligned}
&Y|\boldsymbol{\gamma} \overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
&\eta_1 = g_1(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z}_{11}\boldsymbol{\gamma}_{11} + ... + \mathbf{Z}_{1J_1}\boldsymbol{\gamma}_{1J_1} \\
&\eta_2 = g_2(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_{21}\boldsymbol{\gamma}_{21} + ... + \mathbf{Z}_{2J_2}\boldsymbol{\gamma}_{2J_2} \\
&\eta_3 = g_3(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{Z}_{31}\boldsymbol{\gamma}_{31} + ... + \mathbf{Z}_{3J_3}\boldsymbol{\gamma}_{3J_3} \\
&\eta_4 = g_4(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + \mathbf{Z}_{41}\boldsymbol{\gamma}_{41} + ... + \mathbf{Z}_{4J_4}\boldsymbol{\gamma}_{4J_4}
\end{aligned}
$$

em que $\mathbf{X}_i$ são matrizes de delineamento do modelo associadas aos parâmetros fixos $\boldsymbol{\beta}_i$, com $i = 1, .., n$, enquanto $\mathbf{Z}_{ij}$, são matrizes que pré-multiplicam os parâmetros de efeito aleatório $\boldsymbol{\gamma}_{ij}$.

A estimação dos modelos é feita com base no método na máxima verossimilhança quando os mesmos apresentam apenas parâmetros fixos β_i, ou seja,

$$
l = \sum_{i=1}^{n} \log f\left(y_i|\mu_i, \sigma_i, \nu_i, \tau_i\right),
$$

na presença de efeitos aleatórios $\boldsymbol{\gamma}_{ij}$, o método da máxima verossimilhança penalizada será

$$
l_p = l - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}'\mathbf{G}_{kj}(\lambda_{kj})\boldsymbol{\gamma}_{kj},
$$

sendo $\mathbf{G}_{kj}$ a matriz de precisão dos efeitos aleatórios, com

$$\boldsymbol{\gamma}_{kj} \sim N\left(\mathbf{0}, [\mathbf{G}_{kj}(\boldsymbol{\lambda}_{kj})]^{-1}\right).$$

Na comparação entre modelos, foi utilizado o $GAIC = -2\hat{l} + k \cdot df$ como critério para a escolha do melhor modelo, uma vez que, quanto menor o valor do GAIC, melhor será a indicação do modelo proposto. Assim, seja $\hat{l}$ a função log-verossimilhança do modelo, em que k é a constante penalizadora e o df são os graus de liberdade do modelo proposto; quando $k = 2$ tem-se o AIC e enquanto $k = log(n)$ tem-se o SBC.

A seguir são apresentadas as principais distribuições discretas associadas à modelagem de dados de contagem e algumas delas possuem parâmetros adicionais para captação da assimetria dos dados, como também parâmetros que explicam a variabilidade extra causada pelos fenômenos de sub ou superdisperção.

Modelo de Poisson (PO)

Dos modelos lineares generalizados, a distribuição de Poisson é usualmente utilizada na modelagem estatística de dados de contagem para representar a variável aleatória discreta sob a forma $Y_i \sim Poisson(\mu)$, com parâmetro desconhecido $\mu > 0$, cuja função de densidade de probabilidade é denotada por

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}, \tag{3.1}$$

com $E(Y) = Var(Y) = \mu$.

A distribuição de Poisson é comumente utilizada na estatística no cálculo de probabilidades de ocorrências de um determinado evento, como também utilizada nos modelos lineares generalizados para o ajuste de modelos de regressão.

Double Poisson (DPO)

A distribuição Double Poisson possui uma caraterística importante que é a captação do efeito de sub ou superdispersão dos dados através do parâmetro σ. Assim, Rigby et al. (2017) denotam a função de distribuição de probabilidade da DPO(μ, σ) por

$$P(Y = y|\mu, \sigma) = c(\mu, \sigma)\sigma^{-\frac{1}{2}}e^{-\frac{\mu}{\sigma}}\left(\frac{\mu}{y}\right)^{\frac{y}{\sigma}}\frac{e^{\frac{y}{\sigma}-y}y^y}{y!}, \tag{3.2}$$

com

$$c(\mu, \sigma) = \left[\sum\nolimits_{y=0}^{\infty}\sigma^{-\frac{1}{2}}e^{-\frac{\mu}{\sigma}}\left(\frac{\mu}{\sigma}\right)^{\frac{y}{\sigma}}\frac{e^{\frac{y}{\sigma}-y}y^y}{y!}\right]^{-1}, \tag{3.3}$$

uma vez que $y = 0, 1, 2, \ldots$, $\mu > 0$ e $0 < \sigma < \infty$, sendo $E(Y) \approx \mu$ e $Var(Y) \approx \sigma\mu$.

Em casos especiais, quando na DPO apresenta parâmetro $\sigma = 1$, por exemplo, a distribuição DPO será similar a distribuição de Poisson. Para $\sigma > 1$, a presença de superdispersão dos dados será considerada, ou seja, uma variação extra que não pode ser explicada apenas pela distribuição de Poisson, por outro lado, quando $\sigma < 1$, há presença de subdispersão dos dados.

Distribuição Poisson-normal-inversa (PIG)

A distribuição de Poisson-normal-inversa é descrita em Rigby e Stasinopoulos (2009), como uma combinação entre as distribuições Poisson e Gaussiana inversa. Diferentemente da distribuição binomial negativa, a PIG viabiliza o ajuste dos modelos na presença de assimetria, pois possui uma cauda superior mais pesada, sendo descrita por:

$$P(Y|\mu,\sigma) = \left(\frac{2\alpha}{\pi}\right)^{\frac{1}{2}} \frac{\mu^y e^{\frac{1}{\sigma}} K_{y-\frac{1}{2}}(\alpha)}{(\alpha\sigma)^y y!}, \tag{3.4}$$

com $\alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma}$, $\mu > 0$ e $\sigma > 0$. Sendo $K_\lambda(t) = \left(\frac{1}{2}\right) \int_0^\infty x^{\lambda-1} \exp\left[-\frac{1}{2}t\left(x + \frac{1}{x}\right)\right] dx$ a função modificada de Bessel de terceira forma. Além disso, tem-se que a média da distribuição será $E[Y] = \mu$ e a variância é dada por $Var[Y] = \mu + \sigma\mu^2$.

