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

O problema da superdispersão em dados categorizados politômicos nominais em estudos agrários

Maria Letícia Salvador

Dissertação apresentada para obtenção de título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2019

Maria Letícia Salvador
Licenciatura Plena em Matemática

O problema da superdispersão em dados categorizados politômicos nominais em estudos agrários

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **IDEMAURO ANTONIO RODRIGUES DE LARA**

Dissertação apresentada para obtenção de título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2019

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

Salvador, Maria Letícia

O problema da superdispersão em dados categorizados politômicos nominais em estudos agrários / Maria Letícia Salvador. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2019 .

48 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Seleção de modelos; 2. Máxima verossimilhança; 3. Índice de superdispersão; 4. Dirichlet-multinomial . I. Título.

DEDICATÓRIA

Aos meus pais pais Silvana e Sebastião
dedico com muito amor e gratidão.
A minha irmã Aleteia,
pela torcida e por me fazer acreditar em mim
e nos meus empenhos.
A minha sobrinha Manuela:
sua alegria faz com que a vida fique mais doce.

À Mãe, Rainha e Vencedora Três Vezes Admirável de Schoenstatt,
gratidão pela graça alcançada e por todas as vezes que intercede por mim.

Em especial:
A Deus, por ser essencial em minha vida,
autor do meu destino, minha fortaleza.

AGRADECIMENTOS

A Deus que esteve comigo desde o primeiro dia, pelo apoio, força e segurança que me ajudaram a alcançar esta grande meta.

Aos meus pais, Sebastião e Silvana, pelo apoio, força e amor incondicional, por compreender os momentos de ausência. Sem vocês a realização desse sonho não seria possível.

À minha irmã, pelo companheirismo, pelo amor incondicional, por acreditar em mim.

À minha prima, Maria Beatriz, amiga de todas as horas e conselheira, agradeço pelo amor, por acredita em meu sonho.

Aos meus amigos, em especial Tobias, meu muito obrigado. Por todo o amor, por nunca negarem uma palavra de apoio, força e cumplicidade ao logo dessa etapa em minha vida.

Aos meus colegas da pós-graduação, em especial Glória, Caroline, Pollyane, Cristiane, Valdemiro, Vivian e Pórtya, pelo companheirismo e amizade.

Aos meus colegas de mestrado Eduardo e Welinton, pela ajuda fornecida nas configurações do trabalho.

Ao Prof. Dr. Idemauro Antonio Rodrigues de Lara, a orientação, pela disposição, apoio e o constante estímulo.

Aos professores do departamento de Matemática e Estatística da ESALQ-USP, pelo ensinamento, pelo apoio e pela prontidão e atenção dispensada.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro durante todo o curso.

Finalmente, a todos que direta ou indiretamente, contribuíram para a realização deste trabalho.

EPÍGRAFE

"Por melhor e mais eficiente que eu seja,
eu nunca serei tão bom ou tão eficiente quanto todos nós juntos".
(Pe. Haroldo)

"Tudo posso naquele que me fortalece"
(Felipenses 4,13).

SUMÁRIO

RESUMO

O problema da superdispersão em dados categorizados politômicos nominais em estudos agrários

Variáveis politômicas são comuns em experimentos agronômicos, apresentando natureza nominal ou ordinal. O modelo dos logitos generalizados é uma classe de modelos que pode ser empregada para a análise desses dados. Uma das características deste modelo é a pressuposição de que a variância é uma função conhecida da média e, espera-se, que a variância observada esteja próxima da variância pressuposta pelo modelo assumido. Contudo, quando ela é maior do que a especificada pelo modelo, tem-se o fenômeno da superdispersão. Nesse contexto, o presente trabalho objetivou caracterizar o problema da superdispersão associado a dados nominais em estudos "*cross-sectional*". Como motivação apresentam-se dois estudos adaptados da área de ciências agrárias relativos à fruticultura e zootecnia, ambos planejados no delineamento inteiramente casualizado. Verifica-se indicativo de superdispersão nos dados dos dois exemplos e como uma alternativa metodológica utilizou-se o modelo Dirichlet-multinomial. Por meio do gráfico de diagnóstico *half-normal plot* avaliou-se o ajuste do modelo dos logitos generalizados e do Dirichlet-multinomial. Adicionalmente, foi proposta uma extensão do índice de dispersão para os dados politômicos, com performance avaliada sob simulação. O modelo Dirichlet-multinomial mostrou-se adequado para o ajuste aos dados com superdispersão comparativamente ao modelo dos logitos generalizados. Apesar dos resultados satisfatórios obtidos, ressalta-se que este trabalho é uma introdução ao problema.

Palavras-chave: 1. Seleção de modelos; 2. Máxima verossimilhança; 3. Índice de superdispersão; 4. Dirichlet-multinomial

ABSTRACT

The problem of overdispersion in categorized polymorphic data in agrarian studies

Polytomic variables are common in agronomic experiments, presenting nominal or ordinal nature. The generalized logits model is a class of models that can be used to analyze these facts. One of the characteristics of this model is the assumption that variance is a known function of the mean and. It is expected, that the analyzed variance is close to that assumed by the model. However, when it is larger than the one specified by the model, it has the phenomenon of overdipersion. In this context, the present work aims to characterize the problem of overdispersion associated with nominal data in cross-sectional studies. As motivation, it is showed two adapted studies of the agricultural sciences área, related to fruit growing and zootechnics, both planned in the completely randomized design. The Dirichlet-multinomial model was used as a methodological alternative and was indicated as an overdispersion in the facts of the two examples. The model of the generalized logits and the Dirichlet-multinomial model were evaluated using the half-normal plot. In addition, it was proposed an extension of the dispersion index for the polytomic data, with performance evaluated under simulation. The Dirichlet-multinomial model proved to be adequate for the adjustment to the overdispersed fact compared to the generalized logit model. Despite the satisfactory results obtained, it is emphasized that this work is an introduction to the problem.

Keywords: 1. Selection of models; 2. Maximum likelihood; 3. Overdispersion index; 4. Dirichlet-multinomial

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Dados categorizados decorrem da observação de características dos indivíduos que dizem respeito a uma qualidade ou atributo, expressos em categorias mutuamente exclusivas. Esses dados são frequentes na prática em diversas áreas, em especial, nas Ciências Agrárias em que são realizados experimentos cuja variável resposta refere-se à gravidade de uma doença em frutos, ao comportamento de um animal, características fenotípicas e genéticas, dentre outros.

As variáveis categorizadas podem ser classificadas de acordo com o número de categorias, sendo dicotômicas para duas categorias, ou politômicas para três ou mais categorias. As variáveis politômicas podem ser classificadas de acordo com a sua natureza, sendo ordinal se seguir uma ordem natural, ou nominal, caso contrário. De acordo com Agresti (2002), quando a variável resposta é politômica, o modelo probabilístico mais associado é a distribuição multinomial, que é uma extensão da distribuição binomial.

Por outro lado, modelos que envolvem a distribuição multinomial são extensões dos Modelos Lineares Generalizados (MLG) propostos por Nelder e Wedderburn (1972). Os MLG permitem analisar dados de natureza discreta ou contínua e constituem uma ferramenta poderosa de análise de dados. Um dos objetivos dos MLG é analisar as influências que uma ou mais variáveis explicativas exercem sobre a variável resposta.

Para uma variável dicotômica o modelo clássico de análise é a regressão logística. Com mais de duas categorias de resposta, os modelos mais usuais são: modelo dos logitos generalizados para o caso nominal e os modelos de chances proporcionais, chances parciais e cumulativos para o caso ordinal. Nesses modelos, a variância é uma função conhecida da média e, espera-se que a variância observada esteja próxima da variância pressuposta pelo modelo assumido. Entretanto, nem sempre isso ocorre. Na área das ciências agrárias, por exemplo, quando se trabalham com dados entomológicos, de comportamento de animais, de florescimento de espécies, entre outros, não é raro, constatar que há uma heterogeneidade da variável resposta resultando em uma variância maior do que a especificada pelo modelo proposto. Segundo Hinde e Demétrio (1998) estas situações experimentais em que esta discrepância ocorre, ou seja, a variância observada maior do que a nominal, são típicas do fenômeno da superdispersão.

Porém, segundo Olsson (2002) deve-se ter cautela para não confundir o fenômeno da superdispersão com o ajuste insatisfatório do modelo, que pode ser causado pela escolha errada da função de ligação, por exemplo. Reconhecer que a superdispersão está presente nos dados é primordial para que sejam tomadas algumas alternativas, a fim de trabalhar com o problema na escolha de um modelo com o intuito de garantir uma estimação com maior segurança.

Na literatura, encontram-se modelos capazes de solucionar o problema da superdispersão, como por exemplo os modelos: de quase-verossimilhaça (Wedderburn, 1974) e o modelo de dois estágios (Hinde e Demétrio, 1998), em particular para dados binomiais e de contagem. Neste trabalho utilizou-se a abordagem do modelo de dois estágios, por meio da distribuição composta Dirichlet-multinomial (Mosimann, 1962), que tem sido utilizada em análises de dados politômicos que apresentam variação extra-multinomial.

Assim sendo, este trabalho tem como objetivos específicos:

i) Caracterizar o problema da superdispersão no contexto de dados categorizados politômicos

nominais com ilustração de duas aplicações nas ciências agrárias;

ii) Propor um índice de dispersão como uma medida descritiva e diagnóstica do fenômeno da superdispersão para dados politômicos nominais, bem como avaliar a sua performance sob simulação;

iii) Utilizar o modelo Dirichlet-multinomial, como uma alternativa para os casos de superdispersão em dados multinomiais;

iv) Utilizar o gráfico de diagnóstico *half-nomal plot* para avaliar a qualidade de ajuste do modelo.

2 REVISÃO DE LITERATURA

2.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG), propostos por Nelder e Wedderburn (1972), são uma extensão dos modelos de regressão lineares clássicos. Nesta classe de modelos há mais opções para a distribuição da variável resposta, desde que a mesma pertença à família exponencial de distribuições, além de uma maior flexibilidade na relação funcional entre a média da variável resposta e o preditor linear, que inclui as variáveis explanatórias.

Segundo Cordeiro e Demétrio (2008), os MLG podem ser usados quando se tem uma variável aleatória Y associada a um conjunto de variáveis explanatórias x_i, $i = 1, \cdots, p$. Para uma amostra de n observações $(y_i, \boldsymbol{x}_i)$, em que $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$ é o vetor coluna de variáveis explanatórias, os MLG apresentam três componentes:

(i) Componente aleatório: identifica a variável resposta e sua distribuição.

Sejam Y_i, $i = 1, \cdots, n$ variáveis aleatórias independentes provenientes de uma mesma distribuição que pertence à família exponencial na forma canônica, com média $\mu_1, \mu_2, \cdots, \mu_n$, ou seja,

$$f(y_i, \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\},$$

em que $b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas, θ_i o parâmetro canônico e $\phi > 0$ um parâmetro de dispersão. A média e a variância de Y_i são dados por:

$$\mathrm{E}(\mathrm{Y}_i) = \mu_i = b'(\theta_i) \quad \text{e} \quad \mathrm{Var}(Y_i) = \phi b''(\theta_i) = \phi \mathrm{V}_i.$$

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

(ii) Componente sistemático: especifica as variáveis explanatórias, que entram na estrutura como uma soma linear de seus efeitos, dando origem a um vetor de preditores lineares

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

ou na forma matricial:

$$\boldsymbol{\eta} = \boldsymbol{X}^T\boldsymbol{\beta},$$

em que $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)^T$ é o vetor de preditores lineares, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ é o vetor de parâmetros e $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$ é a matriz de delineamento.

(iii) Função de ligação: uma função monótona $g(.)$ e diferenciável que relaciona o componente sistemático ao aleatório:

$$\eta_i = g(\mu_i)$$

Se $g(\mu_i) = \theta_i = \eta_i$ a função de ligação é denominada canônica, pelo fato do preditor linear η_i modelar diretamente o parâmetro canônico θ_i.

