Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Modelos para Dados de Contagem com Aplicações.

Clarice Camargo Mendes
Orientadora: Profa. Dra. Hildete Prisco Pinheiro

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, para obtenção do Título de Mestre em Estatística.

Campinas - SP
2007

Modelos para Dados de Contagem com Aplicações

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Clarice Camargo Mendes e aprovada pela comissão julgadora.

Campinas, 3 de maio de 2007.

Hildete Pinheiro

Profa. Dra. Hildete Prisco Pinheiro
Orientadora

Banca examinadora:

1. Profa. Dra. Hildete Prisco Pinheiro (Orientadora) - IMECC/UNICAMP
2. Prof. Dr. Gilberto Alvarenga Paula - IME/USP
3. Profa. Dra. Mariana Rodrigues Motta

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica, UNICAMP , como requisito parcial para obtenção do Título de Mestre em Estatística.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Bibliotecária: Maria Júlia Milani Rodrigues – CRB8a / 2116

Mendes, Clarice Camargo

M522m Modelos para dados de contagem com aplicações / Clarice Camargo Mendes -- Campinas, [S.P. :s.n.], 2007.

Orientador : Hildete Prisco Pinheiro

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Regressão de Poisson. 2. Modelos lineares generalizados. 3. Modelos ZIP. I. Pinheiro, Hildete Prisco. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título. (mjmr/imecc)

Título em inglês: Models for count data with applications.

Palavras-chave em inglês (Keywords): 1. Poisson regression. 2. Generalized linear models.

Área de concentração: Bioestatística

Titulação: Mestre em Estatística

Banca examinadora: Profa. Dra. Hildete Prisco Pinheiro (IMECC-UNICAMP)
Prof. Dr. Gilberto Alvarenga Paula (IME-USP)
Profa. Dra. Mariana Rodrigues Motta

Data da defesa: 03/05/2007

Programa de Pós-Graduação: Mestrado em Estatística

Dissertação de Mestrado defendida em 03 de maio de 2007 e aprovada pela Banca Examinadora composta pelos Profs. Drs.

Hildete Prisco Pinheiro

Prof (a). Dr (a). HILDETE PRISCO PINHEIRO

Prof (a). Dr (a). GILBERTO ALVARENGA PAULA

Mariana Rodrigues Motta

Prof (a). Dr (a). MARIANA RODRIGUES MOTTA

Dedico este trabalho ao Valério.

Agradecimentos

Agradeço:

A Deus, pela realização de mais um sonho.

Ao meu marido Valério, por seu amor, sua ajuda constante, seu apoio incondicional e seu companheirismo, que foram absolutamente fundamentais para que eu chegasse até aqui.

Aos meus pais, Carlos Alberto (*in Memoriam*) e Lydia, por me fornecerem todas as condições para que eu completasse meus estudos até a graduação e por sempre me apoiarem em minhas escolhas.

À minha orientadora Profa. Hildete Prisco Pinheiro, pela paciência que teve ao me orientar e por seu incentivo constante.

Ao Prof. Sérgio Furtado dos Reis e seu aluno Márcio Silva Araújo pelo fornecimento do conjunto de dados que inspirou este trabalho, pelas sugestões e incentivo.

Aos professores Gilberto Alvarenga Paula e Mariana Rodrigues Motta, por aceitarem participar da banca examinadora, pelas correções e sugestões.

Aos professores do departamento de Estatística do IMECC que sempre me apoiaram e me incentivaram, especialmente Prof. Mauro Marques, Prof. Armando Infante, Profa. Nancy Garcia, Prof. Aluísio Pinheiro, Prof. Ronaldo Dias, Prof. Filidor Labra, Prof. Mário Gneri, Prof. Hotta e Profa. Marina Vachkovskaia.

A todos os funcionários do IMECC, particularmente à Tânia e aos demais funcionários da Secretaria de Pós-Graduação pelas informações e esclarecimentos.

A todos os colegas do IMECC, pela amizade e incentivo. Especialmente à Samara, Tati e Gabriel pela força e ajuda que sempre me deram.

Resumo

Ao lidarmos com dados de contagem, uma abordagem possível é estimar um Modelo Linear Generalizado com distribuição de Poisson. Freqüentemente nestes modelos costuma surgir o problema da superdispersão, um fenômeno que aparece quando estamos diante de uma variabilidade dos dados maior do que a média.

Temos basicamente três soluções para este problema: abordagem bayesiana, assumindo que o parâmetro do modelo possui uma distribuição de probabilidade; estimação por Quase-verossimilhança, incluindo um fator de dispersão diferente da unidade ou uma função de variância diversa e, finalmente, o emprego de modelos mistos, com a separação de efeitos fixos e aleatórios.

Outra ocorrência comum para dados de contagem é encontrarmos amostras que apresentem um número excessivo de zeros. Detectamos a presença da superdispersão, mas agora ela é devida à ocorrência de mais valores zero na amostra do que seria esperado para dados que seguissem a distribuição de Poisson. Para este caso Lambert (1982) apresenta a chamada regressão de Poisson inflacionada de zeros (ZIP - *Zero Inflated Poisson*).

Através de uma aplicação a dados reais, em estudo referente à alimentação de rãs da espécie *Adenomera*, identificamos os melhores modelos para explicar a quantidade de comida ingerida em função dos efeitos de sexo e da estação do ano. Utilizamos técnicas de diagnóstico para avaliar o impacto que uma determinada observação exerce na estimativa dos parâmetros.

Abstract

When one deals with count data, a possible approach is to fit a generalized linear model with Poisson distribution. Usually it may occur the problem of superdispersion, when the variability of the data is greater then the mean.

There are three basic solutions to this problem: the Bayesian approach, when we assume that the parameter of the model has a distribution of probability; the Quasi-likelihood estimation, including a non-unitary dispersion parameter or a different variance function and, finally, the mixed models.

Another possible occurrence to count data is the presence of samples with an excess of zeros. We detect the presence of the superdispersion, but now it is due to more zero counts than expected from the Poisson distribution. For this case, Lambert (1982) presents the Zero Inflated Poisson (ZIP) model.

As an application to real data, in the study of frogs' nourishment from the species *Adenomera*, we identify the best models to explain the quantity of swallowed food related to sex and season effects. We employ techniques of diagnosis to verify the impact of a specific observation in the parameter estimations.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

A principal motivação para este trabalho foi estudar os modelos empregados para a análise de dados de contagem. O conjunto de dados gentilmente cedido pelo Prof. Dr. Sérgio Furtado dos Reis e seu aluno Márcio Silva Araújo, do Instituto de Biologia da Unicamp, nos motivou a buscar na literatura quais eram as principais abordagens para modelar dados de contagem.