Distribuição Delaporte (DEL)

De acordo com Rigby e Stasinopoulos (2009), a Delaporte é resultante da mistura entre as distribuições de Poisson, obtida com a marginal de Y, em que $Y|\gamma \sim \text{Poisson}(\mu, \gamma)$ com $\gamma \sim \text{SG}(1, \sigma^{\frac{1}{2}}, \nu)$, e uma combinação da distribuição gama deslocada com a função densidade de probabilidade expressa por

$$f_\gamma(\gamma) = \frac{(\gamma-\nu)^{\frac{1}{\sigma}-1}}{\sigma^{1/\sigma}(1-\nu)^{1/\sigma}\Gamma(1/\sigma)} \exp\left[-\frac{(\gamma-\nu)}{\sigma(1-\nu)}\right], \tag{3.5}$$

em que $\gamma > \nu$, onde $\sigma > 0$ e $0 \leq \nu \leq 1$. A parametrização apresentada garante que $E[\gamma] = 1$, o que implicará em $E[Y] = \mu$. Observa-se que $\gamma = \nu + (1-\nu)Z$, sendo $Z \sim \text{GA}(1, \sigma^{\frac{1}{2}})$, então γ terá um limite inferior igual a ν.

Nesse sentido, a distribuição de Delaporte possui três parametrizações, o parâmetro referente a média da distribuição μ, o parâmetro de escala, referente a dispersão σ e finalmente o parâmetro de forma ν. Assim, a função densidade de probabilidade correspondente a $\text{DEL}(\mu, \sigma, \nu)$ será denotada por

$$P(Y = y|\mu,\sigma,\nu) = \frac{\exp^{-\mu\nu}}{\Gamma(1/\sigma)}[1 + \mu\sigma(1-\nu)]^{-1/\sigma} S, \tag{3.6}$$

com

$$S = \sum_{j=0}^{y} \begin{pmatrix} y \\ j \end{pmatrix} \frac{\mu^y \nu^{y-j}}{y!} \left[\mu + \frac{1}{\sigma(1-\nu)}\right]^{-j} \Gamma\left(\frac{1}{\sigma} + j\right), \tag{3.7}$$

para $y = 0, 1, 2, \ldots$, em que $\mu > 0$, $\sigma > 0$, e $0 \leq \nu \leq 1$. Alem disso, tem-se que $E[Y] = \mu$ e a variância é dada por $Var[Y] = \mu + \mu^2\sigma(1-\nu)^2$. Em particular, a Delaporte pode ser estendida para famílias de reparametrizações ao utilizar um parâmetro extra de forma similar ao da binomial negativa tipo I (Rigby et al., 2017).

Distribuição de SICHEL

A distribuição de Sichel é uma distribuição de probabilidade resultante da mistura de distribuições Poisson e Gaussiana Inversa Generalizada, sendo utilizada no ajuste de modelos de Poisson com a presença de superdispersão e assimetria (Rigby e Stasinopoulos, 2009; Tzougas,

2014). Com duas parametrizações, a distribuição de Sichel possui o parâmetro de média μ e dois parâmetros adicionais, σ e ν, que definem a escala e a forma dessa distribuição, respectivamente.

A distribuição de Sichel - **SI(μ,σ,ν)** é denotada em Rigby e Stasinopoulos (2009) por

$$P(Y|\mu,\sigma,\nu) = \frac{\mu^y K_{Yy+\nu}(\alpha)}{y!(\alpha\sigma)^{y+\nu} K_\nu\left(\frac{1}{\sigma}\right)}, \tag{3.8}$$

com $\alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma}$ e $\mu > 0,\ \sigma > 0$ e $-\infty < \nu < \infty$.

A função K_λ, trata-se da modificada de Bessel na terceira forma

$$K_\lambda(t) = \left(\frac{1}{2}\right)\int_0^\infty x^{\lambda-1}\exp\left[-\frac{1}{2}t\left(x+\frac{1}{x}\right)\right]dx \tag{3.9}$$

em que

$$\sigma = \left[\left(\mu^2+\alpha^2\right)^{\frac{1}{2}} - \mu\right]^{-1}. \tag{3.10}$$

Por sua vez, a segunda parametrização da distribuição de Sichel, especificada por **SICHEL(μ,σ,ν)**, pode ser expressa por

$$P(Y|\mu,\sigma,\nu) = \frac{(\mu/c)^y K_{y+\nu}(\alpha)}{y!(\alpha\sigma)^{y+\nu} K_\nu\left(\frac{1}{\sigma}\right)}, \tag{3.11}$$

com $\alpha^2 = \frac{1}{\sigma^2} + \frac{2(\mu/c)}{\sigma}$, $c = \frac{K_{\lambda+1}\left(\frac{1}{\sigma}\right)}{K_\lambda\left(\frac{1}{\sigma}\right)}$, para $\mu > 0,\ \sigma > 0$ e $-\infty < \nu < \infty$. Assim, tem-se a $E(Y) = \mu$ e $Var(Y) = \mu + \mu^2\left[\frac{2\sigma(\nu+1)}{c} + \frac{1}{c^2} - 1\right]$.

3.2.1 Dados Reais

Nesse trabalho foi apresentado uma análise estatística com base nas observações resultantes de estudos por amostragem, durante o ano de 2018, realizados em 4 fazendas: 1 - Morro Grande (130 cruzamentos), 2 - Ipê (474 cruzamentos), 3 - Petrópolis (330 cruzamentos) e 4 - Praia Vermelha (31 cruzamentos), totalizando 965 cruzamentos da raça holandesa, na condição de prenhez, em que todas as fazendas estão localizadas na região de Passos, Minas Gerais, Brasil.