Segundo McCullagh e Nelder (1989), uma escolha importante do MLG é o trinômio: distribuição de probabilidade da variável resposta, matriz do modelo e, por fim, a função de ligação.

2.1.1 Estimação dos Parâmetros

Dentre os vários métodos propostos para a estimação do vetor de parâmetros ($\boldsymbol{\beta}$) o mais utilizado é o de máxima verossimilhança, que tem como propriedades: eficiência, consistência e normalidade assintótica. Para uma amostra de tamanho n, o logaritmo da função de verossimilhança, é dado por:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln f(y_i, \theta_i, \phi) = \phi^{-1} \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)] + c(y_i, \phi). \tag{2.1}$$

Derivando-se a função (2.1) em relação a β_r, tem-se a função escore:

$$U_r = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_r}, \qquad r = 1, \cdots, p. \tag{2.2}$$

Usando a regra da cadeia,

$$\frac{\partial l_i}{\partial \beta_r} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_r},$$

a função (2.2) pode ser expressa por:

$$U_r = \phi^{-1} \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ir}, \qquad r = 1, \cdots, p.$$

As estimativas de máxima verossimilhança de $\boldsymbol{\beta}$ são calculadas igualando U_r a zero. Em geral, as equações são não lineares e são resolvidas via processo iterativo, que podem ser obtidas por meio do algoritmo de *Newton-Raphson* (mínimos quadrados ponderados iterativamente), deve ser executado até atingir um critério de convergência (Dobson, 1990). O processo é sintetizado na equação:

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^T \boldsymbol{W}^{(m)} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(m)} \boldsymbol{z}^{(m)},$$

em que m denota cada passo do processo iterativo, $\boldsymbol{X}$ é a matriz do modelo, $\boldsymbol{W} = \text{diag}\{w_1, \cdots, w_n\}$ é a matriz diagonal de pesos, em que $w_i = \text{V}_i^{-1}(\partial \mu_i / \partial \eta_i)^2$ e $\boldsymbol{z}^{(m)} = \eta_i + (y_i - \mu_i^{(m)}) g'(\mu_i^{(m)})$ é o vetor da variável dependente ajustada.

2.1.2 Função Desvio e Estatística de Pearson

McCullagh e Nelder (1989) descrevem que para um conjunto de dados com n observações de uma variável aleatória Y, pode-se ajustar modelos do mais simples, com apenas um parâmetro (modelo nulo), até o modelo mais complexo, contendo o número máximo de parâmetros possíveis (modelo saturado). Deseja-se encontrar um modelo intermediário que melhor se

ajuste entre esses dois extremos, denominado modelo sob pesquisa (Demétrio, 2002). Neste contexto, o problema está em identificar a utilidade de um parâmetro extra no modelo ou verificar a falta de ajuste induzida pela omissão dele.

Uma medida de discrepância, proposta por Nelder e Wedderburn (1972), para avaliar o ajuste do modelo é a função desvio (*deviance*), cuja expressão é dada por:

$$S_p = 2(\hat{l}_n - \hat{l}_p),$$

sendo que $\hat{l}_n$ e $\hat{l}_p$ denotam o máximo do logaritmo da função de verossimilhança do modelo saturado e sob pesquisa com p parâmetros, respectivamente. Se $\tilde{\theta}_i = \tilde{\theta}_i(y_i)$ e $\hat{\theta}_i = \hat{\theta}_i(\hat{\mu}_i)$ são as estimativas de máxima verossimilhança do parâmetro canônico do modelo saturado e sob pesquisa, pode-se escrever:

$$S_p = \phi^{-1} 2 \sum_{i=1}^{n} [y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)] = \phi^{-1} D_p,$$

em que S_p e D_p são denominados a função do desvio escalonado e a função desvio, respectivamente. De acordo com Cordeiro e Demétrio (2007) a função desvio pode indicar a qualidade de ajuste, pelo fato de quanto melhor o ajuste do modelo, menor deverá ser o valor do desvio.

Assintoticamente, pode-se averiguar a discrepância do ajuste do modelo a um conjunto de dados por meio da estatística χ^2_P de Pearson generalizada, definida por:

$$\chi^2_P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

sendo $V(\hat{\mu}_i)$ a função de variância estimada para a distribuição em estudo.

2.2 Técnicas para diagnósticos em MLG

Meyers, Montgomery e Vining (2010) salientam que as técnicas de diagnósticos são de grande importância para a validação da pressuposição de um modelo estatístico. Em MLG as técnicas são análogas a dos modelos lineares clássicos, com adaptações.

Por exemplo, os vetores $\boldsymbol{y}$ e $\hat{\boldsymbol{\mu}}$ que são considerados para verificar a pressuposição de linearidade dos modelos clássicos, são substituídos pela variável ajustada $\boldsymbol{z}$ e pelo preditor linear $\hat{\boldsymbol{\eta}}$. A variância residual será substituída pela estimativa do parâmetro de dispersão ϕ e a matriz de projeção $\boldsymbol{H}$ é definida por:

$$\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}^{\frac{1}{2}}.$$

sendo $\boldsymbol{W}$ a matriz de pesos definida em (**??**).

2.2.1 Tipos de resíduos

Os resíduos mais utilizados para a análise de diagnóstico dos MLG são:

i) Resíduos ordinários:

$$r_i = y_i - \hat{\mu}_i,$$

em que y_i representa a variável resposta e $\hat{\mu}_i$ a estimativa correspondente.

ii) Resíduos de Pearson generalizados:

$$r_{P_i} = (y_i - \hat{\mu}_i)\sqrt{\frac{w_i}{V(\hat{\mu}_i)}},$$

no qual w_i são os pesos a priori e $V(\hat{\mu}_i)$ a função de variância.

iii) Resíduos de Pearson generalizados estudentizados:

$$r_{Pe_i} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi}V(\hat{\mu}_i)(1-h_i)}},$$

no qual $\hat{\phi}$ é a estimativa do parâmetro de dipersão ϕ e h_i é o i-ésimo elemento da matriz de projeção $\boldsymbol{H}$.

iv) Componente do desvio:

É definido como a raiz quadrada de cada elemento da *deviance* com o sinal do resíduo ordinário:

$$d_{r_i} = \sqrt{d_i} \times sinal(r_i),$$

no qual, $d_i = \{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}$.

v) Resíduos da *deviance* estudentizada:

$$d_{e_i} = \frac{d_{r_i}}{\sqrt{\hat{\phi}(1-h_i)}}.$$

Essas medidas de discrepância são de grande importância para avaliar a diferença entre o valor observado e o valor estimado do modelo.

2.2.2 Técnicas gráficas

Dispositivos gráficos são importantes para a análise de resíduos e diagnósticos. Técnicas específicas são descritas em Pregibon (1981), Williams (1987), McCullagh e Nelder (1989), Demétrio (2002) e um resumo destes procedimentos gráficos podem ser encontrados em Paula (2004). Uma técnica exploratória utilizada são os gráficos normal e semi-normal de probabilidades que são convenientes para averiguar se a função de variância foi corretamente especificada, e detectar a presença de *outliers*. O comportamento esperado dos resíduos para um modelo adequado é aproximadamente uma reta.

Com relação ao gráfico normal de probabilidades, afim de estabelecer um padrão de comparação, Atkinson (1985) propõe a adição de um envelope de simulação, tal que, para um modelo bem ajustado esperam-se que as observações se distribuam dentro do envelope.

A obtenção do envelope de simulação consiste em:

i) Ajustar o modelo e calcular os resíduos r_i, em valor absoluto, e colocá-los em ordem crescente;

ii) Simular 99 amostras para a variável resposta, usando a mesma matriz das variáveis exploratórias;

iii) Ajustar o mesmo modelo a cada variável resposta simulada e extrair os resíduos do modelo, e novamente ordenar os valores absolutos;

iv) Calcular os percentis 5%, 50% e 95%;

v) Plotar os percentis desejados dos valores de diagnósticos observados em cada estatística de ordem esperada e utilizá-los para formar o envelope de simulação.

Em se tratando de dados discretos, Hinde e Demétrio (1998) destacam o uso do gráfico *half-normal plot* (hnp). O hnp consiste em plotar os valores absolutos ordenados de uma medida de diagnóstico adequada *versus* os valores esperados da estatística de ordem, em valor absoluto, da distribuição meio-normal, definida por:

$$\Phi^{-1}\left[\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right],$$

no qual Φ é a função acumulada da distribuição normal padrão, sendo $1 \leq i \leq n$ em que n é o tamanho da amostra.

Ademais, Moral, Hinde e Demétrio (2017) salientam que o objetivo do hnp não é fornecer uma região de aceitação ou rejeição das observações, mas sim servir como guia do que se esperar de um modelo bem ajustado. Além disso, esses gráficos também são úteis para detectar possíveis *outliers*, o fenômeno da superdispersão e se a função de ligação ou se a distribuição do erros são adequadas. Moral (2013) implementou a função hnp() no *software* R (R Core Team, 2018), permitindo o uso do hnp para diversas funções de probabilidade, diferentes funções de ligação e para diferentes tipos de resíduos (Moral, Hinde e Demétrio, 2018).

2.2.3 Verificação da Função de Ligação

Existem dois métodos para verificar a adequabilidade da função de ligação, o formal e o informal. Informalmente, a verificação da função de ligação é realizada por meio de técnicas exploratórias (gráficos). Podem-se usar dois tipos de gráficos, o mais simples consiste no diagrama de dispersão da variável dependente ajustada $\boldsymbol{z}$ contra o preditor linear $\hat{\boldsymbol{\eta}}$, no qual o padrão nulo que é uma reta indicará se a função escolhida é satisfatória. Outra maneira de verificar a função de ligação é por meio do gráfico da variável adicionada. Formalmente, utiliza-se o teste da razão de verossimilhança, que consiste em adicionar $\hat{\boldsymbol{\eta}}^2$ como variável explanatória extra e

examinar se a mudança ocorrida no desvio é significativa. Se isso ocorrer, há evidências de que a função de ligação é inadequada (Paula, 2004).

2.2.4 Verificação da Função de Variância

Segundo Cordeiro e Demétrio (2008), a função de variância é definida pela distribuição da variável resposta, que pode ser afetada pela escolha errada da função de ligação. Informalmente, a verificação da função de variância pode ser realizada por meio do gráfico dos resíduos *versus* os valores ajustados transformados, em uma escala com variância constante. O padrão nulo será uma distribuição aleatória de média zero e amplitude constante. Formalmente, pode-se incorporar um parâmetro λ, $V(\lambda) = \mu^{\lambda}$, e fazer o teste de hipótese $H_0 : \lambda = \lambda_0$ utilizando o teste da razão de verossimilhança ou o escore (Paula, 2004).

2.3 Superdispersão

Em algumas situações práticas, em especial nas ciências agrárias, quando o modelo linear generalizado é ajustado aos dados (na forma de contagem ou de proporção), mesmo assumindo que o preditor linear resultante é adequado, é comum observar a presença da superdispersão. Segundo McCullagh e Nelder (1989) este fenômeno é caracterizado pelo fato da variabilidade extra ser maior do que a prevista pela relação implícita média-variância. Na literatura, a superdispersão tem sido amplamente considerada na análise de dados discretos (Ridout *et al.* 1998, Vieira 2008, Dobson 2010, Paula 2004). Por exemplo, para dados na forma de proporção, em que $\mathrm{Y}_i \sim \mathrm{Binomial}(m_i, \pi_i)$, a variabilidade extra-binomial é observada quando $\mathrm{Var}(\mathrm{Y}_i) > m_i\pi_i(1-\pi_i)$. E para dados na forma de contagem em que $\mathrm{Y}_i \sim \mathrm{Poisson}(\mu_i)$ essa variabilidade extra-Poisson pode ser caracterizada quando $\mathrm{Var}(\mathrm{Y}_i) > \mu_i$.