Iniciamos nosso trabalho com uma visão geral de quais foram as soluções apresentadas para dados de contagem antes do aparecimento dos modelos lineares generalizados. No capítulo 2 apresentamos a teoria dos modelos lineares generalizados, discutindo sua estimação e ajuste do modelo. No capítulo 3 discutimos um pouco mais os modelos ligados a dados de contagem e as soluções mais conhecidas para o problema de superdispersão. Apresentamos também o modelo conhecido como ZIP (*Zero Inflated Poisson*) que lida com amostras de contagem em que ocorre um número excessivo de zeros. No capítulo 4 discutiremos os modelos mistos, que também aparecem na literatura como solução para a superdispersão, uma vez que estes modelos permitem trabalhar de forma mais refinada a variabilidade encontrada nos dados. No capítulo 5 apresentaremos dois exemplos de dados já conhecidos na literatura. No capítulo 6 usamos as noções discutidas e propomos uma metodologia para estudar os dados referentes à alimentação de rãs da espécie *Adenomera*, ponto de partida de nosso estudo.

1.2 Antecedentes dos modelos lineares generalizados

Stigler (1986) inicia seu livro afirmando que o método dos mínimos quadrados representou para a Estatística no século XIX aquilo que o Cálculo representou para a Matemática no século XVIII. Trabalhando com medidas astronômicas e seus erros de medidas, Legendre publicou em 1805 o método dos mínimos quadrados. Outro trabalho de suma importância

para os modelos lineares foi o de Galton, em 1885, em que ele apresentou o termo "regressão" no estudo sobre hereditariedade e estatura. Os trabalhos de Legendre e Galton estavam ligados respectivamente a medidas astronômicas e biológicas, mas ambos tinham em comum a suposição da normalidade dos erros.

A teoria de modelos lineares clássicos supõe a hipótese de normalidade dos erros, mas, como afirmam McCullagh e Nelder (1989), citando Gauss, as propriedades mais importantes dos mínimos quadrados não dependem da normalidade dos erros, mas da suposição de variância constante e independência das observações.

Antes do aparecimento dos modelos lineares generalizados havia algumas técnicas que buscavam lidar com modelos lineares nos quais a variável resposta não tinha distribuição normal. As transformações nas variáveis resposta eram uma tentativa de "normalizar" os dados originais. Desta forma conseguia-se chegar mais perto da suposição da variância constante que é inerente à distribuição normal.

Diversos autores apresentam os modelos de regressão linear clássicos como Draper e Smith (1981) e Montgomery e Peck (1991) estudando os modelos básicos, qualidade do ajuste, análise de resíduos etc. e propondo algumas técnicas de transformação das variáveis para buscar a normalidade dos dados.

Uma solução mais consistente do que as transformações foi dada por Nelder e Wedderburn (1972). Eles generalizaram o emprego de modelos lineares para distribuições pertencentes à família exponencial. Modelos lineares generalizados (MLGs) dizem respeito a modelos lineares em relação a uma função da média da variável, de tal forma que podemos lidar com outras famílias de distribuição de variáveis aleatórias, especificamente, as que pertencem à família exponencial.

O desenvolvimento dos MLGs pode ser visto como uma generalização do procedimento descrito por Finney (1971) para a estimação de máxima verossimilhança do probito. A técnica apresentada por Nelder e Wedderburn (1972) unificou os problemas de estimação de máxima verossimilhança dos parâmetros dos seguintes modelos lineares de regressão: probito, logístico, Poisson, log-lineares etc. Finney (1971) trabalhou especificamente com o probito (um caso particular) e Nelder e Wedderburn (1972) generalizaram a estimação de máxima verossimilhança para a chamada "função de ligação", que "liga" a média da variável ao preditor linear.

McCullagh e Nelder (1989) enfatizam que para os MLGs as propriedades mais importantes são a relação existente entre a variância e a média e a independência entre as observações, e não tanto a forma da distribuição que está sendo analisada.

Na subseção a seguir mostraremos quais eram as transformações mais comuns utiliza-

das para tentar sanar o problema de dados com distribuições diferentes da Normal.

1.2.1 Transformações

Montgomery e Peck (1991) apresentam uma tabela de transformações na variável resposta que buscavam a estabilização da variância como na Tabela 1.1. Desta forma, podia-se continuar trabalhando com a teoria de modelos lineares clássicos. Seja Y a variável aleatória e Y' a transformação sugerida.

Tabela 1.1: Transformações para estabilizar a variância.

Relação entre σ^2 e $E(Y)$	Transformação
$\sigma^2 \propto$ constante	$Y' = Y$
$\sigma^2 \propto E(Y)$	$Y' = \sqrt{Y}$
$\sigma^2 \propto E(Y)(1 - E(Y))$	$Y' = sen^{-1}(\sqrt{Y})$
$\sigma^2 \propto [E(Y)]^2$	$Y' = \ln Y$
$\sigma^2 \propto [E(Y)]^3$	$Y' = Y^{-1/2}$
$\sigma^2 \propto [E(Y)]^4$	$Y' = Y^{-1}$

No caso da distribuição de Poisson, a variância da variável aleatória Y é igual à sua média, mas a variância da raiz quadrada de Y é independente desta. Bartlett (1936) apresenta uma justificativa para esta afirmação colocando a variância de $\sqrt{Y}$ como sendo constante. Seja Y variável aleatória com distribuição Poisson(λ) e uma amostra $Y_1, ..., Y_2$. Temos que $\hat{\sigma}^2(Y) = E(Y - \hat{\lambda})^2$ e seja uma função $f : \mathbb{R} \to \mathbb{R}$, contínua. Vamos estudar o comportamento da razão entre a variância amostral de uma função de Y e a variância amostral de Y:

$$\frac{\hat{\sigma}^2(f(Y))}{\hat{\sigma}^2(Y)} = \frac{E(f(Y) - \hat{m})^2}{E(Y - \hat{\lambda})^2} = \frac{\sum_{i=1}^{n}(f(Y_i) - \hat{m})^2}{\sum_{i=1}^{n}(Y_i - \hat{\lambda})^2}, \tag{1.1}$$

em que $\hat{m}$ é a média amostral de $f(Y)$. Agora, no caso de $f(Y) = \sqrt{Y}$ e $\lambda >> 1$, temos que a aproximação linear

$$\tilde{f}(t) = \frac{t}{2\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{2}$$

é bastante satisfatória para um largo intervalo contendo λ. Por exemplo, tomemos o intervalo $[\lambda/2; 3\lambda/2]$,

$$\tilde{f}\left(\frac{\lambda}{2}\right) - f\left(\frac{\lambda}{2}\right) = \frac{\lambda/2}{2\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{2} - \sqrt{\frac{\lambda}{2}} = \frac{3 - 2\sqrt{2}}{4}\sqrt{\lambda} = 0,043 \quad \text{e}$$

$$\tilde{f}\left(\frac{3\lambda}{2}\right) - f\left(\frac{3\lambda}{2}\right) = \frac{3\lambda/2}{2\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{2} - \sqrt{\frac{3\lambda}{2}} = \frac{5 - 2\sqrt{6}}{4}\sqrt{\lambda} = 0,025.$$