Os animais do plantel foram selecionados de acordo com os valores genéticos, os quais foram submetidos ao sistema de criação *free stall*, isto é, com baias individuais, área de alimentação, água de qualidade, limpeza periódica e sistema de refrigeração. Além disso, as fêmeas que passaram pelo protocolo reprodutivo após 7 dias de detecção estral são as mesmas que mantiveram e conceberam os embriões. Nessa estrutura foram avaliadas as variáveis: produtividade leiteira (L.Maior - em quilos), número de lactação (Lact: 1 - 8), idade das vacas até a última inseminação (Idade: 1 - 8), quantidade de inseminações artificiais (Ia: 1 - 11). Além disso, fatores como touro doador do sêmen, estação do ano em que ocorreu a última inseminação (1 - Primavera: setembro, outubro, novembro, 2 - Verão: dezembro, janeiro, fevereiro , 3 - Outono: março, abril, maio, 4 - Inverno: junho, julho, agosto), e estágios de vida da vaca (Estatus: 1 - novilha, 2 - em leite, 3 - seca).

Para análise descritiva e ajuste do modelo foram consideradas as medianas das variáveis quantitativas, exceto para a variável Leite Maior, que por ter uma distribuição simétrica foi empregada a média. Em síntese, tais valores foram tabulados considerando-se o aninhamento dos fatores fazenda, touro, estação da última inseminação e status da vaca, sendo o número de prenhez a variável resposta.

3.2.2 Modelagem

O conjunto de dados apresentados na subseção 3.2.1 está relacionado a dados de contagens. Nesse sentido, a modelagem estatística poderá propiciar um melhor entendimento quanto à resposta número de prenhez, em que a taxa de ocorrência não é constante, visto que essa característica irá depender tanto de fatores genéticos de cada indivíduo observado, quanto das condições ambientais presentes nas diferentes localidades. Assim, para o ajuste dos modelos de efeitos fixos foram propostas as distribuições Poisson, Doubler Poisson, Delaporte, Sichel e Poisson-nomal-inversa, sendo que essas três últimas distribuições possuem parâmetros adicionais que captam efeitos de dispersão e assimetria, utilizando-se nesses casos os preditores lineares $g(\sigma) = \sigma$ e $g(\nu) = \nu$. Dessa forma, baseados nos modelos tem-se o preditor linear dado por

$$g(\mu) = \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3k} + \beta_4 \text{Ia} + \beta_5 \text{L.Maior} + \beta_6 \text{Lact}, \tag{3.12}$$

em que β_0 é a constante geral, β_{1i} é o efeito da i-ésima fazenda, β_{2j} é o j-ésimo efeito da estação do ano em que ocorreu a última inseminação, β_{3k} é o efeito do k-ésimo estágio de vida da vaca, β_4, β_5 e β_6 estão associados as variáveis quantitativas, número de inseminações artificiais, produtividade leiteira e número de lactações, respectivamente.

Nos casos dos modelos que consideram a distribuição de Poisson, tem-se que o parâmetro de dispersão fixo associado ao ajuste é igual a 1 e que para um modelo bem ajustado é esperado que o desvio residual esteja próximo dos graus de liberdade do resíduo, isto é, o valor esperado da distribuição de referência (χ^2). Porém, isso geralmente não ocorre devido à presença da superdispersão, variação nos dados muito maior que a acomodada pelo modelo ou na presença de subdispersão, variância muito menor que a especificada pela distribuição assumida. Possíveis causas desses fenômenos seriam a variabilidade do material experimental, correlações entre respostas individuais, agregação (*clusters*) e variáveis omitidas no preditor linear (Hinde e Demétrio, 1998a).

Dessa forma, para tentar captar as diferentes estruturas presentes no conjunto de dados, foi feita a inclusão do efeito aleatório de touro no preditor linear, pois considerou-se que esses indivíduos representam uma amostra de todos os possíveis touros doadores de sêmen utilizados nas fazendas. Por conseguinte, tem-se para o ajuste dos modelos associados as distribuições PO, DPO, PIG, DEL e SICHEL apresentarão o seguinte preditor linear

$$g(\mu) = \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3k} + \beta_4 \text{Ia} + \beta_5 \text{L.Maior} + \beta_6 \text{Lact} + \xi_{\text{t}}, \tag{3.13}$$

cujos parâmetros dos efeitos fixos foram definidos no modelo dado em (3.12), e ξ_t é o efeito aleatório a nível de touro, com $\xi_t \sim \text{N}(0, \sigma^2_{\text{t}})$.

Para o cálculo das estimativas dos parâmetros dos modelos sugeridos nessa subseção, foram utilizados o método da máxima verossimilhança, implementada no pacote gamlss (Rigby e Stasinopoulos, 2005). As comparações entre os ajustes foi realizada por meio do critério da informação de *Akaike* generalizada (GAIC).

3.3 Resultados e Discussões

Com o objetivo de multiplicar o número de descendentes de um animal com alto valor genético, as técnicas de superovulação consistem em submeter uma fêmea a um tratamento hor-

monal, com a finalidade de produzir mais oócitos em um único estro. Após a superovulação da vaca, essa passa pelo procedimento de inseminação artificial em tempo fixo (IATF), que é considerada uma biotecnologia reprodutiva capaz de promover o aumento dos índices reprodutivos, desempenhando efeito significativo sobre a eficiência bioeconômica nas propriedades com atividades pecuárias (Goncalves et al., 2001). O aumento do número de descendentes obtidos por meio das biotecnologias viabiliza melhores definições quanto aos padrões genéticos dos animais e promove avaliações das diferentes progênies (Carvalho et al., 2019).