Para contagens, algumas medidas foram propostas para caracterizar a variabilidade extra-Poisson, sendo a mais comum entre elas o índice de dispersão proposto por Winkelmann (1995). Então, se $y_1, \cdots, y_n$ são n observações de contagens da variável aleatória Y, o índice de dispersão é definido por:

$$\mathrm{ID(Y)} = \frac{\mathrm{Var(Y)}}{\mathrm{E(Y)}},$$

que avalia a distorção da variância em relação a média. Como na distribuição Poisson pressupõe-se $\mathrm{E(Y)} = \mathrm{Var(Y)}$ decorre que, quando $\mathrm{ID} > 1$, $\mathrm{ID} < 1$ e $\mathrm{ID} = 1$ há indícios de superdispersão, subdispersão e equidispersão, respectivamente. Um índice alternativo à equação (**??**) foi prposto por Ridout *et al.* (1998):

$$\mathrm{ID(Y)} = \frac{\mathrm{Var(Y)} - \mathrm{E(Y)}}{\mathrm{E(Y)}},$$

afim de que o valor de referência seja zero, ou seja, se $\mathrm{ID} = 0$ há indícios de equidispersão. Uma vez constatada a superdispersão é preciso investigar suas causas.

De acordo com Hinde e Demétrio (1998), o fenômeno da superdispersão pode ocorrer devido a vários motivos tais como:

(i) Variabilidade do experimento;

(ii) Amostragem por conglomerados;

(iii) Correlação entre respostas individuais;

(iv) Omissão de covariáveis que possam explicar a falta de homogeneidade;

(v) Excesso de zeros nos dados em estudo.

Segundo os autores, diferentes causas de superdispersão podem coexistir, tornando-se difícil inferir a causa precisa da superdispersão. Com isso, medidas de diagnóstico se mostram ferramentas importantes para estudar a superdispersão e é preciso estar atento para a sua ocorrência, pois na presença da superdispersão e com o uso de modelos não apropriados pode-se obter estimativas erradas dos erros padrões, e consequentemente, avaliar incorretamente a significância dos parâmetros de regressão. Além do fato que se deve tomar cuidado para não confundir a superdispersão com o ajuste insatisfatório do modelo (Olsson, 2002),que pode ser devido a:

(i) Escolha errada do preditor linear ($\boldsymbol{\eta}$);

(ii) Presença de outilliers;

(iii) Escolha inapropriada da função de ligação;

(iv) Dimensão amostral tal que os pressupostos da teoria assintótica não são satisfeitos.

Hinde e Demétrio (1998) apresentam uma revisão dos modelos e métodos de estimação para dados discretos que apresentam superdispersão, com ênfase em contagens e proporção. Os autores apresentam duas propostas, a saber:

(i) Modelos que admitem uma forma mais geral para a função de variância, permitindo um parâmetro adicional. Nessa classe, os parâmetros podem ser estimados por diferentes métodos de estimação, sendo os mais comuns: quase-verossimilhança e pseudo-verossimilhança.

(ii) Modelos de dois estágios para a variável resposta. Nessa classe, assume-se uma distribuição para a variável resposta e adicionalmente uma distribuição de probabilidade para os parâmetros do modelo. Exemplos dessa classe são o modelo binomial-negativo e Poisson-normal para dados em forma de contagens e o modelo beta-binomial e binomial-normal para dados de proporções.

Considere, como ilustração dados de proporção sob à suposição de superdispersão, um modelo de classe (i) tem a seguinte forma de variância:

$$\mathrm{Var}(\mathrm{Y}_i) = m_i\pi_i(1-\pi_i)\{1+\phi[m_i-1]^{\delta_1}[\pi_i(1-\pi_i)]^{\delta_2}\}. \tag{2.3}$$

De (2.3), defini-se várias funções de variância para diferentes valores δ_1, δ_2 e ϕ. Se $\phi = 0$, tem-se a variância do modelo binomial padrão com $\mathrm{Var}(\mathrm{Y}_i) = m_i\pi_i(1-\pi_i)$, se $\delta_1 = \delta_2 = 0$, tem-se $\mathrm{Var}(\mathrm{Y}_i) = m_i\pi_i(1-\pi_i)[1+\phi]$ que é o modelo binomial reparametrizado com superdispersão

constante. Finalmente, se $\delta_1 = 1$ e $\delta_2 = 0$, tem-se $\text{Var}(\text{Y}_i) = m_i\pi_i(1-\pi_i)[1+\phi(m_i-1)]$ que é a função de variância da distribuição beta-binomial.

Alternativamente, para modelos de dois estágios da classe (ii), assume-se que $\text{Y}_i \mid \pi_i \sim \text{Binomial}(m_i, \pi_i)$, em que π_i é uma variável aleatória que assume valores entre zero e um com $\text{E}(\pi_i) = \mu_i$ e $\text{Var}(\pi_i) = \phi\mu_i(1-\mu_i)$. Assim, tem-se que $\text{E}(\text{Y}_i) = m_i\mu_i$ e $\text{Var}(\text{Y}_i) = m_i\mu_i(1-\mu_i)[1+\phi(m_i-1)]$ que representa um modelo com superdisersão. Um caso especial é considerar que $\pi_i \sim \text{Beta}(\alpha_i, \beta_i)$, com $\alpha_i + \beta_i$ constante, o que resulta em Y_i tem distribuição beta-binomial.

Para o caso de dados politômicos nominais, Morel e Nagaraj (1992) discutem o uso do modelo Dirichlet-multinomial para modelar a superdispersão, sendo este uma extensão multivariada do modelo beta-binomial de dois estágios.

2.4 Distribuições de probabilidades para dados politômicos nominais

2.4.1 Distribuição Multinomial

A distribuição multinomial é uma extensão da distribuição binomial, associada a situações experimentais com variável politômica.

Considere um experimento aleatório e n realizações independentes, no qual o espaço amostral admite J possíveis resultados, mutuamente exclusivos, $A_1, A_2, \cdots, A_J$. Dessa forma, a cada realização do experimento ocorrerá somente um dos eventos com probabilidade $\pi_j = \text{P}(A_j)$, sendo $\sum_{j=1}^{J}\pi_j = 1$. Seja $\mathbf{Y}= \ (\text{Y}_1, \text{Y}_2, \cdots, \text{Y}_J)$ um vetor aleatório, em que cada um dos componentes Y_j é uma variável aleatória que descreve o número de vezes que A_j foi observado nas n realizações do experimento, o vetor $\mathbf{Y}$ segue uma distribuição multinomial com parâmetros n e $\boldsymbol{\pi}$, em que $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_J)$.

Assim, a função de distribuição de probabilidade de Y_j é dada por:

$$f_{\text{Y}_1,\cdots,\text{Y}_J}(y_1, \cdots, y_J) = \text{P}(\text{Y}_1 = n_1, \cdots, \text{Y}_j = n_j) = f(\boldsymbol{y}|\boldsymbol{\pi}) = \frac{n!}{n_1!\, n_2! \cdots n_J!}\prod_{j=1}^{J}\pi_j^{n_j},$$

em que $\sum_{j=1}^{J} n_j = n$.

Tem-se que (**??**) define a distribuição multinomial com valor esperado:

$$\text{E}(\text{Y}_j) = n\pi_j.$$

e variância e covariância dadas por:

$$\text{Var}(\text{Y}_j) = n\pi_j(1-\pi_j).$$

$$\text{Cov}(\text{Y}_j, \text{Y}_k) = -n\pi_j\pi_k, \quad j \neq k, \quad j,k = 1, \cdots, J.$$

2.4.2 Distribuição Dirichlet

A distribuição Dirichlet foi apresentada por Connor e Mosimann (1969) para modelar dados em proporção e independentes. A distribuição Dirichlet é uma extensão multivariada da

distribuição Beta, empregada no estudo de distribuições conjuntas de variáveis aleatórias que pertençam ao intervalo $[0, 1]$.

Se $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_J)$ é um vetor aleatório que segue a distribuição Dirichlet, então a função de probabilidade conjunta é definida por:

$$f(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} (\pi_j)^{\alpha_j - 1},$$

em que os parâmetros $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_j)$ são estritamente positivos, $\sum_{j=1}^{J} \pi_j = 1$, a região $\Omega = \{\boldsymbol{\pi}; \pi_j \in (0, 1), j = 1, \cdots, J; \sum_{j=1}^{J} \pi_j = 1\}$ e $\Gamma(.)$ é a função Gama.

Seja $\varphi = \sum_{j=1}^{J} \alpha_j$. O valor esperado, a variância e a covariância são dados respectivamente,

$$\mathrm{E}(\pi_j) = \frac{\alpha_j}{\varphi} = \mu_j, \quad j = 1, \cdots, J. \tag{2.4}$$

$$\mathrm{Var}(\pi_j) = \frac{\alpha_j(\varphi - \alpha_j)}{\varphi^2(1 + \varphi)} = \mu_j(1 - \mu_j)\rho \quad j = 1, \cdots, J, \tag{2.5}$$

em que, $\rho = \frac{1}{1+\varphi}$.

$$\mathrm{Cov}(\pi_j, \pi_k) = -\frac{\alpha_j \alpha_k}{\varphi^2(1 + \varphi)}, \quad j \neq k, \quad j, k = 1, \cdots, J. \tag{2.6}$$

2.4.3 Distribuição Dirichlet-multinomial

A distribuição Dirichlet-multinomial é uma distribuição composta que foi introduzida por Mosimann (1962).

Sejam $\boldsymbol{Y}$ e $\boldsymbol{\pi}$ dois vetores aleatórios como no contexto de (2.4.1), tal que, em um primeiro estágio, $\mathbf{Y}|\boldsymbol{\pi} \sim$ Multinomial(n, $\boldsymbol{\pi}$) e, em um segundo estágio, $\boldsymbol{\pi}$ segue a distribuição Dirichlet como em (2.4.2) sob o espaço amostral Ω. Assim, $\mathbf{Y}$ segue uma distribuição Dirichlet-multinomial com função de probabilidade, dada por:

$$\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\alpha}) &= \int_{\Omega} f(\boldsymbol{y}|\boldsymbol{\pi}) f(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi} \\
&= \int_{\Omega} \frac{n!}{n_1!\, n_2! \cdots n_J!} \prod_{j=1}^{J} (\pi_j)^{n_j} \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} (\pi_j)^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= \frac{n!}{n_1!\, n_2! \cdots n_J!} \frac{\Gamma(\sum_{j=1}^{J} \alpha_{ij})}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \int_{\Omega} \prod_{j=1}^{J} (\pi_{ij})^{n_j + \alpha_j - 1} d\boldsymbol{\pi} \\
&= \frac{n!}{n_1!\, n_2! \cdots n_J!} \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\Gamma(n + \sum_{j=1}^{J} \alpha_j)} \prod_{j=1}^{J} \frac{\Gamma(n + \alpha_j)}{\Gamma(\alpha_j)},
\end{aligned} \tag{2.7}$$

em que, os parâmetros $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_J)$ são estritamente positivos e $\Gamma(.)$ é a função Gama.