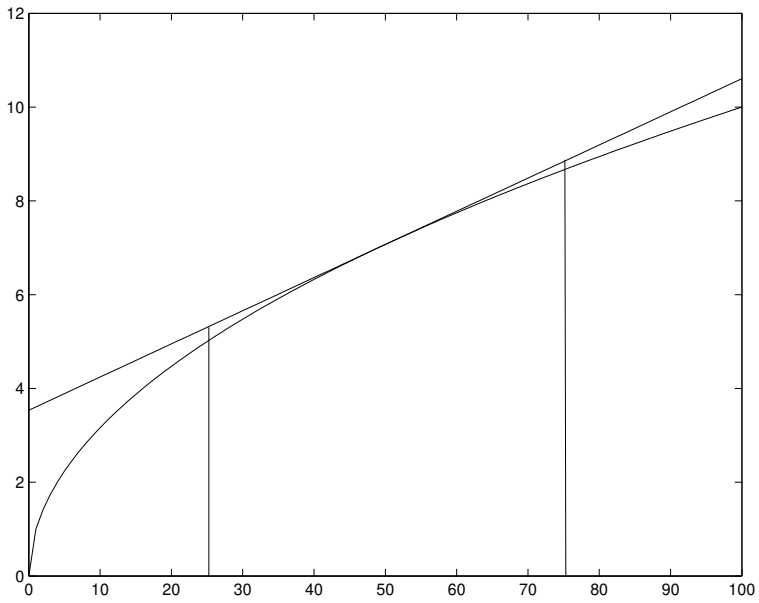

Figura 1.1: Aproximação linear para a função raiz quadrada

Ou seja, conseguimos uma boa aproximação linear para $f(t) = \sqrt{t}$, como podemos também atestar pelo Figura 1.1. Vamos conseguir uma aproximação linear para $\hat{m}$:

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} f(Y_i) \cong \frac{1}{n}\left[\frac{1}{2\sqrt{\lambda}}\sum_{i=1}^{n} Y_i + \frac{n}{2}\sqrt{\lambda}\right] = \frac{1}{2\sqrt{\lambda}}\hat{\lambda} + \frac{\sqrt{\lambda}}{2} = \tilde{f}(\hat{\lambda})$$

$$\sum_{i=1}^{n}(f(Y_i) - \hat{m})^2 \cong \sum_{i=1}^{n}(\tilde{f}(Y_i) - \tilde{f}(\hat{m}))^2 = \frac{1}{4\lambda}\sum_{i=1}^{n}(Y_i - \hat{\lambda})^2 \tag{1.2}$$

Com $f(Y) = \sqrt{Y}$ e substituindo o resultado (1.2) em (1.1) temos

$$\frac{\hat{\sigma}^2(\sqrt{Y})}{\hat{\sigma}^2(Y)} \cong \frac{1}{4\lambda}.$$

Para $Y \sim Poisson(\lambda)$ temos finalmente $\hat{\sigma}^2(\sqrt{Y}) = \frac{1}{4}$, independente de λ.

Tukey (1957) faz uma análise gráfica conjunta das possíveis transformações, o que ele chama de família de transformações que será generalizado por Box e Cox (1964). A transformação de Box e Cox lida com a correção da não-normalidade e/ou da variância não constante. A variável Y é transformada em $Y^{(\lambda)}$ segundo a fórmula

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda \dot{Y}^{\lambda-1}} \ , & \lambda \neq 0 \\ \\ \dot{Y} \ln Y \ , & \lambda = 0 \end{cases} \tag{1.3}$$

em que $\dot{Y}$ é a média geométrica das observações dada por:

$$\dot{Y} = \ln^{-1}\left(\frac{\sum_{i=1}^{n} \ln Y_i}{n}\right).$$

O modelo a ser estimado é então:

$$\hat{\mathbf{Y}}^{(\lambda)} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad \text{ou} \quad \mathbf{Y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

O método consiste em ajustar-se modelos para $Y^{(\lambda)}$ de acordo com (1.3) para diferentes valores de λ. Não se pode selecionar λ diretamente comparando a soma dos quadrados dos resíduos de $Y^{(\lambda)}$ em $\mathbf{X}$, pois para cada λ estaríamos medindo a soma dos quadrados dos resíduos em uma escala diferente. A transformação sugerida em (1.3) escalona as respostas de tal forma que as somas dos quadrados dos resíduos são diretamente comparáveis. O melhor valor de λ é o que fornece a menor soma dos quadrados dos resíduos.

Capítulo 2

Modelos Lineares Generalizados

2.1 Introdução

Consideremos um vetor de variáveis aleatórias $\mathbf{Y}$ e uma matriz $\mathbf{X}$ de valores fixos e independentes. Dizemos que um modelo é linear nos parâmetros se

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros a ser estimado e $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Um modelo é linear generalizado se existe uma função g tal que $g(E(\mathbf{Y})) = \mathbf{X}\boldsymbol{\beta}$, e a distribuição da variável $\mathbf{Y}$ pertence à família exponencial.

Por exemplo, se $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, $g(\boldsymbol{\mu}) = \boldsymbol{\mu}$ nos dará $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Então, $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\epsilon}$, onde $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Desta forma, recaímos no modelo linear, ou seja, o caso normal é um caso particular dos modelos lineares generalizados (MLGs).

Os MLGs representam assim uma classe de modelos não lineares que abrangem diferentes distribuições. Também nos MLGs são modelados os efeitos fixos das varáveis explicativas, ou regressores. Primeiramente, temos o preditor linear

$$\eta_i = x_i'\boldsymbol{\beta},$$

em que x_i' representa os regressores da i-ésima observação e $\boldsymbol{\beta}$ é o vetor de parâmetros a serem estimados, que explicam a relação existente entre os regressores e uma função da média das observações.

Em seguida, define-se a função de ligação tal que $g(\mu_i) = \eta_i$ e finalmente especifica-se a forma da variância em termos da média μ_i. Os dois últimos passos estão ligados à informação sobre a distribuição dos Y_i, que pertence à família exponencial.

Segundo McCullogh e Searle (2001), a construção de um MLG envolve 3 decisões:

1. Qual é a distribuição dos dados?
2. Qual função da média será modelada como função linear dos regressores?
3. Quais serão os regressores?

Supõe-se inicialmente que o vetor $\mathbf{Y}$ consista de medidas independentes de uma variável aleatória cuja distribuição pertença à família exponencial, ou seja:

$$Y_i \sim f_{Y_i}(y_i),\ i = 1, ..., n.$$

Uma função que comumente associa a média aos regressores é a chamada função de ligação canônica, especificada na parametrização da família exponencial por θ_i. Mas podem existir outras funções de ligação, diferentes da canônica.