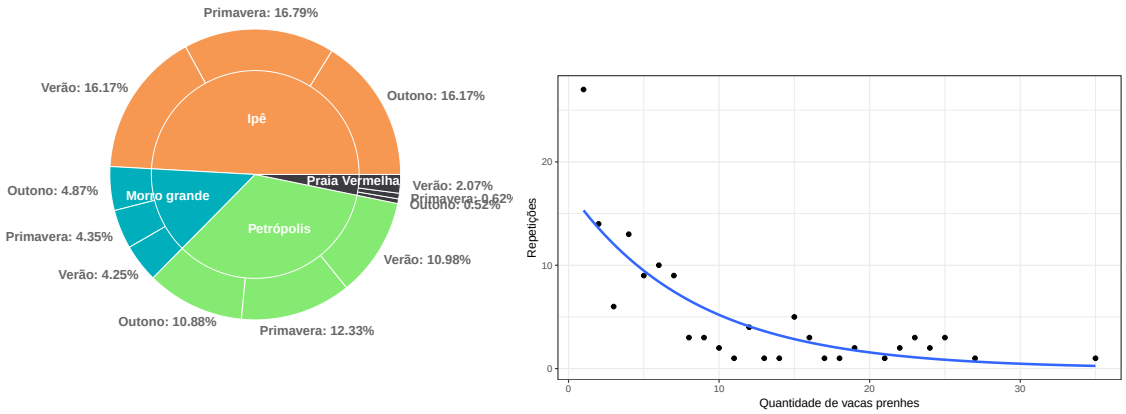
Como abordado na subseção 3.2.1, das 1483 vacas registradas no protocolo reprodutivo, 965 mostraram resposta de prenhez. Assim, na Figura 3.1a tem-se uma representação gráfica referente ao número de vacas prenhes, considerando cada fazenda por estações do ano.

Observa-se que os rebanhos das propriedades Ipê e Petrópolis estão estabilizados, ou seja, com número equilibrado de animais em processo de lactação. Ademais, devido à manutenção contínua de leite, os produtores realizam planejamentos das parições entre as estações (primavera/verão/outono), visto que nessas épocas não ocorrem grandes variações na proporcionalidade do número de prenhez. O mesmo comportamento pode ser notado na fazenda Morro Grande, porém esta ainda encontra-se em processo de estabilização de rebanho, sendo, aproximadamente, 15% do plantel em lactação. Por outro lado, a propriedade Praia Vermelha, expôs baixas quantidades de fêmeas nessa condição de prenhez, provavelmente pela elevada renovação do rebanho, ocasionando uma redução do plantel de vacas adultas e aumentando o número de fêmeas novilhas, causando uma sub-estimação das médias no que diz respeito ao número de vacas prenhes na fazenda.

A ausência de observações na estação do inverno é justificada por ser um período que apresenta menores ofertas de alimentos com boa qualidade. Além disso, elevadas pressões climáticas podem intensificar o surgimento de doenças respiratórias principalmente nas crias.

Ao analisar apenas a variável resposta, observa-se que houve cruzamentos que apresentaram mais de uma repetição; isso ocorre devido ao interesse da fazenda em multiplicar os animais de maior interesse, seja por motivos comerciais ou reprodutivos (Figura 3.1b). Em geral, a taxa média de serviços (número de inseminações realizadas) é acima de três. Assim, elevadas repetições de IA, devem ser analisadas com cautela, uma vez que animais com predisposição à problemas de eficiência reprodutiva devem ser descartados do sistema.

No estudo trabalhou-se com 3 categorias de animais, tais como, novilhas (vacas que estavam em reprodução pela primeira vez); em leite (animais que já pariram e encontram-se em estágio de lactação com até 6 meses de prenhez); vacas secas (fêmeas que estão no período final de gestação, aproximadamente, nos 3 últimos meses e que foram retiradas do processo de lactação para descanso do aparelho mamário, viabilizando o animal receber nova gestação).



(a) (b)

Figura 3.1. Dados de IA: a) Percentuais de vacas prenhes observados em três estações do ano (Verão/primavera/outono), por propriedade avaliada; e b) Dispersão da quantidade de vacas prenhes por número de repetições de cada cruzamento

Na Figura 3.2A, tem-se que as fêmeas que estavam em lactação, aparentemente, apresentaram melhores desempenhos quanto aos resultados de prenhez, uma vez que essa categoria apresentou maiores valores médios de prenhez. No entanto, ao comparar as vacas lactantes com as secas, nota-se que não houve um bom desempenho dessa última ao estímulo reprodutivo.

Embora as fazendas empreguem os mesmos protocolos reprodutivos, encontra-se na Figura 3.2B uma variabilidade entre os comportamentos dessas que sugestionou maiores resultados de prenhez nas fazendas Ipê e Petrópolis. Pode-se verificar na Figura 3.2C que, ao realizar inseminações artificiais em vacas secas e lactantes na época do verão, não obtêve-se bons resultados. De acordo com Sartori et al. (2002), vacas em lactação têm um aumento na temperatura corporal em resposta ao estresse térmico, principalmente na época do verão.