Utilizando as propriedades de esperança condicional e as expressões (2.4), (2.5) e (2.6), o valor esperado, a variância e a covariância dos componentes de $\mathbf{Y}$, são dados por:

$$\mathrm{E}(Y_j) = \mathrm{E}[\mathrm{E}(Y_j|\boldsymbol{\pi})] = n\mathrm{E}(\pi_j) = n \cdot \frac{\alpha_j}{\sum_{j=1}^{J} \alpha_j} = n\mu_j.$$

$$\begin{aligned}
\mathrm{Var}(Y_j) &= \mathrm{E}[\mathrm{Var}(Y_j|\boldsymbol{\pi})] + \mathrm{Var}[\mathrm{E}(Y_j|\boldsymbol{\pi}))] = \mathrm{E}[n\pi_j(1-\pi_j)] + \mathrm{Var}[n\pi_j] \\
&= n\mathrm{E}[\pi_j(1-\pi_j)] + n^2\mathrm{Var}[\pi_j] = n\mathrm{E}(\pi_j) - n\mathrm{E}(\pi_j^2) + n^2\mathrm{Var}(\pi_j) \\
&= n\mathrm{E}(\pi_j) - n[\mathrm{Var}(\pi_j) + \mathrm{E}(\pi_j)^2] + n^2\mathrm{Var}(\pi_j) \\
&= n\mathrm{E}(\pi_j)[1 - \mathrm{E}(\pi_j)] + n(n-1)\mathrm{Var}(\pi_j) \\
&= n\mu_j(1-\mu_j) + n(n-1)\mu_j(1-\mu_j)\rho \\
&= n\mu_j(1-\mu_j)[1 + (n-1)\rho],
\end{aligned}$$

em que, $\mu_j = \frac{\alpha_j}{\sum_{j=1}^{J} \alpha_j}$.

$$\mathrm{Cov}(\mathrm{Y}_j, \mathrm{Y}_k) = -\frac{n(n+\varphi)}{(1+\varphi)}\mu_j\mu_k, \quad j \neq k, \quad j,k = 1, \cdots, J.$$

2.5 Modelos associados a dados politômicos nominais

2.5.1 Modelo dos logitos generalizados

Quando a variável resposta é politômica nominal, o modelo mais usual é o modelo dos logitos generalizados. Este modelo compara cada categoria de resposta com uma categoria de referência, frequentemente a última.

Considere um vetor aleatório com distribuição multinomial, em que seus componentes representam as ocorrências de categoria de resposta.

Para uma situação experimental, com uma amostra aleatória dessa distribuição e seja $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$ o vetor de variáveis explanatórias, sendo $\pi_j(\boldsymbol{x}) = P(\mathrm{Y} = j|\boldsymbol{x})$, com $\sum_{j=1}^{J} \pi_j(\boldsymbol{x}) = 1$ e $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jp})^T$ o vetor de parâmetros desconhecidos e de interesse, em que $j = 1, \cdots, J$.

Fixando-se a J-ésima categoria como referência, tem-se que o modelo é definido por (Agesti, 2002):

$$\boldsymbol{\eta} = \ln\left(\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right) = \lambda_j + \boldsymbol{\beta}_j^T\boldsymbol{x} = \lambda_j + \boldsymbol{\beta}_{j1}x_1 + \cdots \boldsymbol{\beta}_{jk}x_k, \tag{2.8}$$

em que $j = 1, \cdots J-1$.

No modelo dos logitos generalizados o intercepto ($\boldsymbol{\lambda}$) e vetor de parâmetros de regressão ($\boldsymbol{\beta}$) são diferentes para cada logito, o que implica que os efeitos das covariáveis variam de acordo com a categoria de resposta (Agresti, 2002).

Segundo Giolo (2017), embora o modelo (2.8) considere $J-1$ logitos para todos os possíveis pares de categorias, nota-se que os $J-1$ logitos considerados pelo modelo determinam os logitos para todos os outros pares de categoria.

Com relação à estimação dos parâmetros do modelo (2.8), Agresti (2002) mostra que o método empregado é o da máxima verossimilhança, em que o logaritmo da função de verossimilhança é dado por:

$$\ell = \ln \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \pi_j(x_i)^{y_{ij}} \right] = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} y_{ij}(\lambda_j + \boldsymbol{\beta}_j \boldsymbol{x}_i) - \ln \left[1 + \sum_{j=1}^{J-1} \exp(\lambda_j + \boldsymbol{\beta}_j \boldsymbol{x}_i) \right] \right\}. \quad (2.9)$$

A equação (2.9) não tem forma analítica fechada e deve-se usar o método de Newton-Raphson para obter as estimativas de máxima verossimilhança.

Assumindo-se o modelo (2.8), as probabilidades preditas pelo modelo dos logitos generalizados são dadas por:

$$\hat{\pi}_j(\boldsymbol{x}) = \frac{\exp\left(\hat{\lambda}_j + \hat{\boldsymbol{\beta}}_j \boldsymbol{x}\right)}{1 + \sum_{j=1}^{J-1} \exp\left(\hat{\lambda}_j + \hat{\boldsymbol{\beta}}_j \boldsymbol{x}\right)},$$

,

em que, $j = 1, \cdots, J-1$ e $\sum_{j=1}^{J} \pi_(\boldsymbol{x}) = 1$.

Agora, se a reposta de interesse for politômica ordinal, existem outros modelos que podem ser utilizados, como por exemplo, o modelo de probabilidade cumulativa e o modelo de chances proporcionais (Agresti 2007).

2.5.2 Modelo Dirichlet Multinomial

Uma maneira de se tratar o problema da superdispersão em dados categorizados multinomiais é por meio do modelo Dirichlet-multinomial (Morel e Nagaraj, 1992). Assume-se então, a distribuição descrita na Seção 2.4.3 para o vetor de variável resposta. Tal como na Seção 2.5.1, seja também $\boldsymbol{x} = (matx_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$ o vetor de variáveis explanatórias, e $\pi_j(\boldsymbol{x}) = P(\mathrm{Y} = j|\boldsymbol{x})$ a probabilidade de ocorrência na j-ésima categoria, com $\sum_{j=1}^{J} \pi_j(\boldsymbol{x}) = 1$ e $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jp})^T$ o vetor de parâmetros desconhecidos e de interesse, em que $j = 1, \cdots, J$. Para incorporar o efeito da variáveis explanatórias utiliza-se a função de ligação log-linear, dada por:

$$\boldsymbol{\eta} = \ln(\alpha_j) = \ln(\lambda_j + \boldsymbol{\beta}_j^T \boldsymbol{x}),$$

sendo que $\alpha_j = \exp(\lambda_j + \boldsymbol{\beta}_j^T \boldsymbol{x})$. Dessa forma o modelo Dirichlet-multinomial é expresso por:

$$\pi_j(\boldsymbol{x}) = \frac{\alpha_j(\boldsymbol{x})}{\sum_{j=1}^{J} \alpha_j(\boldsymbol{x})}, \quad (2.10)$$

em que $j = 1, \cdots, J$.

Verifica-se, que em comparação com o modelo dos logitos generalizados, o modelo Dirichlet-multinomial não compara cada categoria de resposta com uma categoria de referência.

Com relação à estimação dos parâmetros do modelo (2.10) ela pode ser realizada por meio do método da máxima verossimilhança, no qual o logaritmo da função de verossimilhança é definido por (Chen e Li, 2013):

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{x}, \boldsymbol{n}) = \sum_{i=1}^{n} \left[\ln \Gamma \left(\sum_{j=1}^{J} \boldsymbol{\beta}_j(\boldsymbol{x}_i) \right) - \ln \Gamma \left(\sum_{j=1}^{J} n_{ij} + \sum_{j=1}^{J} \boldsymbol{\beta}_j(x_i) \right) + \sum_{j=1}^{J} \ln \Gamma(n_{ij} + \boldsymbol{\beta}_j(\boldsymbol{x}_i)) \right.$$
$$\left. - \ln \Gamma(\boldsymbol{\beta}_j(\boldsymbol{x}_i)) \right].$$

Segundo Paul *et al.* (1989), pode-se reparametrizar a equação (2.7) (função de probabilidade da distribuição Dirichlet-multinomial) considerando $\gamma = \frac{1}{\sum_{j=1}^{J} \alpha_j}$. Nesse contexto, se $\gamma = 0$, o modelo Dirichlet-multinomial reduz-se a modelo multinomial. O parâmetro γ quando positivo é característico do modelo Dirichlet-multinomial, sendo usual quando há superdispersão de dados. Assim, um teste de razão de verossimilhanças para discriminar entre as duas estruturas de modelos tem como hipótese:

$$\begin{cases} H_0 : \gamma = 0 \ \text{(modelo multinomial).} \\ H_a : \gamma > 0 \ \text{(modelo Dirichlet-multinomial).} \end{cases}$$

A estatística do teste de razão de verossimilhanças é dada por:

$$\text{TRV} = -2\ln(\ell_1 - \ell_0), \tag{2.11}$$

em que ℓ_0 é o logaritmo da função de verossimilhanças sob a hipótese nula, e ℓ_1 representa o logaritmo da função de verossimilhanças sob a hipótese alternativa. Sob a hipótese nula verdadeira, TRV tem distribuição qui-quadrado χ^2_g, em que g representa o o número de graus de liberdade que é igual a diferença do número de parâmetros (np) do modelos multinomial e do modelo Dirichlet-multinomial.

Freitas (2001) salienta que, se o teste de hipóteses for significativo há indícios da presença de superdispesão.

3 MATERIAIS E MÉTODOS

3.1 Materiais

3.1.1 Estudo de motivação I

Como um primeiro exemplo de motivação considera-se um experimento desenvolvido por Castro (2016) no período de março a julho de 2014 com suínos machos, realizado em delineamento inteiramente casualizado.

O objetivo foi avaliar o comportamento desses animais expostos a duas condições de criação: com enriquecimento (CE) e sem enriquecimento (SE). O ambiente enriquecido foi onde as baias foram equipadas com diferentes objetos como correntes suspensas e recipiente plástico suspenso e solto. E a ausência deste fator são baias que não contém nenhum objeto.

Neste experimento utilizou-se um grupo de animais na fase de crescimento, correspondendo a aproximadamente 90 dias, com um total de 128 suínos machos, divididos ao acaso em 8 baias contendo 16 animais em cada baia. Neste trabalho considera-se uma análise dos dados referente ao décimo segundo dia às dez horas, no qual a variável resposta refere-se ao comportamento animal, com as categorias: "deitado", "em pé" e "sentado". Adicionalmente, para efeito de ilustração, considera-se um conjunto de dados simulado com base nas probabilidades obtidas pelo experimento real, contendo 60 baias cada uma com 20 animais, sendo que 30 baias foram submetidas a condição de criação com enriquecimento e o restante ao fator sem enriquecimento. Esse conjunto foi simulado induzindo um efeito de superdispersão.

3.1.2 Estudo de motivação II

Como um segundo estudo de motivação considera-se parte de um experimento desenvolvido por Voigt (2013), realizado durante o ano de 2011 conduzido em uma casa de vegetação. O experimento foi realizado no delineamento inteiramente casualizado, envolvendo a laranjeira da variedade "x11", que tem como principal característica o fato de apresentar período juvenil curto.

O objetivo deste experimento foi avaliar o florescimento de plantas adultas desta variedade quando enxertadas com os porta-enxertos limão "Cravo" e citrumelo "*Swingle*", durante o período de um ano, sendo que nove plantas foram exertadas sob o porta-enxerto limão "Cravo" e sete sob o citrumelo "*Swingle*".

Para efeito de aplicação, considera-se neste trabalho, os dados referentes à estação Inverno. A variável resposta refere-se à contagem para cada classificação dos ramos das plantas em três categorias mutuamente exclusivas: ramos terminais, ramos laterais, ramos sem flor ou abortada.

3.2 Métodos

Considerando os experimentos descritos na Seção (3.1), para estabelecer notação seja Y_{ijk} a variável resposta da i-ésima observação, na j-ésima categoria, no k-ésimo tratamento seguindo a distribuição multinomial, com função de ligação canônica (logito), no qual o modelo é dado por:

$$\eta_{jk} = \ln\left(\frac{\pi_{jk}}{\pi_{JK}}\right) = \lambda_j + \beta_{jk}\text{tratamento}_k, \tag{3.1}$$

em que β_{jk} é o parâmetro associado ao efeito de tratamento, com $j = 1, 2, 3$ e $k = 1, 2$.

No primeiro estudo de motivação, para o conjunto de dados real e simulado, fixou-se como categoria de referência no processo de estimação a classificação de comportamento "sentado". E para o segundo estudo de motivação fixou-se a classificação de ramos terinais.