Podemos escrever cada função densidade de distribuição na forma

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

em que temos a esperança das variáveis dada por $E(Y_i) = \mu_i = b'(\theta_i)$, como mostramos a seguir. Por definição,

$$E(Y_i) = \int_{-\infty}^{\infty} y_i dF(y_i).$$

A integral acima é a de *Lebesgue-Stieltjes*. Uma definição desta integral encontra-se em Durret (1991), página 416, ou em qualquer literatura sobre Teoria da Medida. No caso discreto, esta integral assume a forma de uma somatória, a saber,

$$E(Y_i) = \sum_i y_i P(Y_i = y_i).$$

Usaremos para exemplificar o caso de uma variável contínua em que temos:

$$E(Y_i) = \int_{-\infty}^{\infty} y_i f(y_i; \theta_i, \phi) dy_i.$$

Como f é função densidade, então

$$\int_{-\infty}^{\infty} f(y_i; \theta_i, \phi) = 1. \tag{2.1}$$

Supondo que todas as condições de regularidade são satisfeitas, de (2.1) temos

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

$$\int_{-\infty}^{\infty} e^{\phi\theta_i y_i} e^{c(y_i,\phi)} dy_i = e^{\phi b(\theta_i)}$$

$$b(\theta_i) = \frac{1}{\phi} \log \int_{-\infty}^{\infty} e^{\phi\theta_i y_i} e^{c(y_i,\phi)} dy_i,$$

então

$$\begin{aligned} b'(\theta_i) &= \frac{1}{\phi}\frac{1}{\int_{-\infty}^{\infty} e^{\phi\theta_i y_i} e^{c(y_i,\phi)} dy} \int_{-\infty}^{\infty} y_i \phi e^{\phi\theta_i y_i} e^{c(y_i,\phi)} dy_i \\ &= e^{-\phi b(\theta_i)} \int_{-\infty}^{\infty} y_i e^{\phi\theta_i y_i} e^{c(y_i,\phi)} dy_i \\ &= \int_{-\infty}^{\infty} y_i e^{\phi(\theta_i y_i - b(\theta_i))} e^{c(y_i,\phi)} dy_i \\ &= E(Y) = \mu. \end{aligned}$$

A seguir, mostramos que a variância é dada por $Var(Y_i) = \phi^{-1} V_i$, sendo que $V_i = d\mu_i/d\theta_i$ é a chamada função de variância, que caracteriza a distribuição e ϕ^{-1} é o parâmetro de dispersão. Note que $\phi^{-1} > 0$.

Como $Var(Y_i) = E(Y_i^2) - \mu_i^2 = E(Y_i^2) - [b'(\theta_i)]^2$, e também $b'(\theta_i) = \mu_i$, então

$$\begin{aligned} b''(\theta_i) &= \frac{d\mu_i}{d\theta_i} \\ &= (-\phi b'(\theta_i)) \int_{-\infty}^{\infty} y_i f(y_i;\theta_i,\phi) dy_i + \phi \int_{-\infty}^{\infty} y_i^2 f(y_i;\theta_i,\phi) dy_i \\ &= -\phi [b'(\theta_i)]^2 + \phi E(Y_i^2). \end{aligned}$$

Rearranjando os termos da equação acima, chegamos à expressão da variância:

$$\phi^{-1}\frac{d\mu_i}{d\theta_i} = E(Y_i^2) - [b'(\theta_i)]^2 = Var(Y_i).$$

Em relação à distribuição Binomial para uma variável aleatória temos:

$$Y \sim Binomial(n, \mu),$$

$$f_Y(y) = \left(\begin{array}{c} n \\ y \end{array} \right) \mu^y (1-\mu)^y$$

$$= \exp\left\{ \left[y \log\left(\frac{\mu}{1-\mu} \right) + n \log(1-\mu) \right] + \log \left(\begin{array}{c} n \\ y \end{array} \right) \right\}.$$

Temos então que:

$\phi = 1$ (parâmetro de dispersão),

$\theta = \log\left(\frac{\mu}{1-\mu} \right)$ (função de ligação canônica),

$b(\theta) = n\log(1+e^{\theta}) = -n\log(1-\mu)$,

$b'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}} = n\mu$,

$c(y,\phi) = \log\left(\begin{array}{c} n \\ y \end{array}\right)$,

$Var(y) = \phi^{-1}\frac{d\mu}{d\theta} = n\mu(1-\mu)$,

$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \mathbf{x}_i'\boldsymbol{\beta}$,

$\mu = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$.

A função $b''(\theta)$ deve depender somente de μ. Para o caso da distribuição Binomial, temos então que $Y^* = Y/n$ pertence à família exponencial.

Vejamos como ficam as funções referentes à distribuição de Poisson para uma variável:

$$Y \sim Poisson(\mu),$$

$$f_Y(y) = e^{-\mu}\frac{\mu^y}{y!} = \exp\{[y\log\mu - \mu] - \log(y!)\}.$$

Temos então que:

$\phi = 1$ (parâmetro de dispersão),

$\theta = \log\mu$ (função de ligação canônica),

$b(\theta) = e^{\theta} = e^{\log\mu} = \mu$,

$b'(\theta) = e^{\theta} = \mu$,

$c(y,\phi) = -\log(y!)$,

$Var(y) = \phi^{-1}\frac{d\mu}{d\theta} = \mu$,

$\eta = g(\mu) = \log\mu = \mathbf{x}_i'\boldsymbol{\beta}$,

$\mu = e^{\mathbf{x}_i'\boldsymbol{\beta}}$.

2.2 Estimação dos Parâmetros

Se $Y_1, ..., Y_n$ é uma amostra de variáveis aleatórias independentes com médias diferentes, logo, não são identicamente distribuídas. A função de verossimilhança é dada por

$$\begin{aligned} L(\mathbf{y};\theta,\phi) &= \prod_{i=1}^{n} \exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\} \\ &= \exp\{\sum_{i=1}^{n} \phi[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i,\phi)\}. \end{aligned}$$

Temos então que o logaritmo da função de verossimilhança é dado por

$$\ell(\mathbf{y};\theta,\phi) = \log\{L(\mathbf{y};\theta,\phi)\} = \sum_{i=1}^{n} \phi[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i,\phi). \tag{2.2}$$

Usando os resultados gerais de verossimilhança que valem sob as condições de regularidade satisfeitas pela família exponencial,

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \quad \text{e} \quad -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = E\left(\frac{\partial \ell}{\partial \theta}\right)^2, \tag{2.3}$$

Agresti (2006) mostra que para uma observação y_i a sua contribuição para o logaritmo da verossimilhança é dada por

$$\ell_i = \phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi) \quad \text{e}$$

$$\frac{\partial \ell_i}{\partial \theta_i} = \phi[y_i - b'(\theta_i)] \ , \ \frac{\partial^2 \ell_i}{\partial \theta_i^2} = -\phi b''(\theta_i).$$

Aplicando estes resultados em (2.3) para uma única observação,

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = E(\phi[Y_i - b'(\theta_i)]) = 0 \quad \text{ou} \quad E(Y_i) = b'(\theta_i) \quad \text{e}$$

$$E\left(\frac{\partial \ell}{\partial \theta}\right)^2 = E(\phi[Y_i - b'(\theta_i)])^2 = \phi^2 Var(Y_i) = \phi b''(\theta_i).$$

Desta forma $Var(Y_i) = b''(\theta_i)/\phi$, ou seja, a função $b(\cdot)$ determina os momentos de Y_i.

A estimação dos parâmetros do vetor $\boldsymbol{\beta}$ pelo método da máxima verossimilhança consiste em encontrar o vetor $\hat{\boldsymbol{\beta}}$ que maximize a função de verossimilhança, ou o logaritmo desta.