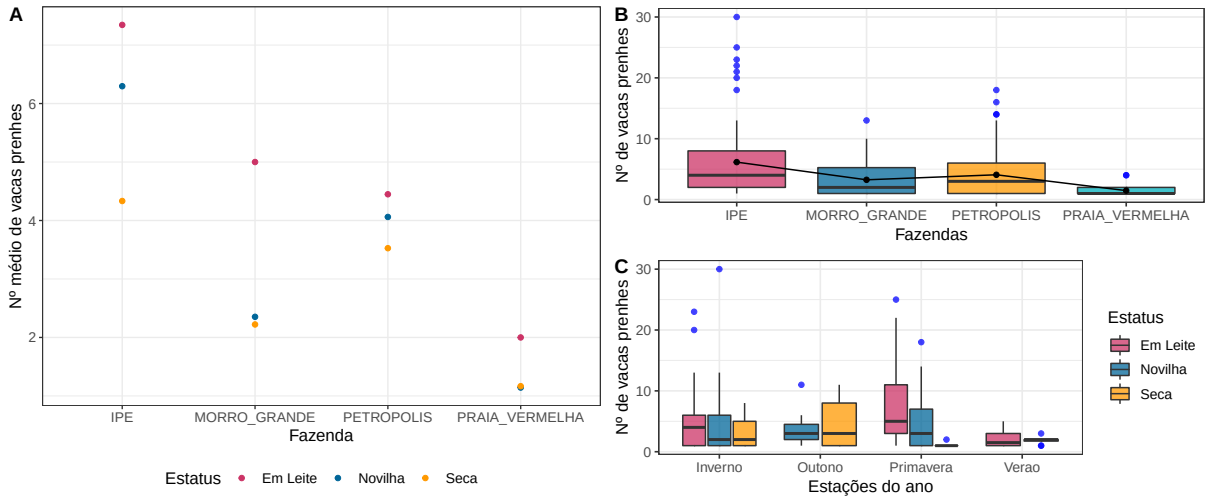


Figura 3.2. Dados de IA: (A); Número médio de vacas prenhes, por fazenda, considerando os estágios de vida das vacas receptoras; (B) Boxplot do número de vacas prenhes por fazenda; (C) Boxplot do número de vacas prenhes por estações do ano e estágios de vida da vaca receptora. Pontos em azul evidenciam valores discrepantes

As interpretações acerca dos coeficientes de correlação entre as variáveis observadas devem ser cautelosas, uma vez que esses valores podem fornecer números que não condizem com a realidade. De acordo com Gujarati e Porter (2011), essas medidas de associação entre duas variáveis são quantificadas na presença de um ou mais fatores, o que acarretará influências nos coeficientes. Um coeficiente mais confiável seria a correlação parcial, pois essa associação elimina todas as possíveis influências externas, informando de fato se há dependência. Desse modo, na Figura 3.3A têm-se as correlações parciais entre as covariáveis quantitativas em estudo. Observou-se também uma relação negativa e significativa entre os números de lactações e inseminações artificiais, ou seja, à medida que há mais vacas lactantes, as fazendas diminuíram o número de inseminações. Uma relação esperada seria a associação positiva e significativa entre as variáveis Leite Maior e número de lactações, pois quando a vaca está em condição de prenhez essa terá um aumento quanto à sua produção leiteira (Figura 3.3B). Para avaliar as possíveis influências nos resultados de prenhez em vacas holandesas, foi inicialmente ajustado o modelo Poisson apenas com as variáveis de efeito fixo (Subseção 3.2.2). De acordo com Ross (2009), a distribuição de Poisson tem a pressuposição da taxa de ocorrência constante. No entanto, a estimativa do parâmetro de dispersão $\hat{\phi} = 2,81$ foi muito maior que a esperada pelo modelo, o que evidencia superdispersão. Uma explicação plausível para a fenômeno ocorrido deve-se ao fato da falta de mesma probabilidade para que todas as vacas fiquem prenhes, pois cada indivíduo tem características genéticas distintas, além disso, fatores ambientais também podem ocasionar a variabilidade encontrada. Nesse contexto, têm-se os demais modelos de efeitos fixos, em que de acordo com a verificação da qualidade do ajuste nota-se que o modelo PIG foi o que apresentou menor estimativa do GAIC, sendo esse o considerado como mais adequado dentre essas classes de modelos (Tabela 3.1).

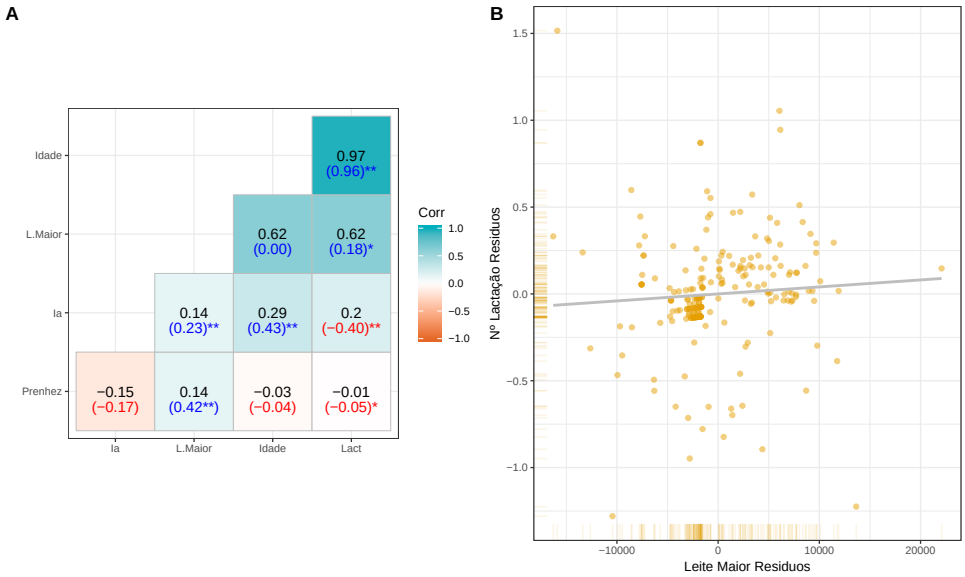


Figura 3.3. Dados de Inseminação Artificial - IA: A) Gráfico das correlações simples (em preto) e parciais (vermelho - correlação negativa, azul - correlação positiva); B) Gráfico dos resíduos entre Leite Maior e Número de Lactação, com uma reta de regressão linear

Tabela 3.1. Tabela resumo das estimativas dos parâmetros com os respectivos erros-padrão (SE), para os modelos apenas de efeitos fixos propostos