A verificação do efeito de tratamento do modelo (3.1) é realizada por meio do teste da razão de verossimilhança para modelos encaixados. As hipóteses testadas são:

$$\begin{cases} H_0 : \boldsymbol{\eta} = \lambda_j \\ H_a : \boldsymbol{\eta} = \lambda_j + \beta_{jk}\text{tratamento}_k \end{cases}$$

A estatística do teste da razão verossimilhança é dada por:

$$\text{TRV} = -2\ln\left[\frac{L_{H_0}}{L_{H_a}}\right], \tag{3.2}$$

em que L_{H_0} é o logaritmo da função de verossimilhança sob a hipótese nula (modelo sem efeito de tratamento), e L_{H_a} representa o logaritmo da função de verossimilhança sob a hipótese alternativa (modelo que apresenta o efeito de tratamento). Sob a hipótese nula o teste da razão de verossimilhança tem distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros (np) dos modelos a serem comparados.

O modelo multinomial requer algumas condições, como o fato de que a variância observada esteja próxima da esperada pelo modelo. Porém como já citado, pode-se ocorrer casos em que a variância observada excede a ajustada pelo modelo, e este fato pode ser um indicativo da presença da superdispersão como discutido na Seção (2.3). Para o qual, a caracterização deste fenômeno são feitas por medidas decorrentes, como a *deviance* residual, bem como a comparação das variâncias observadas e ajustadas pelo modelo assumido. Constatada a presença da superdispersão faz-se necessário o uso de modelos que levam em conta a variação extra-multinomial, sendo assim, para a mesma estrutura do preditor linear do modelo (3.1) é considerada a distribuição Dirichlet-multinomial.

O critério utilizado para verificar qual o melhor modelo a ser ajustado é feito por meio do teste de hipótese baseado na razão de verossimilhança proposto por Paul *et al.* (1989), como descrito na Seção (2.5.2). A verificação da qualidade do ajuste do modelo com diferentes distribuições é realizada por meio do *half-normal plot* disponível no pacote hnp (Moral, Hinde e Demétrio, 2017).

As análises são realizadas com o auxílio do *software* R (R Core Team, 2018), por meio do pacote "nnet" (Vanables e Ripley, 2013) e do pacote "MGLM"(Kim, Zhang e Zhou, 2018) para o ajuste dos modelos. As estimativas dos parâmetros são realizadas pelo método da máxima verossimilhança.

3.2.1 Índice de Dispersão

Na Seção (2.3) apresentaram-se dois índices de dispersão para dados de contagem, propostos por Wilkemann (1995) e Ridout, Demétrio e Hinde (1998).

Nessa seção, apresenta-se uma proposta de índice de dispersão para dados politômicos nominais, sendo construída como se segue:

1. Considere uma variável resposta $\boldsymbol{Y}$ que segue a distribuição multinomial. Calcula-se um índice de dispersão para cada categoria:

$$\mathrm{ID}_j = \frac{\mathrm{Var}_j(\text{Observada})}{\mathrm{Var}_j(\text{Esperada})}$$

2. Por fim, calcula-se a média dos índices dispersão obtidos para cada uma das categorias:

$$\phi = \frac{\sum_{j=1}^{J} \mathrm{ID}_j}{\mathrm{J}} \tag{3.3}$$

em que, J será o número total de categorias.

Assim, se $\phi > 1$ indica que os dados estão superdispersos e se $\phi = 1$ diz-se que os dados estão equidispersos, ou seja, variância observada se aproxima da ajustada pelo modelo.

Com a finalidade de se averiguar o comportamento do índice de dispersão proposto neste trabalho, realizou-se um estudo de simulação (Seção 4).

4 ESTUDO DE SIMULAÇÃO

A fim de avaliar a performance do índice de dispersão proposto foi realizado um estudo de simulação inicial.

Para a simulação usou-se como motivação o estudo apresentado na Seção (3.1.1). Assim, considera-se N=60 como sendo o número de grupos, $n = 20$ representa o número de indivíduos por grupo, $j = 1, 2, 3$ indica o número de categorias e $k = 1, 2$ indica os níveis de tratamento.

Para a distribuição multinomial os vetores de probabilidades utilizados no processo de simulação associados a cada nível de tratamento foram $\pi_{1j} = (0,66; 0,28; 0,06)$ e $\pi_{2j} = (0,59; 0,37; 0,03)$. Para a distribuição Dirichlet-multinomial adicinalmente, foram fixados os parâmetros $\alpha_{1j} = (0,1; 0,4; 0,1)$ e $\alpha_{2j} = (0,2; 0,5; 0,3)$. As simulações foram realizadas utilizando o *software* R (R Core Team, 2018).

Com base nestes parâmetros fixados, foram feitas simulações sob dois cenários: dados idealizados com esquidispersão e superdispersão. No qual simulou-se mil conjunto de dados para cada cenário considerado.

Para o primeiro cenário fixou-se os vetores de probabilidade da distribuição multinomial, no qual os conjuntos de dados foram simuladas por meio da função rmultinom ( ). No segundo cenário, simulou-se o conjunto de dados por meio da função rdirm ( ), fixando-se os parâmetros da distribuição Dirichlet-multinomial.

Para cada conjunto simulado os modelos multinomial e Dirichlet-multinomial foram ajustados. Sendo possível calcular para cada conjunto de dado simulado a variância observada e a ajustada pelo modelo multinomial e o índice de dispersão, no qual o índice será a razão da variância observada pelo conjunto de dado simulado em estudo sob a variância esperada pelo modelo multinomial.

Tabela 4.1: Medidas descritivas referente ao índice de dispersão com base nos dados simulados, de acordo com cada cenário considerado no processo de estimação.

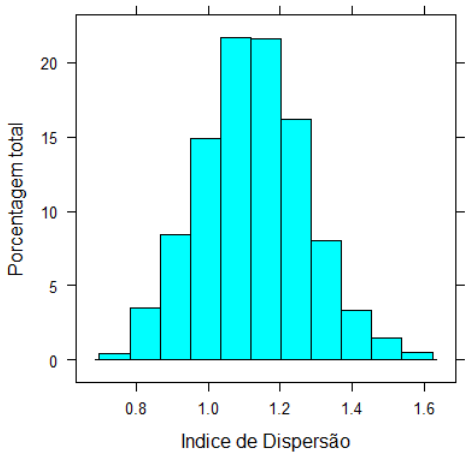
(a) Cenário 1: conjuntos de dados simulados idealizado com equidispersão

Máximo	1,59
Mínimos	0,73
Amplitude	0,86
Média	1,12
Desvio padrão	0,15

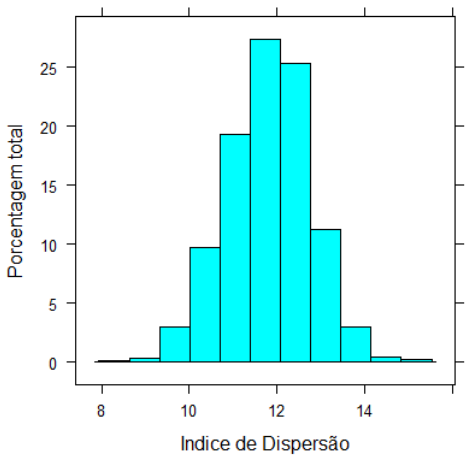
(b) Cenário 2: conjuntos de dados simulados idealizado com superdispersão

Máximo	15,25
Mínimos	8,22
Amplitude	7,03
Média	11,82
Desvio padrão	0,95

Observou-se que quando se ajusta o modelo multinomial para ambos os cenários, por meio do primeiro cenário, a Tabela 4.1(a) indicou que o índice de dispersão (ϕ) varia aproximadamente de 0,7 a 1,5, para dados equidispersos. Por outro lado, no segundo cenário, verificou-se que o índice de dispersão varia aproximadamente de 8 a 15 conforme mostra a Tabela 4.1(b), para dados superdispersos. Com base nos histogramas apresentados na Figura 4.1 (a) e (b), notou-se que a distribuição do índice de dispersão tem padrão de uma distribuição normal.



(a) Histograma do índice de dispersão para o primeiro cenário.



(b) Histograma do índice de dispersão para o segundo cenário.

Figura 4.1: Histograma referente ao índice de dispersão com base nos conjuntos de dados simulados acordo com os dois cenários considerados no processo do estudo de simulação.

Comparando as simulações realizadas pelos dois processos, observou-se que a média do índice de dispersão obtido pelo segundo cenário é aproximadamente 10,5 vezes maior do que a obtida pelo primeiro. Considerando-se o desvio padrão, tem-se que este é aproximadamente seis vezes maior para o segundo cenário quando comparado ao obtido no primeiro cenário, evidenciando para o segundo cenário quando se considera o modelo multinomial este apresenta um valor alto de desvio padrão, indicando que há uma variação extra entre os índices de dispersão.

5 RESULTADOS E DISCUSSÃO

5.1 Estudo de motivação I: dados reais

Apresenta-se inicialmente uma análise exploratória com o objetivo de compreender o comportamento dos dados, por meio das medidas descritivas apresentadas na Tabela 5.1.

Tabela 5.1: Resumo descritivo em relação ao número de animais de acordo com a classificação de comportamento dos suínos baseado na condição de criação de acordo com o experimento desenvolvido por Castro (2016), no décimo segundo dia, às dez horas.

	Condições de criação			
	Com enriquecimento		Sem enriquecimento	
Comportamento	Média	Variância	Média	Variância
Deitado	9,50	8,33	10,50	8,33
Em pé	6,00	6,00	4,50	13,67
Sentado	0,50	1,00	1,00	0,67

Os resultados preliminares da Tabela 5.1 mostram que há em média um maior número de animais na classificação de comportamento "deitado", seguidos das classificações "em pé" e "sentado", para ambas as condições de criação. Em seguida, ajusta-se o modelo multinomial.

Tabela 5.2: Teste da Razão de verossimilhanças para o efeito de tratamento (com e sem enriquecimento) por meio do modelo dos logitos generalizados, em que np = número de parâmetros, ℓ = log-verossimilhanças e TRV = Estatística do teste da razão de verossimilhanças.

Modelos	Preditor linear	np	ℓ	TRV	p-valor
1	$\eta_j = \lambda_j$	2	-30,57		
2	$\eta_j = \lambda_j + \beta_{jk}\text{tratamento}_k$	4	-29,70	1,73	0,4

De acordo com o teste da razão de verossimilhanças (TRV), Tabela 5.2, considerando um nível de 5% de significância, nota-se que as condições de criação não interferem na classificação de comportamento do animais, ou seja, não houve efeito de tratamento.

Tabela 5.3: Variâncias observadas e obtidas após o ajuste do modelo multinomial para o fator enriquecimento com relação a classificação de comportamento dos suínos, de acordo com o experimento desenvolvido por Castro (2016), no décimo segundo dia, às dez horas.

	Condições de criação			
	Com enriquecimento		Sem enriquecimento	
Comportamento	Variância Observada	Variância Ajustada	Variância Observada	Variância Ajustada
Deitado	8,33	6,77	8,33	7,24
Em pé	6,00	6,06	13,67	7,03
Sentado	1,00	1,76	0,67	0,91

As variâncias observadas e preditas pelo modelo, Tabela 5.3, constata-se que estas estão bem próximas, exceto para o caso da classificação de comportamento "em pé" sob a condição de criação com enriquecimento. Ademais, nota-se que o valor do índice de dispersão é $\phi = 1,8$ e que o valor da *deviance* residual $(30,70)$ excedeu o número de graus de liberdade do resíduo (14). Porém, o índice de dispersão ainda é próximo de um.