Se temos n observações independentes,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i = \sum_{i=1}^{n} \phi[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi).$$

Maximizando a verossimilhança,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0.$$

Aplicando a regra da cadeia para uma observação:

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Sabemos que

$$\frac{\partial \ell_i}{\partial \theta_i} = \phi[y_i - b'(\theta_i)] \ , \ \mu_i = b'(\theta_i) \ \Rightarrow \ \frac{\partial \ell_i}{\partial \theta_i} = \phi[y_i - \mu_i];$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \phi Var(Y_i) \ \Rightarrow \ \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\phi Var(Y_i)};$$

$$\eta_i = g(\mu_i) \ \Rightarrow \ \frac{\partial \mu_i}{\partial \eta_i} \text{ depende de } g(\mu_i);$$

$$\eta_i = \sum_j \beta_j x_{ij} \ \Rightarrow \ \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Portanto, para uma observação i, temos que

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \phi(y_i - \mu_i)\frac{1}{\phi Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} = \frac{(y_i - \mu_i)}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}. \tag{2.4}$$

As equações de verossimilhança ficam

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} = 0 \quad , \quad j = 1, ..., m.$$

A matriz de covariância assintótica do estimador de verossimilhança $\hat{\boldsymbol{\beta}}$ é dada pela inversa da matriz de informação de Fisher, $\mathcal{J}$. Vale para a família exponencial

$$\begin{aligned} E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) &= -E\left(\frac{\partial \ell_i}{\partial \beta_j}\right)\left(\frac{\partial \ell_i}{\partial \beta_k}\right) \\ &= -E\left(\frac{(Y_i - \mu_i)x_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\frac{(Y_i - \mu_i)x_{ik}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\right) \\ &= -\frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \end{aligned}$$

Mas temos que,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i \text{ e } w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{Var(Y_i)},$$

então

$$E\left(-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \sum_{i=1}^{n} x_{ij} w_i x_{ik}.$$

Escrevendo na forma matricial, expressamos a matriz de Informação de Fisher

$$\mathcal{J} = -E\left(\frac{\partial \ell^2(\boldsymbol{\beta}; \mathbf{y})}{\partial \beta_j \partial \beta_k}\right) = \mathbf{X}'\mathbf{W}\mathbf{X}, \tag{2.5}$$

tal que $\mathbf{W} = diagonal\{w_1, ..., w_n\}$.

Assim a matriz de variância-covariância de $\hat{\boldsymbol{\beta}}$ é dada por

$$Cov(\hat{\boldsymbol{\beta}}) = \mathcal{J}^{-1} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}.$$

A solução das equações de verossimilhança para $\boldsymbol{\beta}$ é feita por um método iterativo. Agresti (2006) inicialmente nos mostra como ficaria a estimação usando o método de Newton-Raphson. Este é um método clássico para o cálculo de zeros de funções. Segundo Cláudio e Marins (1988), se queremos encontrar o valor x em que a função $f(x)$ se anula, a fórmula de Newton ou o método das tangentes é dado por:

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}. \tag{2.6}$$

em que t indica a t-ésima iteração.

Na estimação de verossimilhança estamos procurando os valores de $\boldsymbol{\beta}$ que maximizem $\ell(\boldsymbol{\beta})$, ou seja, queremos encontrar os valores de $\boldsymbol{\beta}$ que anulem $\ell'(\boldsymbol{\beta})$.

Aplicando em (2.6) temos

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \frac{\ell'(\boldsymbol{\beta})}{\ell''(\boldsymbol{\beta})}.$$

A primeira derivada do logaritmo da verossimilhança é a chamada função escore

$$\mathbf{u}' = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, ..., \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_m}\right),$$

e a segunda derivada do logaritmo da verossimilhança é a chamada matriz Hessiana $\boldsymbol{\mathcal{H}}$ com entradas

$$\mathcal{H}_{ab} = \frac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b}, \quad \text{ou}$$

matriz de informação de Fisher observada.

O passo t no processo iterativo $t = 0, 1, 2, ...$ aproxima $\ell(\boldsymbol{\beta})$ perto de $\boldsymbol{\beta}^{(t)}$, e o passo $(t+1)$ no processo iterativo é dado por

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [\boldsymbol{\mathcal{H}}^{(t)}]^{-1}\mathbf{u}^{(t)}.$$

As iterações procedem até que as alterações em $\ell(\boldsymbol{\beta}^{(t)})$ sejam muito pequenas.

Vamos dar um exemplo usando um problema conhecido. O valor que maximiza a verossimilhança de uma observação y de uma distribuição Binomial (n, π) é y/n. Vejamos este mesmo cálculo usando o método de Newton-Raphson.

Sabemos que

$$\ell(\pi) = y \log \pi + (n-y)\log(1-\pi) + \log \left(\begin{array}{c} n \\ y \end{array} \right),$$

$$\ell'(\pi) = \frac{y}{\pi} - \frac{(n-y)}{(1-\pi)} = \frac{(y - n\pi)}{\pi(1-\pi)},$$

$$\ell''(\pi) = -\left[\frac{y}{\pi^2} + \frac{(n-y)}{(1-\pi)^2}\right],$$

e cada passo da iteração tem a forma

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{y}{(\pi^{(t)})^2} + \frac{(n-y)}{(1-\pi^{(t)})^2}\right]^{-1} \frac{(y - n\pi^{(t)})}{\pi^{(t)}(1-\pi^{(t)})}.$$

Se iniciarmos com $\pi^{(0)} = 1/2$ verificamos que $\pi^{(1)} = y/n$ e os próximos passos continuam iguais à y/n, ou seja, já chegamos na resposta correta para $\hat{\pi}$.

Um outro método para encontrar os estimadores de máxima verossimilhança é o do escore de Fisher. Ao invés de utilizarmos a informação de Fisher observada, usamos a informação de Fisher esperada, $\mathcal{J}$. A expressão do cálculo iterativo fica:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + [\mathcal{J}^{(t)}]^{-1}\mathbf{u}^{(t)} \quad \text{ou}$$

$$\mathcal{J}^{(t)}\boldsymbol{\beta}^{(t+1)} = \mathcal{J}^{(t)}\boldsymbol{\beta}^{(t)} + \mathbf{u}^{(t)}. \tag{2.7}$$

Existe uma relação entre este método e o de estimação por mínimos quadrados ponderados. Lembrando que

$$u_j = \frac{\partial \ell}{\partial \beta_j} = \frac{y_i - \mu_i}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \quad \text{e} \quad w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{Var(Y_i)},$$

e multiplicando o numerador e o denominador da primeira equação acima por $\frac{\partial \mu_i}{\partial \eta_i}$, temos que

$$\begin{aligned} u_j &= \frac{y_i - \mu_i}{Var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \frac{\partial \eta_i}{\partial \mu_i} \\ &= x_{ij} w_i (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}. \end{aligned}$$

Assim, o lado direito de (2.7) pode ser escrito como

$$\begin{aligned} \mathcal{J}^{(t)} \boldsymbol{\beta}^{(t)} + \mathbf{u}^{(t)} &= \mathbf{X}' \mathbf{W}^{(t)} \mathbf{X} \boldsymbol{\beta}^{(t)} + \mathbf{u}^{(t)} \\ &= \mathbf{X}' \mathbf{W}^{(t)} \sum_{i=1}^{n} x_{ij} \beta_j^{(t)} + \sum_{i=1}^{n} x_{ij} w_i (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} \\ &= \mathbf{X}' \mathbf{W}^{(t)} \mathbf{z}^{(t)}, \end{aligned}$$

onde $\mathbf{z}^{(t)}$ é dada por

$$\begin{aligned} \mathbf{z}^{(t)} &= \sum_{i=1}^{n} x_{ij} \beta_j^{(t)} + (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} \\ &= \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}. \end{aligned}$$