Parâmetros	PO (SE)	DEL (SE)	DPO (SE)	SICHEL (SE)	PIG (SE)
β_0	$1.8968\ (0.1574)^{***}$	$1.8927\ (0.2027)^{***}$	$1.8995\ (0.2799)^{***}$	$1.6543\ (0.2967)^{***}$	$1.6684\ (0.2734)^{***}$
β_{11}	$-0.6535\ (0.0993)^{***}$	$-0.3402\ (0.1286)^{**}$	$-0.6812\ (0.1827)^{***}$	$-0.5982\ (0.1818)^{**}$	$-0.6081\ (0.1650)^{***}$
β_{13}	$-0.3998\ (0.0790)^{***}$	$-0.1536\ (0.1227)$	$-0.4108\ (0.1109)^{***}$	$-0.3067\ (0.1375)^{*}$	$-0.3162\ (0.1380)^{*}$
β_{14}	$-1.1106\ (0.1983)^{***}$	$-0.4142\ (0.2313)$	$-1.2547\ (0.6127)^{*}$	$-0.9652\ (0.4167)^{*}$	$-0.9743\ (0.2857)^{***}$
β_{21}	$0.2424\ (0.0755)^{**}$	$0.2085\ (0.1141)$	$0.2514\ (0.1014)^{*}$	$0.1896\ (0.1314)$	$0.1931\ (0.1356)$
β_{22}	$-0.9025\ (0.1585)^{***}$	$-0.9120\ (0.1876)^{***}$	$-0.9928\ (0.5615)$	$-0.8250\ (0.4006)^{*}$	$-0.8190\ (0.2303)^{***}$
β_{23}	$0.1117\ (0.1004)$	$0.4190\ (0.1574)^{**}$	$0.1230\ (0.2014)$	$0.1979\ (0.2023)$	$0.1914\ (0.1694)$
β_{31}	$0.0010\ (0.1444)$		$0.0164\ (0.2753)$	$0.1033\ (0.2815)$	$0.1014\ (0.2486)$
β_{33}	$-0.5955\ (0.1004)^{***}$	$-0.7261\ (0.1259)^{***}$	$-0.6273\ (0.2034)^{**}$	$-0.5875\ (0.2055)^{**}$	$-0.5896\ (0.1706)^{***}$
β_4	$-0.0806\ (0.0188)^{***}$	$-0.0607\ (0.0235)^{*}$	$-0.0873\ (0.0371)^{*}$	$-0.0561\ (0.0350)$	$-0.0577\ (0.0305)$
β_5	$0.0000\ (0.0000)^{***}$	$0.0000\ (0.0000)^{***}$	$0.0000\ (0.0000)^{*}$	$0.0000\ (0.0000)^{*}$	$0.0000\ (0.0000)^{**}$
β_6	$-0.0457\ (0.0413)$	$-0.1769\ (0.0449)^{***}$	$-0.0484\ (0.0814)$	$-0.0159\ (0.0849)$	$-0.0179\ (0.0729)$
$log(\sigma_0)$		$-0.8861\ (0.5209)$	$0.9985\ (0.0828)^{***}$	$-0.7469\ (0.2179)^{***}$	$-0.7776\ (0.1828)^{***}$
ν		$-2.3746\ (2.2867)(*)$		$-1.2017\ (1.7964)$	
N.O	219	219	219	219	219
GAIC	1254.0791	1406.1955	1118.2114	1057.7249	1055.8701

Significâncias ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$); N.O (número de observações); $logit(\nu)$ em (*)

Devido às estruturas de correlações e diversas fontes de variações presentes nos dados, foram propostos os modelos de efeito aleatório no preditor linear a nível de touro (Tabela 3.2). A identificação da variabilidade presente nas fazendas estudadas serviu para modelar não apenas o parâmetro da média da distribuição assumida, mas também a dispersão. Nesse contexto, ao ajustar o modelo DPO, que considerou diferentes dispersões por fazenda, percebe-se que o ajuste apresentou um GAIC inferior aos demais modelos, indicando que por mais ajustado que esteja a dispersão, não é necessário a inclusão de mais parâmetros, uma vez que os demais ajustes sem esse efeito propiciaram melhores desempenhos no que tange a modelagem, visto que o critério da qualidade de ajuste foi melhor.

As avaliações dos modelos Sichel, PIG e Delaporte mostraram comportamentos similares quanto ao GAIC, contudo, o último foi o que apresentou menor valor. Adicionalmente, ao analisar os componentes de variância dos ajustes, observa-se que o Delaporte foi o que conseguiu captar maior variabilidade relacionada ao efeito aleatório de touro, isto é, proporcionou uma redução da variabilidade residual, que agora será decomposta em parte ao acaso e parte referente ao animal doador do sêmen.

Dessa forma, por mais que se tenha uma consistência quanto às significâncias dos efeitos estimados nos diferentes modelos, ao observar o *worm plot*, percebe-se que o ajuste do modelo Delaporte foi o mais indicado, visto que a sua adequação é verificada quando os pontos no gráfico se encontram em linha reta e a forma de "*worme*" não é detectada, indicando que os dados não diferem dos valores preditos pelo modelo sob a distribuição teórica (Figure 3.4).

Assim, conforme a avaliação do modelo escolhido, pode-se dizer que houve indícios de que o estágio de vida da vaca foi significativo no resultado de prenhez. A utilização de novilhas como gestoras no protocolo de superovulação é comum, uma vez que essa categoria apresenta uma taxa gestacional variando entre 10% e 23% (Hasler, 2014). Entretanto, os efeitos estimados mostraram que tanto as novilhas quanto as lactantes apresentaram desempenhos similares, ou

seja, assim como encontrado no trabalho de Demetrio et al. (2007), usar fêmeas novilhas ou lactantes se têm os mesmos resultados no que diz respeito ao número de prenhez. Por outro lado, nota-se que ao usar vacas secas houve uma redução na quantidade de vacas prenhes.