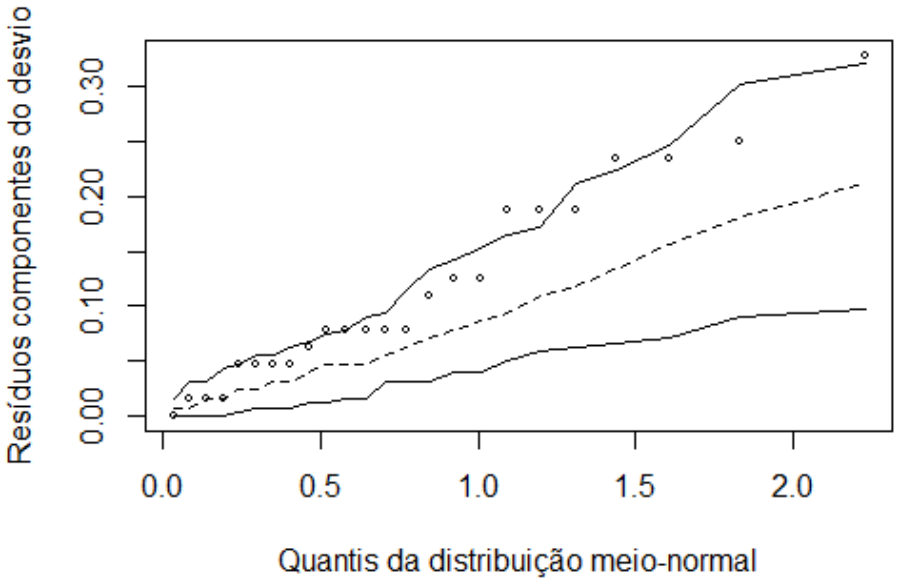


Figura 5.1: Gráficio de diagnóstico (*half-normal plot*) para avaliar a qualidade do ajuste do modelo dos logitos generalizados com efeito de tratamento (com e sem enriquecimento) aplicado aos dados do estudo de comportamento animal.

Por meio do gráfico *half-normal plot* (Figura 5.1), verifica-se que a maioria doa encontram-se dentro do envelope de simulação, indicando razoável ajuste do modelo multinomial.

5.2 Estudo de motivação I: dados simulados

Nesta seção apresentam-se resultados para um conjunto de dado simulado com base nas probabilidades obtidas pelo experimento real. Este conjunto de dados foi escolhido ao acaso por meio da simulação realizada considerando o cenário idealizado com superdispersão (Seção 4). Considerou-se 60 baias cada uma contendo 20 animais, sendo 30 baias submetidos a condição de criação com enriquecimento e o restante ao fator sem enriquecimento.O objetivo destas análises com dados simulados é ampliar a discussão entre duas classes de modelos apresentadas neste trabalho. Inicia-se com uma análise exploratória, seguindo os mesmos procedimentos da análise estabelecida pelos dados reais, com base nas medidas descritivas apresentadas na Tabela 5.4.

Tabela 5.4: Análise descritiva dos dados em relação ao número de animais de acordo com a classificação de comportamento dos suínos baseado na condição de criação em um conjunto de dados simulados com 20 animais por baia.

	Condições de criação			
	Com enriquecimento		Sem enriquecimento	
Comportamento	Média	Variância	Média	Variância
Deitado	1,77	15,90	3,60	27,42
Em pé	15,53	45,22	10,06	37,71
Sentado	2,27	25,09	5,93	37,65

Por meio da Tabela 5.4, verificou-se que há em média um maior número de animais na

classificação de comportamento “em pé”, seguidos dos que foram classificados como “sentado” e “deitado”. Observou-se também que a variância é maior que a média, para todas as condição de criação e de classificação de comportamento dos suínos. Em seguida, ajustou-se o modelo multinomial.

Tabela 5.5: Teste da Razão de verossimilhanças para o efeito de tratamento (com e sem enriquecimento) por meio do modelo dos logitos generalizados, em que np = número de parâmetros, ℓ = log-verossimilhanças e TRV = Estatística do teste da razão de verossimilhança.

Modelos	Preditor linear	np	ℓ	TRV	p-valor
1	$\eta_j = \lambda_j$	2	-635,83		
2	$\eta_j = \lambda_j + \beta_{jk}\text{tratamento}_k$	4	-583,11	105,43	<0,01

Pelo teste da razão de verossimilhanças (Tabela 5.5), considerando um nível de 5% de significância, notou-se que a princípio deve-se considerar o modelo que apresentou efeito de tratamento.

Tabela 5.6: Estimativas e erros-padrões dos parâmetros em relação ao modelo com efeito do fator enriquecimento, em que SE é a condição de criação sem enriquecimento.

Parâmetro	Estimativa	Erro-padrão	p-valor
λ_1(deitado)	-0,50	0,12	<0,001
λ_2(em pé)	0,53	0,09	<0,001
β_{11}(deitado SE)	0,25	0,22	0,255
β_{21}(em pé SE)	1,40	0,16	<0,001

As estimativas dos parâmetros do modelo selecionado bem como os erros padrões são apresentados na Tabela 5.6. A um nível de 5% de significância, observa-se que os parâmetros do modelo selecionado foram estatisticamente significativos, exceto para parâmetro β_{11}(deitado SE). Considerando a classificação de comportamento “sentado” como categoria de referência observou-se que esta se diferencia da classificação de comportamento “em pé”.

Tabela 5.7: Variâncias observadas e obtidas após o ajuste do modelo multinomial incluindo a condição de criação a classificação de comportamento dos suínos, com base no conjunto de dado simulado.

	Condições de criação			
	Com enriquecimento		Sem enriquecimento	
Comportamento	Variância Observada	Variância Ajustada	Variância Observada	Variância Ajustada
Deitado	15,90	4,50	27,42	2,46
Em pé	45,22	7,49	37,71	4,91
Sentado	23,09	6,33	37,65	3,07

De acordo com o modelo selecionado, verifica-se que o valor da *deviance* residual $(1014,36)$ excedeu o número de graus de liberdade do resíduo (116). Além disso as variâncias observadas foram maiores do que as obtidas após o ajuste do modelo multinomial, conforme mostra a Tabela 5.7. Para esses dados, o índice de dispersão foi $\phi = 11,68$. Com base nestas três medidas de diagnóstico, há evidências de superdispersão nos dados.

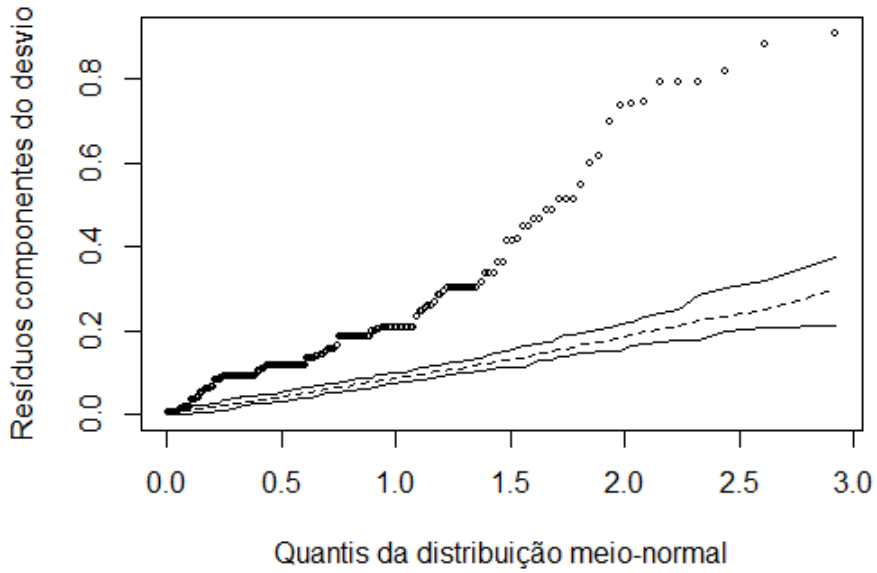


Figura 5.2: Gráfico de diagnóstico (*half-normal plot*) para avaliar a qualidade do ajuste do modelo dos logitos generalizados com efeito de tratamento (com e sem enriquecimento), com base em dados simulados.

Por meio do gráfico *half-normal plot* (Figura 5.2), verificou-se que boa parte dos pontos estão fora do envelope de simulação, evidenciando que o modelo selecionado com efeito do fator enriquecimento não se ajustou bem aos dados. Neste contexto, considerou-se o modelo Dirichlet-multinomial.

Tabela 5.8: Comparação entre os modelos multinomial e Dirichlet-multinomial, por meio dos valores do AIC, da log-verossimilhanças (ℓ), do número de parâmetro (np) e o valor do teste da razão de verossimilhanças (TRV).

Modelos	np	ℓ	AIC	TRV	p-valor
Multinomial	4	-583,11	1174,22		
Dirichlet-multinomial	6	-233,27	478,54	699,67	<0,01

O teste da razão de verossimilhanças apresentado na Tabela 5.8, indica a seleção do modelo Dirichlet-multinomial, sendo também o modelo que apresenta menor AIC.

Tabela 5.9: Variâncias observadas e obtidas após o ajuste do modelo Dirichlet-multinomial com relação ao tratamento e a classificação de comportamento dos suínos.

	Condições de criação			
	Com enriquecimento		Sem enriquecimento	
Comportamento	Variância Observada	Variância Ajustada	Variância Observada	Variância Ajustada
Deitado	15,90	30,27	27,42	28,42
Em pé	45,22	47,69	37,71	43,03
Sentado	25,09	25,08	37,65	34,96

Comparando as variâncias observadas e as obtidas após o ajuste do modelo Dirichlet-multinomial (Tabela 5.9), observou-se que estas estão próximas, além do mais, o valor do índice

de dispersão foi $\phi = 0,9$, indicando que há evidências de que o modelo Dirichlet-multinomial se ajustou bem aos dados com superdispersão.

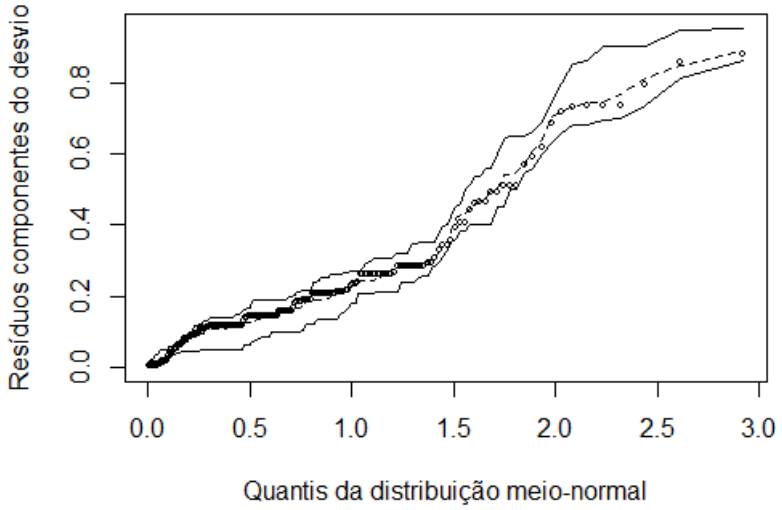


Figura 5.3: Gráfico de diagnóstico (*half-normal plot*) para avaliar a qualidade do ajuste do modelo Dirichlet-multinomial com efeito de tratamento (com e sem enriquecimento), com base em dados simulados.

Finalmente, por meio do gráfico *half-normal plot* (Figura 5.3) verificou-se que os pontos se acomodam dentro do envelope de simulação indicando que o modelo Dirichlet-multinomial se ajustou bem aos dados.

Tabela 5.10: Teste da Razão de verossimilhanças para o efeito de tratamento (com e sem enriquecimento) por meio do modelo Dirichlet-multinomial, em que np = número de parâmetros, ℓ = log-verossimilhanças e TRV = Estatística do teste da razão de verossimilhanças.

Modelos	Preditor linear	np	ℓ	TRV	p-valor
1	$\eta_j = \lambda_j$	3	-242,64		
2	$\eta_j = \lambda_j + \beta_{jk}\text{tratamento}_k$	6	-233,37	18,74	<0,01

O teste da razão de verossimilhanças (Tabela 5.10), indicou que houve efeito de tratamento, ou seja, as condições de criação a que estes animais são submetidos interferem no seu comportamento.

Tabela 5.11: Estimativas e erros-padrões dos parâmetros em relação ao modelo Dirichlet-multinomial com efeito do fator enriquecimento.