Para todo i, z_i é uma forma linearizada da função de ligação g, avaliada em y_i ou seja,

$$g(y_i) \cong g(\mu_i) + (y_i - \mu_i) g'(\mu_i) = \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} = z_i.$$

Desta forma, a equação (2.7) para o escore de Fisher assume a forma

$$\mathbf{X}' \mathbf{W}^{(t)} \mathbf{X} \boldsymbol{\beta}^{(t+1)} = \mathbf{X}' \mathbf{W}^{(t)} \mathbf{z}^{(t)},$$

que são as equações usadas para ajustar um modelo linear por mínimos quadrados ponderados para a variável resposta $\mathbf{z}^{(t)}$ quando a matriz do modelo é $\mathbf{X}$ e $\mathbf{W}^{(t)}$ é a inversa da matriz de covariância. A solução das equações é dada por:

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}' \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{(t)} \mathbf{z}^{(t)}.$$

A variável resposta ajustada $\mathbf{z}$ tem elementos i aproximados por $z_i^{(t)}$ para o passo t do processo iterativo. Em cada passo faz-se uma regressão de $\mathbf{z}^{(t)}$ em $\mathbf{X}$ com pesos $\mathbf{W}^{(t)}$ para se obter uma nova estimativa $\boldsymbol{\beta}^{(t+1)}$. Esta estimativa fornece um novo preditor linear $\boldsymbol{\eta}^{(t+1)} = \mathbf{X} \boldsymbol{\beta}^{(t+1)}$ e uma nova resposta ajustada, $\mathbf{z}^{(t+1)}$, para o próximo passo.

Assim, o estimador de máxima verossimilhança resulta do uso iterativo de mínimos quadrados ponderados, nos quais a matriz de pesos muda a cada passo. Este é o chamado método iterativo de mínimos quadrados re-ponderados.

Uma maneira simples de começar o processo iterativo é usar os dados $\mathbf{y}$ como estimativa para $\boldsymbol{\mu}$. Isto determina a primeira estimativa para a matriz de pesos $\mathbf{W}$ e portanto a estimativa inicial de $\boldsymbol{\beta}$.

Se a função de ligação utilizada for a função canônica, aparecem algumas simplificações. Temos

$$\eta_i = \theta_i = \sum_{j=1} \beta_j x_{ij}.$$

Em (2.2) podemos encontrar casos em que o valor de $\phi = 1$, como por exemplo para a regressão de Poisson. A parte da verossimilhança que envolve os dados fica então $\sum y_i\theta_i$, que simplificamos para

$$\sum_i y_i\theta_i = \sum_i y_i\left(\sum_j \beta_j x_{ij}\right) = \sum_j \beta_j\left(\sum_i y_i x_{ij}\right).$$

Desta forma, a estatística suficiente para estimar $\boldsymbol{\beta}$ é $\sum_i y_i x_{ij}$ para $j = 1, ..., m$. Para a função de ligação canônica,

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i).$$

Assim, (2.4) pode ser escrita como

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{Var(Y_i)} b''(\theta_i) x_{ij} = (y_i - \mu_i) x_{ij}\phi.$$

Quando ϕ é idêntico para todas as observações, as equações de verossimilhança são

$$\sum_i x_{ij} y_i = \sum_i x_{ij}\mu_i \quad , \quad j = 1, ..., m.$$

Estas equações igualam a estatística suficiente para os parâmetros do modelo aos seus valores esperados. Com a função de ligação canônica, as segundas derivadas do logaritmo da verossimilhança ficam

$$\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} = -\phi x_{ij} \frac{\partial \mu_i}{\partial \beta_k},$$

que não dependem mais das observações y_i, portanto,

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = E\left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right).$$

Desta forma, $\boldsymbol{\mathcal{H}} = -\boldsymbol{\mathcal{J}}$. Para modelos com ligações canônicas, os algoritmos de Newton-Raphson e do escore de Fisher são idênticos.

2.3 Ajuste do modelo

2.3.1 Escolha do melhor modelo

Existem várias abordagens relacionadas à busca do melhor modelo, como por exemplo sugere Myers et al. (2001). Começamos com a escolha das variáveis regressoras consideradas na matriz $\mathbf{X}$. Por exemplo, inicia-se com o modelo saturado, que envolve os efeitos principais de todas variáveis regressoras e todas as interações entre estas. Em seguida, verificam-se quais coeficientes são estatisticamente significantes. Consideraremos os testes de hipótese para os elementos individuais do vetor $\hat{\boldsymbol{\beta}}$, levando em conta a normalidade assintótica dos estimadores de máxima verossimilhança:

$$H_0 : \beta_j = 0, \;\; j = 1, \ldots, p;$$

$$\left(\frac{\hat{\beta}_j}{dp(\hat{\beta}_j)}\right)^2 \sim \chi^2_1.$$

O desvio padrão de cada parâmetro β_j escrito como $dp(\hat{\beta}_j)$ é dado pela raiz quadrada do j-ésimo elemento da diagonal principal do inverso da matriz de Informação de Fisher.

Com relação à estimação de máxima verossimilhança, para estimar a variância e também o ajuste do modelo, uma das medidas empregadas é a função desvio, ou *deviance*. Casela e Berger (2001) mostram que, sob as condições de regularidade satisfeitas pela família exponencial, a estatística $-2\log\lambda(\mathbf{y})$ converge para uma distribuição χ^2_{n-p}, em que

$$\lambda(\mathbf{y}) = \frac{L(\hat{\mu}; \theta, \phi)}{L(\mathbf{y}; \theta, \phi)},$$

n é o número total de indivíduos na amostra e p o número total de parâmetros estimados. Porém, este resultado não vale quando o número de parâmetros aumenta junto com o n.

Temos que $L(y; \theta, \phi)$ é a verossimilhança do modelo saturado, enquanto que $L(\hat{\mu}; \theta, \phi)$ é a verossimilhança do modelo ajustado. A *deviance* é dada por

$$\begin{aligned} D(\mathbf{y}, \hat{\mu}) &= -2\log\lambda(\mathbf{y}) = -2[\ell(\hat{\mu}; \theta, \phi) - \ell(y; \theta, \phi)] = 2[\ell(y; \theta, \phi) - \ell(\hat{\mu}; \theta, \phi)] \\ &= 2\phi\left\{\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] - \sum_{i=1}^{n}[\hat{\mu}_i\theta_i - b(\theta_i)]\right\}. \end{aligned} \tag{2.8}$$

Por exemplo, para a distribuição normal de uma variável com variância conhecida σ^2,

$$f(y, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

O logaritmo da verossimilhança é dado por

$$\ell(\mu, y) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}.$$

Se $\mu = y$, o valor máximo alcançado pelo logaritmo da verossimilhança é

$$\ell(y, y) = -\frac{1}{2}\log(2\pi\sigma^2),$$

e o desvio

$$D(y, \mu) = 2[\ell(y, y) - \ell(\mu, y)] = \frac{(y-\mu)^2}{\sigma^2}.$$

Para os modelos considerados, $\ell(y, y)$ é o valor máximo do logaritmo da verossimilhança quando os valores ajustados são iguais aos valores observados. Portanto, o melhor modelo ajustado é aquele com o máximo $\ell(\mu, y)$, ou seja, menor valor de $D(y, \mu)$. Após a seleção do melhor conjunto de variáveis explicativas, o próximo passo é comparar a *deviance* dos possíveis modelos, verificando em qual deles ela é mínima.