Tabela 3.2. Tabela resumo das estimativas dos parâmetros com os respectivos erros-padrão (SE), para os modelos com efeito aleatório propostos

Parâmetros	DPO (SE)	DEL (SE)	SICHEL (SE)	PIG (SE)
β_0	$1.8188\ (0.2398)^{***}$	$1.6655\ (0.2906)^{***}$	$1.6900\ (0.2887)^{***}$	$1.6684\ (0.2781)^{***}$
β_{11}	$-0.6134\ (0.1358)^{***}$	$-0.5703\ (0.1798)^{**}$	$-0.5806\ (0.1742)^{**}$	$-0.6085\ (0.1651)^{***}$
β_{13}	$-0.3270\ (0.1243)^{**}$	$-0.3412\ (0.1357)^{*}$	$-0.3346\ (0.1328)^{*}$	$-0.3159\ (0.1381)^{*}$
β_{14}	$-1.2916\ (0.1865)^{***}$	$-0.9738\ (0.4263)^{*}$	$-0.9735\ (0.4070)^{*}$	$-0.9762\ (0.2860)^{***}$
β_{21}	$0.2021\ (0.0996)^{*}$	$0.1436\ (0.1287)$	$0.1899\ (0.1270)$	$0.1929\ (0.1356)$
β_{22}	$-1.0853\ (0.5305)^{*}$	$-0.8006\ (0.3741)^{*}$	$-0.8212\ (0.3719)^{*}$	$-0.8440\ (0.2308)^{***}$
β_{23}	$0.1580\ (0.1852)$	$0.2056\ (0.1979)$	$0.2116\ (0.1946)$	$0.1914\ (0.1699)$
β_{31}	$-0.1844\ (0.2276)$	$0.1050\ (0.2739)$	$0.0641\ (0.2735)$	$0.1010\ (0.2561)$
β_{33}	$-0.6627\ (0.1768)^{***}$	$-0.5829\ (0.2032)^{**}$	$-0.5844\ (0.1979)^{**}$	$-0.5900\ (0.1708)^{***}$
β_4	$-0.0259\ (0.0258)$	$-0.0534\ (0.0342)$	$-0.0571\ (0.0344)$	$-0.0577\ (0.0306)$
β_5	$0.0001\ (0.0001)$	$0.0001\ (0.0001)^{*}$	$0.0001\ (0.00001)^{*}$	$0.0001\ (0.0001)^{**}$
β_6	$0.0118\ (0.0698)$	$-0.0026\ (0.0806)$	$-0.0167\ (0.0827)$	$-0.0177\ (0.0739)$
$log(\sigma)$	$1.5311\ (0.1433)^{***}$	$0.3483\ (0.2147)$	$-0.7335\ (0.2708)^{**}$	$-0.7780\ (0.1826)^{***}$
$log(\sigma_1)$	$-1.1628\ (0.2381)^{***}$			
$log(\sigma_3)$	$-0.6677\ (0.2380)^{**}$			
$log(\sigma_4)$	$-2.5245\ (0.3089)^{***}$			
μ	$-0.0000\ (0.0436)$	$0.0009\ (0.0664)$	$-0.0091\ (0.0641)$	$-0.0000\ (0.0572)$
σ_t^2	0.0000	0.1632	0.1553	0.0002
ν		$-0.3120\ (0.2039)(*)$	$-2.1727\ (1.2360)$	
N.O	219	219	219	219
GAIC	1090.7320	1053.7933	1056.7400	1055.8715

Significâncias ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$); N.O (Número de observações)
σ_i sendo a i-ésima dispersão por fazenda; $logit(\nu)$ em (*)

Possíveis explicações para os valores que se encontram fora da linha de referência do gráfico de resíduo (Figura 3.4) podem ter sido ocasionados pelos erros não controlados inerentes ao estudo observacional. Existem variações externas que podem estar influenciando nos resultados de prenhez, por exemplo, a recuperação dos sêmens, que foi realizada de forma não cirúrgica, acarretando variações relacionadas ao profissional que irá conduzir o procedimento, uma vez que a capacidade técnica do inseminador de reconhecer a correta detecção do estro, pode induzir o sistema reprodutório da fêmea. Além disso, fatores quanto às condições de transporte dos embriões realizado em filtros e locais, mesmo com ambientes propícios para a eficiência reprodutiva, podem acarretar variações amostrais (Demetrio et al., 2007).

Existem diferentes protocolos de sincronização de estro com vantagens e desvantagens para cada protocolo mas o ponto crítico em relação à sincronização da fêmea é que o momento do estro deverá coincidir com o tempo de inseminação da vaca, assim esses animais terão um ambiente uterino mais preparado para a inseminação artificial (Pereira et al., 2017). Nesse sentido, a falta de sincronizações do estro das fêmeas também reduzirá grandemente a probabilidade da gravidez ser estabelecida, ou seja, promovendo menores quantidades de prenhez e atribuindo aos resíduos pontos discrepantes.

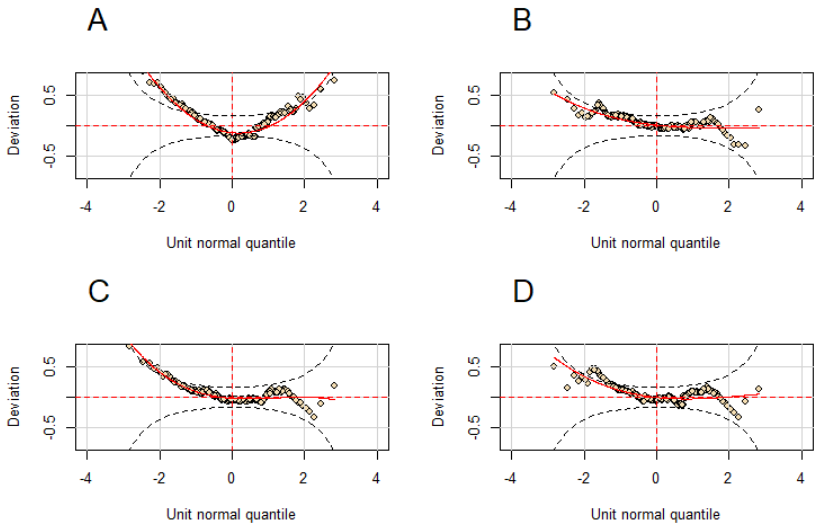


Figura 3.4. Dados de TE: *Warm plot* para os modelos de efeito aleatório A) doubler; B) Delaporte; C) Sichel; D) PIG

Baseado nos efeitos que foram significativos no modelo Delaporte de efeito aleatório, tem-se os coeficientes estimados apresentados na Tabela 3.3.