Parâmetro	Estimativa	Erro-padrão	p-valor
λ_1(deitado)	-1,16	0,26	<0,01
λ_2(em pé)	-0,27	0,27	0,31
λ_3(sentado)	-0,85	0,27	<0,01
β_{11}(deitado SE)	-1,16	0,44	<0,01
β_{21}(em pé SE)	-0,41	0,46	0,37
β_{31}(sentado SE)	-1,65	0,47	<0,01

As estimativas e erros padrões do modelo selecionado são apresentados na Tabela 5.11.

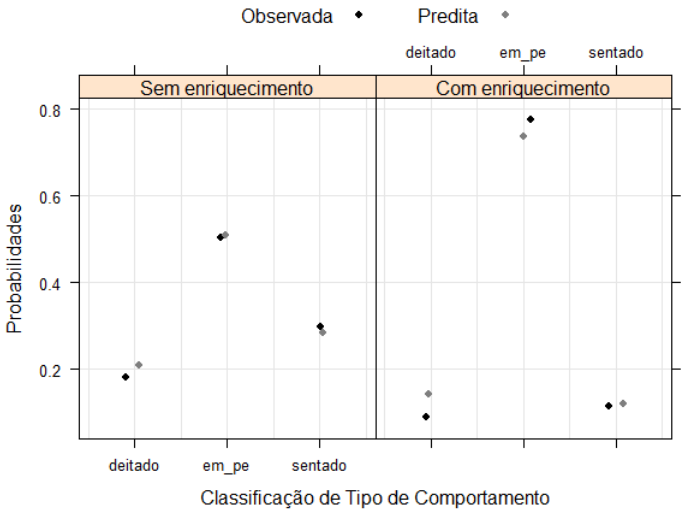


Figura 5.4: Probabilidades observadas e ajustadas pelo modelo Dirichlet-multinomial com relação ao efeito de enriquecimento com relação a classificação do comportamento dos suínos.

As probabilidades preditas e observadas pelo modelo Dirichlet-multinomial são ilustradas pela Figura 5.4. Verifica-se, com base neste cenário simulado, que a probabilidade mais favorável de ocorrência foi a classificação de comportamento "em pé", para ambas as condições de criação, seguidas pelas probabilidades de ocorrência da posição sentado e deitado. Dado que o animal foi submetido ao fator com enriquecimento, tem-se que a probabilidade de ocorrer a classificação de comportamento "em pé" é de 0,736. Porém, se considerar o fator sem enriquecimento a probabilidade de ocorrência é de 0,508. Comparando as probabilidades de ocorrência das categorias com relação as condições de criação desses animais, observa-se que há uma discrepância entre essas probabilidades sendo maior quando os animais são submetidos a condição de criação com enriquecimento.

5.3 Estudo de Motivação II

A análise descritiva para os dados do experimento desenvolvido por Voigt(2013) na estação inverno, é apresentada na Tabela 5.12.

Tabela 5.12: Resumo descritivo do número de ramos em relação à classificação, de acordo com o experimento desenvolvido por Voigt (2013) na estação inverno.

	Porta - enxertos			
	Limão "Cravo"		Citrumelo "*Swingle*"	
Classificação de ramos	Média	Variância	Média	Variância
Terminal	12,33	71,50	27,71	430,57
Lateral	89,00	939,25	70,43	496,25
Sem flor ou abortada	3,67	10,00	4,43	7,62

Verificou-se que há em média um maior número de ramos classificados como lateral, seguidos de terminal e sem flor ou abortada. Além disso notou-se que a variância é maior do que a média para ambos os porta-enxertos. A seguir ajustou-se o modelo multinomial.

Tabela 5.13: Teste da Razão de verossimilhanças para o efeito de porta-enxerto por meio do modelo dos logitos generalizados, em que np = número de parâmetros, ℓ = log-verossimilhanças e TRV = Estatística do teste da razão de verossimilhanças.

Modelo	Preditor linear	np	ℓ	TRV	p-valor
1	$\eta_j = \lambda_j$	2	-169,55		
2	$\eta_j = \lambda_j + \beta_{jk}$	4	-136,62	65,87	<0,01

De acordo como o teste da razão da verossimilhanças (TRV), considerando um nível de 5% de significância, (Tabela 5.13), notou-se que o modelo a ser considerado inicialmente é o que apresenta efeito de tratamento.

Tabela 5.14: Estimativas e erros-padrões dos parâmetros em relação ao modelo multinomial com efeito de porta-enxerto na estação inverno.

Parâmetro	Estimativa	Erro-padrão	p-valor
λ_1(lateral)	1,98	0,10	<0,01
λ_2(sem flor ou abortada)	-1,21	0,20	<0,01
β_{11}(lateral "*Swingle*")	-1,04	0,13	<0,01
β_{21}(sem flor ou abortada "*Swingle*")	-0,62	0,28	0,025

As estimativas dos parâmetros do modelo selecionado bem como os erros padrões são apresentados na Tabela 5.14. E com base na categoria de referência (ramos terminais) verificou-se que esta diferencia das demais categorias.

Tabela 5.15: Variâncias observadas e obtidas após o ajuste do modelo multinomial com efeito de porta-enxerto e a classificação de ramos, na estação inverno, de acordo com os dados do estudo de motivação II.

	Porta-enxertos			
	Limão "cravo"		Citrumelo "*swingle*"	
Classificação de ramos	Variância Observada	Variância Ajustada	Variância Observada	Variância Ajustada
Terminal	71,50	11,51	430,57	38,25
Lateral	939,25	103,46	496,95	106,08
Sem flor ou abortada	10,00	1,11	7,62	1,28

De acordo com o modelo multinomial, verificou-se que o valor da *deviance* residual (159,11) excede o número de graus de liberdade do resíduo (28). Além disso as variâncias observadas foram maiores do que as obtidas após o ajuste do modelo, conforme a Tabela 5.15. O índice de dispersão foi $\phi = 7,7 > 1$. Com base nestas três medidas de diagnósticos, há evidências de superdispersão nos dados.

Tabela 5.16: Comparação ente os modelos multinomial e Dirichlet-multinomial, por meio dos valores do AIC, da log-verossimilhanças (ℓ), do número de parâmetro (np) e o teste da razão de verossimilhanças (TRV).

Modelos	np	ℓ	AIC	TRV	p-valor
Multinomial	4	-136,62	281,24		
Dirichlet-multinomial	6	-68,34	208,68	76,56	<0,001

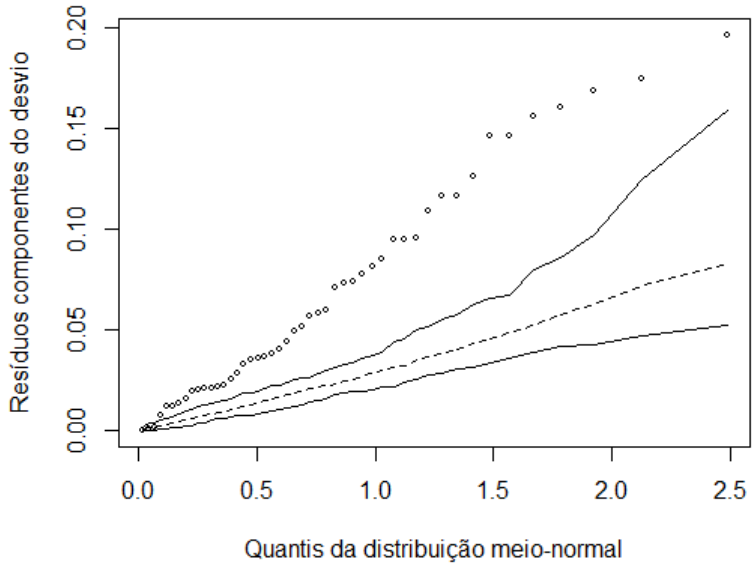


Figura 5.5: Gráfico de diagnóstico (*half-normal plot*) para avaliar a qualidade do ajuste do modelo multinomial com efeito de porta-enxerto, ajustado aos dados do estudo de motivação II.

Por meio do gráfico *half-normal plot* (Figura 5.5), verificou-se que a maior parte dos pontos encontram-se fora do envelope de simulação, indicando que o modelo selecionado não se ajustou bem aos dados. Neste contexto, considerou-se o modelo Dirichlet-multinomial.

Tanto o teste da razão de verossimilhanças quanto o AIC apresentados na Tabela 5.16, indicam a seleção do modelo Dirichlet-multinomial.

Tabela 5.17: Variâncias observadas e obtidas após o ajuste do modelo Dirichlet-multinomial com efeito de porta-enxertos e a classificação de ramos, na estação inverno, de acordo com o estudo de motivação II.

	Porta-enxertos			
	Limão "carvo"		Citrumelo "*swingle*"	
Classificação de ramos	Variância Observada	Variância Ajustada	Variância Observada	Variância Ajustada
Terminal	71,50	64,59	430,57	323,34
Lateral	939,25	3685,39	496,95	2239,60
Sem flor ou abortada	10,00	3,12	7,62	3,82

Comparando as variâncias observadas e obtidas após o ajuste do modelo Dirichlet-multinomial (Tabela 5.17), observou-se que estas estão próximas, exceto para a classificação de ramos laterais, para ambos os porta-enxertos. O índice de dispersão foi $\phi = 1,35$, evidencia que o modelo Dirichelet-multinomial foi adequado para este dados com superdispersão.

Por fim, com base no gráfico *half-normal plot* (Figura 5.6), constatou-se que os pontos se acomodam dentro do envelope de simulação indicando que o modelo Dirichlet-multinomial se ajustou bem aos dados. Dessa forma, realizou-se o teste da razão de verossimilhanças com o intuito de averiguar a presença do efeito de tratamento no modelo Dirichlet-multinomial.

Por meio do teste da razão de verrosimilhanças (Tabela 5.18), o efeito de porta-enxerto não é significativo a um nível de 5% de significância.

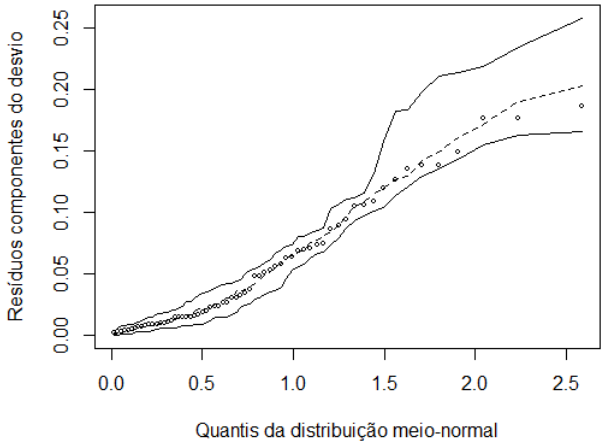


Figura 5.6: Gráfico de diagnóstico (*half-normal plot*) para avaliar a qualidade do ajuste do modelo Dirihlet-multinomial, ajustado aos dados do estudo de motivação II.

Tabela 5.18: Teste da Razão de verossimilhanças para o efeito de porta-enxerto por meio do modelo Dirichlet-multinomial, em que np = número de parâmetros, ℓ = log-verossimilhanças e TRV = Estatística do teste da razão de verossimilhanças.

Modelo	Preditor linear	np	ℓ	TRV	p-valor
1	$\eta_j = \lambda_j$	3	-102,15		
2	$\eta_j = \lambda_j + \beta_{jk}$	6	-98,33	7,63	0,0543

Tabela 5.19: Estimativas e erros-padrões dos parâmetros em relação ao modelo Dirichlet-multinomial com efeito de porta-enxerto na estação inverno.