Outra medida usada para comparar o ajuste de modelos é o o critério AIC:

$$AIC = -2\ell(\mu, y) + 2m,$$

em que m é o número de parâmetros estimados. Também procuramos o modelo que tenha o menor valor para o critério AIC.

Como medida de discrepância temos também o Chi-quadrado de Pearson generalizado,

$$\boldsymbol{\chi}^2 = \sum \frac{(y-\hat{\mu})^2}{V(\hat{\mu})},$$

onde $V(\hat{\mu})$ é a função de variância estimada para a distribuição com a qual estamos lidando. Para a distribuição Normal esta estatística é idêntica à soma dos quadrados dos resíduos, enquanto que para as distribuições Poisson e Binomial trata-se da estatística de Pearson original.

Em relação à estimação do parâmetro de dispersão, foram sugeridas várias abordagens, que mudam segundo a distribuição de probabilidade considerada. No caso das distribuições da família exponencial há interesse em estimar ϕ para as distribuições Normal, Normal Inversa e Gama. Para a Poisson e a Binomial, o parâmetro de dispersão é um. Por exemplo, a estimativa $\hat{\phi}^{-1} = D(y, \hat{\mu})/(n-p)$, ou seja, a *deviance* dividida pelos graus de liberdade, é usada como indicadora da presença de superdispersão no caso da Binomial e da Poisson. Já para o caso da Gama, esta não é uma estimativa consistente segundo Paula (2004).

Para os modelos de Poisson, antes de prosseguirmos no ajuste precisamos verificar se a *deviance* dividida pelos graus de liberdade é próxima da unidade. Se não for, podemos ter escolhido uma distribuição inadequada para os dados, ou estar diante do problema da superdispersão, que será melhor analisado no próximo capítulo.

2.3.2 Técnicas de diagnóstico

As chamadas técnicas de diagnóstico referem-se basicamente à análise dos resíduos e de pontos na amostra que possam ter uma influência muito grande no ajuste do modelo.

Podemos pensar basicamente em 4 resíduos:

1. $Y_i - \hat{\mu}_i$, ou resíduos de resposta;

2. $\hat{r}_{P_i} = (Y_i - \hat{\mu}_i)/\sqrt{\hat{V}_i}$, resíduos de Pearson;

3. $(Y_i - \hat{\mu}_i)/(d\mu_i/d\eta_i)$, resíduos de *working*, que vêm do último estágio do processo iterativo;

4. $d(Y_i, \hat{\mu}_i)$, resíduos *deviance*, que são a raiz quadrada de cada componente da função desvio $\sqrt{D(y_i, \hat{\mu}_i)}$, (definida em (2.8)), e o sinal dos resíduos são os mesmos de $(Y_i - \hat{\mu}_i)$.

Em relação aos resíduos dos MLGs, Paula (2004) mostra que os que estão mais próximos da normalidade são os resíduos padronizados da *deviance*.

Paula (2004) também mostra como, das técnicas de diagnóstico existentes para o modelo linear clássico, pode-se obter técnicas para os MLG's. Inicialmente Pregibon (1981) estendeu a técnica de diagnóstico para o modelo logístico e mostrou que podemos usar a diagonal principal da matriz de projeção $\mathbf{H}$ para detectar pontos de alavanca. Para os modelos generalizados,

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}.$$

Estas técnicas podem ser extendidas para os outros modelos da classe dos MLGs.

Uma medida que pode ser usada para os modelos não lineares é chamada de afastamento da verossimilhança (*likelihood displacement*):

$$LD_i = 2\{L(\hat{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}_{(i)})\}.$$

Ou seja, para cada uma das n observações da amostra verifica-se qual é o afastamento da verossimilhança total $L(\hat{\boldsymbol{\beta}})$ em relação ao valor da verossimilhança que seria obtida se a observação i desconsiderada, $L(\hat{\boldsymbol{\beta}}_{(i)})$. Uma maneira de obter estes valores é através de uma aproximação por série de Taylor até segunda ordem.

Por exemplo, para uma função G de duas componentes (x e y), queremos obter o valor aproximado em série de Taylor desprezando os valores de terceira ordem em diante.

Temos então

$$\begin{aligned} G(x,y) &\cong G(x_0,y_0) + < (G_x, G_y), (x-x_0, y-y_0) > \\ &+ \frac{1}{2}(x-x_0 \quad y-y_0) \begin{pmatrix} G_{xx} & G_{xy} \\ G_{xy} & G_{yy} \end{pmatrix} \begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix}. \end{aligned}$$

Vamos então expandir a função $L(\boldsymbol{\beta})$ em torno de $\hat{\boldsymbol{\beta}}$:

$$L(\boldsymbol{\beta}) \cong L(\hat{\boldsymbol{\beta}}) + L'(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' L''(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Passamos $L(\hat{\boldsymbol{\beta}})$ para o lado esquerdo da equação e temos que se L é a função de verossimilhança, ao maximizá-la, $L'(\hat{\boldsymbol{\beta}}) = 0$ e portanto,

$$2\{L(\boldsymbol{\beta}) - L(\hat{\boldsymbol{\beta}})\} \cong (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' L''(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Sabemos por (2.5) que o valor esperado de $-L''(\hat{\boldsymbol{\beta}})$ corresponde à matriz de informação de Fisher, podemos substituir $-L''(\hat{\boldsymbol{\beta}})$ na equação anterior pelo seu valor esperado e trocar $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}_{(i)}$ obtendo assim um valor aproximado para LD_i

$$LD_i \cong 2\{L(\hat{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}_{(i)})\} \cong (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \{\phi \mathbf{X}' \hat{\mathbf{W}} \mathbf{X}\} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}). \tag{2.9}$$

Para obter os valores de $\hat{\boldsymbol{\beta}}_{(i)}$, Pregibon (1981) utilizou a aproximação de um passo do processo iterativo de *scoring* de Fisher quando o mesmo é iniciado em $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^1_{(i)} = \boldsymbol{\beta} - \frac{\hat{r}_{P_i}\sqrt{\hat{w}_i \phi^{-1}}}{(1-\hat{h}_{ii})} (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_i.$$