Tabela 3.3. Tabela resumo das estimativas dos parâmetros com os respectivos erros-padrão (SE), para o modelo Delaporte com efeito aleatório

Parâmetros	DEL (SE)
β_0	1.5340 (0.2311)
β_{11}	−0.5720 (0.1701)
β_{13}	−0.2667 (0.1382)
β_{14}	−1.0881 (0.2790)
β_{21}	0.1298 (0.1360)
β_{22}	−0.8050 (0.2436)
β_{23}	0.2262 (0.1706)
β_{31}	0.1100 (0.2136)
β_{33}	−0.6070 (0.1701)
β_5	0.0001 (0.0001)
σ	0.3360 (0.4436)
ν	−0.4573 (0.4243)
μ	−0.0008 (0.0627)
σ_t^2	0.0991
N.O	219
GAIC	1053.8475

N.O (número de observações)

3.4 Conclusão

A metodologia GAMLSS permite escolher várias funções de ligação para os efeitos das variáveis preditoras sobre a variável dependente. Conforme mostrado nesse trabalho, houve uma consistência entre as significâncias nos modelos, o que acredita ser indícios de concordância entre a prática e os resultados apresentados aqui. Assim, foram modelados diferentes estruturas

de preditores lineares do parâmetro de dispersão e verificado que a inclusão das dispersões por fazendas não foram interessantes, visto que o modelo não teve bom ajuste dos resíduos.

Na presença de estruturas de correlação e o fenômeno da superdispersão, incluiu-se o efeito aleatório de touro no preditor linear para tentar captar essa variabilidade extra, sendo considerado o modelo Delaporte como mais adequado, pois além de contemplar o ajuste da média, dispersão e assimetria presente nos dados, esse se mostrou de forma satisfatória quanto às análises de diagnóstico.

No sentido prático, pode-se concluir que os fatores que influenciaram no resultado de prenhez foram a estação do ano, estágio de vida da vaca, produtividade leiteira e as fazendas.

Referências

Carvalho, J. S., Cavalcanti, M. O., Chaves, M. S., e Rizzo, H. (2019). Eficiência da inseminação artificial em tempo fixo em fêmeas zebuínas na mesorregião sudeste do pará, brasil. *Revista de Ciências Agrárias Amazonian Journal of Agricultural and Environmental Sciences*, 62.

Demetrio, D., Santos, R., Demetrio, C., e Vasconcelos, J. L. M. (2007). Factors affecting conception rates following artificial insemination or embryo transfer in lactating holstein cows. *Journal of Dairy Science*, 90(11):5073–5082.

Draper, N. R. e Smith, H. (1981). Applied regression analysis 2nd ed.

Goncalves, P. B. D., de FIGUEIREDO, J., e de F.(Ed.) FREITAS, V. (2001). *Biotécnicas aplicadas à reprodução animal*. Livraria Varela.

Gujarati, D. N. e Porter, D. C. (2011). *Econometria Básica-5*. Amgh Editora.

Hasler, J. F. (2014). Forty years of embryo transfer in cattle: A review focusing on the journal theriogenology, the growth of the industry in north america, and personal reminisces. *Theriogenology*, 81(1):152–169.

Hastie, T. e Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, pages 1005–1016.

Hinde, J. e Demétrio, C. G. (1998a). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.

Hoffmann, R. (2006). *Análise de Regressão - Uma Introdução a Econometria*. Hucitec, S o Paulo, 4 edition.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.

Pereira, M., Sanches Jr, C., Guida, T., Wiltbank, M., e Vasconcelos, J. (2017). Comparison of fertility following use of one versus two intravaginal progesterone inserts in dairy cows without a cl during a synchronization protocol before timed ai or timed embryo transfer. *Theriogenology*, 89:72–78.

Rigby, B. e Stasinopoulos, M. (2009). *A flexible regression approach using GAMLSS in R.*

Rigby, R., Stasinopoulos, D., Heller, G., e De Bastiani, F. (2017). Distributions for modelling location, scale, and shape: using gamlss in r. *URL www. gamlss. org.(last accessed 5 March 2018)*.

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

Ross, S. (2009). *Probabilidade: um curso moderno com aplicaçoes*. Bookman Editora.

Sartori, R., Sartor-Bergfelt, R., Mertens, S., Guenther, J., Parrish, J., e Wiltbank, M. (2002). Fertilization and early embryonic development in heifers and lactating cows in summer and lactating and dry cows in winter. *Journal of dairy science*, 85(11):2803–2812.

Stasinopoulos, M., Rigby, B., Akantziliotou, C., e Voudouris, V. (2005). Generalized additive models for location scale and shape. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54:507–554.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., e De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*. Chapman and Hall/CRC.

Tzougas, G. (2014). *The Design of an Optimal Bonus-Malus System Based on the Sichel Distribution*.

4 CONSIDERAÇÕES FINAIS

Em síntese, o objetivo desse trabalho foi abordar técnicas inovadoras referente a utilização de modelos que captam a diversificação da natureza dos dados, em especial relacionados a estudos de eficiência reprodutiva em animais bovinos com aptidão para produção de leite.

Conforme apresentado, foi possível averiguar por meio de simulação que dados discretos com excessos de zeros podem ser ajustados com a distribuição COMPoisson sem que haja necessidade de transformações dos mesmos, visto que essa é uma prática comum nas mais diversas áreas quando as observações não seguem uma distribuição Gaussiana, o que ocasionará uma não normalidade nos resíduos.

Mensurações que tenham como variável resposta o número de prenhez podem ser modelados via metodologia GAMLSS, uma vez que conseguem captar várias formas de dispersões sem que haja perdas de informações. Dessa forma, os modelos de efeitos aleatórios foram satisfatórios, porém, foi indicado o ajuste com a distribuição Delaporte por contemplar um comportamento esperado pelos resíduos.

No sentido prático, pode-se dizer que os fatores produtividade leiteira, estágio de vida da vaca, estações do ano e fazendas que desempenharam o serviço foram significativos no resultado de prenhez.

Para trabalhos futuros serão abordados outras distribuições, bem como estruturas de variâncias e covariâncias, a fim de verificar possíveis dependências nos dados.