Parâmetro	Estimativa	Erro-padrão	p-valor
λ_1(lateral)	2,977	0,457	<0,01
λ_2(sem flor ou abortada)	0,004	0,442	0,993
λ_3(terminal)	1,038	0,445	0,020
β_{11}(lateral "*Swingle*")	-0,206	0,677	0,761
β_{21}(sem flor ou abortada "*Swingle*")	0,274	0,646	0,671
β_{31}(terminal "*Swingle*")	0,666	0,677	0,326

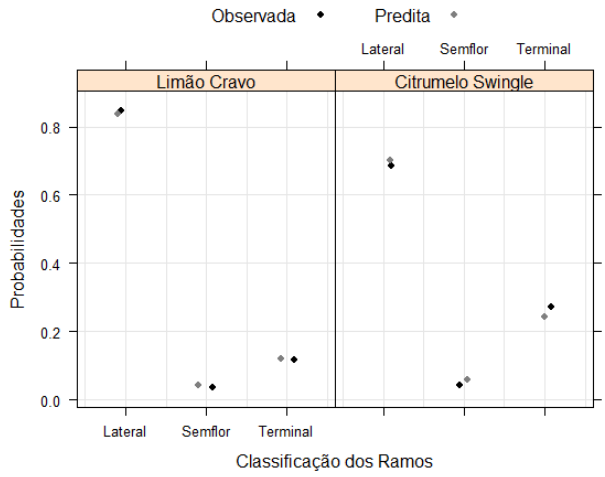


Figura 5.7: Probabilidades observadas e ajustadas pelo modelo Dirichlet-multinomial com relação aos porta-enxertos e a classificação de ramos.

As estimativas e os erros padrões são apresentados na Tabela 5.19, verificam que os parâmetros β_{jk} são não significativos, em função do efeito marginal de tratamento.

As probabilidades preditas e observadas pelo modelo Dirichlet-multinomial são ilustradas pela Figura 5.7. Verifica-se que a categoria mais favorável de vir a ocorrer é o de ramos laterais, sendo esta estimada em 0,837 para plantas com porta-enxerto limão "cravo" e 0,700 para o citrumelo "*Swingle*". A segunda categoria mais provável de ocorrer na estação inverno é a sem flor ou abortada, sendo sua probabilidade maior nas plantas com porta-enxerto citrumelo "*Swingle*", mas estas diferenças não são significativas.

6 CONSIDERAÇÕES FINAIS

Este trabalho apresenta um estudo introdutório para a análise de dados politômicos, utilizando a abordagem de dados agrupados com superdispersão em um estudo "*cross sectional*". A alternativa proposta é o uso do modelo Dirichlet-multinomial, um modelo em dois estágios, que apresenta parâmetro adicional em comparação ao multinomial, permitindo acomodar a variabilidade extra.

Este trabalho também apresenta uma proposta de índice de dispersão como uma medida de diagnóstico de superdispersão em dados politômicos nominais, em que se avaliou sua performance por meio do estudo de simulação inicial.

Tanto o diagnóstico da superdispersão quanto a escolha do modelo apropriado são importantes para evitar conclusões errôneas. Como por exemplo na análise dos dados relativos ao estudo de motivação II, se não se considera a presença da superdispersão, pode-se concluir com o uso do modelo dos logitos generalizados que o efeito de tratamento é significativo. No entanto com o modelo Dirichlet-multinomial este fato não ocorre.

Apesar do modelo Dirichlet-multinomial ter apresentado um ajuste satisfatório e da evidência do índice de dispersão inicialmente ter apresentado uma performance satisfatória, este trabalho apresenta apenas um estudo inicial para a modelagem de dados categorizados multinomiais com superdispersão, que são fontes de pesquisas futuras.

Como perspectivas de outros trabalhos pretende-se realizar um estudo mais detalhado do índice de dispersão proposto avaliando sua performance por meio do intervalo de confiança. Além de dar ênfase para a abordagem de outros modelos tais como: modelos log-lineares, o modelo multinomial negativo e o Dirichlet-multinomial generalizado. Também pretende-se estender a problemática da superdispersão considerando estudos longitudinais e a dados politômicos ordinais.

REFERÊNCIAS

Agresti, A., 2002 *Categorical Data Analysis*. New York, second edition.

Agresti, A., 2007 *An Introduction to Categorical Data Analysis*. New York, second edition.

Atkinson, A. C. A. C., 1985 *Plots, transformations, and regression : an introduction to graphical methods of diagnostic regression analysis*. n Oxford Statistical Science Series.

Castro, A. C. D., 2016 *Comportamento e desempenho sexual de suínos reprodutores criados em ambientes enriquecidos*. Ph.D. thesis, Universidade de São Paulo/ Escola Superior de Agricultura " Luiz de Queiroz ".

Chen, J. and H. Li, 2013 Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Annals of Applied Statistics **7**: p.418–442.

Connor, R. J. and J. E. Mosimann, 1969 Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. Public Health **64**: 194– 206.

Cordeiro, G. M., 2007 Modelos Lineares Generalizados Minicurso para o 12° SEAGRO e a 52ª Reunião Anual da RBRAS. p. 161p., Santa Maria, RS, UFSM.

Cordeiro, G. M. and C. G. Demétrio, 2008 Modelos Lineares Generalizados e Extensões. p. 392p., Piracicaba - SP.

Demétrio, C., 2002 Modelos Lineares Generalizados em Experimentação Agronômica.

Dobson, A. J., 2010 *An Introduction to generalized linear models*. Chapman & Hall/CRC, New York, second edition.

Freitas, S. M., 2001 *Modelos para proporções com superdispersão provenienetes de ensaios toxicológicos no tempo*. Ph.D. thesis, Universidade de São Paulo / Escola Superior Agricola "Luiz de Queiroz".

Giolo, S. R., 2017 *Introdução à análise de dados categóricos com aplicações*. São Paulo.

Hinde, J. and C. G. B. Demétrio, 1998 Overdispersion : Models and Estimation. In *A Short Course for SINAPE 1998*, p. 73p.

Hinde, J. and C. G. B. Demtrio, 1998 Overdispersion : Models and estimation. Computational Statistics & Data Analysis **27**: p.151–170.

Kim, J., Y. Zhang, J. Day, and H. Zhou, 2018 MGLM : An R Package for Multivariate Categorical Data Analysis **10**: p.73–90.

McCullagh, P. P. and J. A. Nelder, 1989 *Generalized linear models*. Chapman and Hall, second edition.

Moral, R. A., J. Hinde, and C. G. B. Demétrio, 2017 Half-Normal Plots and Overdispersed Models in R : The hnp Package. Journal of Statistical Software **81**: 23p.

Moral, R. d. A., 2013 *Modelagem estatística e ecológica de relações tróficas em pragas e inimigos naturais*. Ph.D. thesis, Universidade de São Paulo, Piracicaba.

Moral, R. d. A., J. Hinde, and C. G. B. Demétrio, 2018 Half-Normal Plots with Simulation Envelopes.

Morel, J. G. and N. K. Nagaraj, 1992 A Finite Mixture Distribution for Modelling Multinomial Extra Variation. Biometrika Trust, Oxford University Press **80**: p.363–371.

Mosimann, J. E., 1962 On the Compound Multinomial Distribution , the Multivariate β-Distribution , and Correlations Among Proportions. Biometrika Trust, Oxford University Press **49**: p.65–82.

Myers, R. H., D. C. Montgomery, G. G. Vining, and T. J. Robinson, 2010 *Generalized linear models : with applications in engineering and the sciences*. John Wiley &Sons.

Nelder, J. A. and R. W. M. Wedderburn, 1972 Generalized Linear Models **135**: p.370–384.

Olsson, U., 2002 *Generalized Linear Models An Applied Approach*. Lund: Studentlitteratur.

Paul, S. R., K. Y. Liang, and S. G. Self, 1989 On Testing Departure from the Binomial and Multinomial Assumptions **45**: p.231–236.

Paula, G. A., 2004 MODELOS DE REGRESSÃO com apoio computacional. Technical report.

Pregibon, D., 1981 Logistic Regression Diagnostics. The Annals of Statistics **9**: p.705–724.

Ridout, M., C. G. B. Demétrio, and J. Hinde, 1998 Models for count data with many zeros. In Proceedings of the XIXth international biometric conference **19**: p.179–192.

Venables, W. N. and B. D. R. Springer, 2002 *Modern Applied Statistics with S*. Fourth edition.

Vieira, A., 2008 *Modelagem simultânea de média e dispersão e aplicações na pesquisa agronômica*. Ph.D. thesis, Universidade de São Paulo - USP/ ESALQ.

Voigt, V., 2013 *Caracterização fenotípica e avaliação da expressão de genes envolvidos na indução e no florescimento da laranjeira 'x11'*. Ph.D. thesis, Universidade de São Paulo.

Wedderburn, R. W. M., 1974 Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. Biometrika **61**: p.439–447.

Williams, D. A., 1987 Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. Applied Statistics **36**: p.181–191.

Winkelmann, R., 1995 Duration dependence and dispersion in count-data models. Journal of Business & Economic Statistics **13**: p.467–474.

ANEXOS

Linhas de Comando R

```
# Pacotes utilizados
library(MGLM)
library(tidyr)
library(dplyr)

# Configurações da simulação
n <
k <
p1 < c()
p2 < c()
B < 1000

# Cria matrizes dos resultados
lls < matrix(ncol = , nrow = B)
nps < matrix(ncol = , nrow = B)
aics < matrix(ncol = , nrow = B)
bics < matrix(ncol = , nrow = B)
chisq_stat < matrix(ncol = , nrow = B)
chisq_pval < matrix(ncol = , nrow = B)
IS < matrix(ncol = , nrow = B)
meIS < matrix(ncol = , nrow = B)
aux <

for (i in 1:B) {
#Conjunto de dados simulados
y1 < t(rmultinom(n, k, p1))
y2 < t(rmultinom(n, k, p2))
#Nomeando as Colunas
colnames(y2) < c('categoria1', 'categoria2', 'categoria3')
colnames(y1) < c('categoria1', 'categoria2', 'categoria3')
# Montando o data set
tratamento < c(rep('tratamento1', 30), rep('tratamento2', 30))
dados_simulados1 < data.frame(tratamento, rbind(y1, y2))
dados < data.frame(rbind(y1, y2))

# Ajusta o modelo multinomial
modelMN1 < try(MGLMreg(formula =
            cbind(categoria1, categoria2, categoria3) ~ tratamento,
                        data = dados_simulados1,
```

```
                                dist = "MN",
                                LRT = FALSE))

# Ajusta o modelo Multinomial Dirichlet
modelDM1 <- try(MGLMreg(formula =
            cbind(categoria1, categoria2, categoria3) ~ tratamento,
                                data = dados_simulados1,
                                dist = "DM",
                                LRT = FALSE))

# Obtem medidas de goodness of fit
t <- system.file(
  modelo <- try(list("Multinomial" = modelMN1,
                     "DirichketMult" = modelDM1)))
  if (! (class(modelMN1) == "try error" ||
         class(modelDM1) == "try error")) {
  lls[i, ] <- sapply(modelo, logLik)
  nps[i, ] <- sapply(modelo, function(m) length(m@coefficients))
  aics[i, ] <- sapply(modelo, AIC)
  bics[i, ]<- sapply(modelo, BIC)
  chisq_stat[i]<- 2*diff(as.numeric(as.vector(lls[i,])))
  chisq_pval[i] <- pchisq(as.numeric(as.vector(chisq_stat[i])),
  diff(as.numeric(as.vector(nps[i,]) )),
                              lower.tail = FALSE)

  } else {
  cont = aux+1
  lls[i, ] <- c( logLik(modelMN1), "NA")
  nps[i, ] <- c( length(modelMN1@coefficients), "NA")
  aics[i,] <- c( AIC(modelMN1), "NA")
  bics[i,] <- c( BIC(modelMN1), "NA")
  chisq_stat[i] <- c("NA")
  chisq_pval[i] <-  c("NA")
    }

  IS[i,] <- apply(dados, 2, function(y) {
    varEsp <-  mu * (m  mu) / m
    varObs <- var(y)
    varObs/varEsp #indice de superdispersao
    })
    meIS[i] <- mean(IS[i,])
}
```