Substituindo a expressão anterior em (2.9) temos

$$\begin{aligned} LD_i &\cong \left(\frac{\hat{r}_{P_i}\sqrt{\hat{w}_i\phi^{-1}}}{(1-\hat{h}_{ii})}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_i\right)' \{\phi \mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\} \left(\frac{\hat{r}_{P_i}\sqrt{\hat{w}_i\phi^{-1}}}{(1-\hat{h}_{ii})}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_i\right) \\ &= \frac{\hat{r}^2_{P_i}\hat{w}_i}{(1-\hat{h}_{ii})^2}((\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_i)'(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})((\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_i) \\ &= \frac{\hat{r}^2_{P_i}\hat{w}_i}{(1-\hat{h}_{ii})^2}\mathbf{x}_i'(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_i \\ &= \left(\frac{\hat{r}_{P_i}}{\sqrt{(1-\hat{h}_{ii})}}\right)^2 \frac{1}{(1-\hat{h}_{ii})}\sqrt{\hat{w}_i}\mathbf{x}_i'(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_i\sqrt{\hat{w}_i} \\ &= t^2_{S_i}\frac{1}{(1-\hat{h}_{ii})}\hat{h}_{ii} \ . \end{aligned}$$

Acima, $t^2_{S_i}$ é o quadrado do resíduo de Pearson e $\hat{h}_{ii}$ é o i-ésimo componente da diagonal principal da matriz $\hat{\mathbf{H}}$. Esta é uma medida aproximada da distância de verossimilhança. Paula (2004) discute que esta aproximação tem sido investigada por pesquisadores, e que apesar dela subestimar o verdadeiro valor de LD_i, ainda assim ela é suficiente para identificar pontos aberrantes.

Podemos construir um gráfico com os valores de LD_i para cada observação para identificar quais observações têm uma influência grande no valor da verossimilhança. Venables e Ripley (1999) sugerem como gráfico de diagnóstico os valores dos resíduos de *deviance* contra os valores lineares preditos.

Uma outra ferramenta de análise é o gráfico de envelope sugerido por Paula (2004). O objetivo do gráfico de envelope é gerar a distribuição empírica do resíduo através de simulações sob o modelo postulado e comparar com os resíduos padronizados da *deviance* observados:

$$t_{D_i} = \frac{\phi^{1/2} d(y_i, \hat{\mu}_i)}{\sqrt{(1 - \hat{h}_{ii})}}.$$

A banda de confiança de 95% do gráfico do envelope é calculada ordenando-se os resíduos da simulação, que pode ser, por exemplo, de cem modelos, usando a distribuição em questão.

2.4 Intervalos de Confiança

Para calcularmos os intervalos de confiança para os parâmetros ou funções destes, recorreremos ao Método Delta. Seja T_n uma estatística, em que refere-se ao tamanho da amostra. Para grandes amostras, T_n é aproximandamente normal, distribuída em torno de θ, com erro padrão aproximado $\sigma/\sqrt{n}$. À medida em que n diverge para infinito, a função de distribuição acumulada de $\sqrt{n}(T_n - \theta)$ converge para a função distribuição acumulada de uma variável aleatória normal com média 0 e variância σ^2. Este é um exemplo de limite em distribuição, onde

$$\sqrt{n}[T_n - \theta] \xrightarrow{\mathcal{D}} N(0; \sigma^2).$$

A estatística $g(T_n)$ também tem uma distribuição limite. Para mostrarmos isto, vamos supor que g é pelo menos duas vezes diferenciável em relação a θ. Usaremos a expansão de Taylor para g, avaliada numa vizinhança de θ. Para algum θ^* entre t e θ,

$$\begin{aligned} g(t) &= g(\theta) + (t - \theta) g'(\theta) + (t - \theta)^2 g''(\theta^*)/2 \\ &= g(\theta) + (t - \theta) g'(\theta) + O_p(|t - \theta|^2). \end{aligned}$$

Se $t = T_n$,

$$\begin{aligned} g(T_n) &= g(\theta) + (T_n - \theta)g'(\theta) + (T_n - \theta)^2 g''(\theta^*)/2 \\ &= g(\theta) + (T_n - \theta)g'(\theta) + O_p(|T_n - \theta|^2). \end{aligned}$$

Multiplicando os dois lados por $\sqrt{n}$ e rearranjando os termos,

$$\sqrt{n}(g(T_n) - g(\theta)) = \sqrt{n}(T_n - \theta)g'(\theta) + \sqrt{n}O_p(|T_n - \theta|^2).$$

O último termo à direita será desprezível, uma vez que temos que $|T_n - \theta|$ é limitada com probabilidade alta ou seja,

$$\lim_{N\to\infty} \lim_{n\to\infty} P(|T_n - \theta| \leq N) = 1.$$

Assim temos que $|T_n - \theta| \xrightarrow{P} 0$.

Levando-se em conta que no limite o último termo à direita da expressão acima tende a zero, concluímos que $\sqrt{n}(g(T_n) - g(\theta))$ tem a mesma distribuição limite que $\sqrt{n}(T_n - \theta)g'(\theta)$, ou mais precisamente,

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{\mathcal{D}} N(0; \sigma^2(g'(\theta))^2).$$

Este é o chamado método delta para obter a distribuição assintótica de funções de estatísticas, com g duas vezes diferenciável. Como em geral $\sigma^2 = \sigma^2(\theta)$ e $g'(\theta)$ dependem do parâmetro desconhecido θ, a variância assintótica também não é conhecida. Mas quando $\sigma^2 = \sigma^2(\cdot)$ e $g'(\cdot)$ são funções contínuas em θ e $\sigma(T_n)$, $g'(T_n)$ são estimadores consistentes de $\sigma(\theta)$ e $g'(\theta)$. Dessa forma, podemos construir os intervalos de confiança e realizar testes usando o resultado que $\sqrt{n}(g(T_n) - g(\theta))/\sigma(T_n)|g'(T_n)|$ tem distribuição assintótica normal padrão. Ou seja,

$$IC_{100(1-\alpha)\%} : g(T_n) \pm z_\alpha \frac{\sigma(T_n)}{\sqrt{n}}|g'(T_n)|,$$

onde z_α é o valor na tabela da Normal padrão tal que $P(Z > z_\alpha = \alpha$.

Como estamos diante de estimadores de máxima verossimilhança, estas propriedades se verificam e podemos utilizar o método delta. Podemos também generalizar este método para funções de vetores aleatórios. Suponha que $\mathbf{T}_n = (T_{n_1}, ..., T_{n_N})'$ tenha distribuição assintótica normal multivariada com média $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)'$ e matriz de covariância $\boldsymbol{\Sigma}/n$. Suponha que as derivadas das funções $g(t_1, ..., t_N)$ sejam $\boldsymbol{\phi} = (\phi_1, ..., \phi_N)'$ e não se anulem em $\boldsymbol{\theta}$. Então

$$\phi_i = \left.\frac{\partial g}{\partial t_i}\right|_{t=\theta},$$

$$\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{\mathcal{D}} N(0; \boldsymbol{\phi}^{'}\boldsymbol{\Sigma}\boldsymbol{\phi}) \text{ e}$$

$$g(T_n) \xrightarrow{\mathcal{D}} N\left(g(\theta); \frac{\boldsymbol{\phi}^{'}\boldsymbol{\Sigma}\boldsymbol{\phi}}{n}\right).$$

Por exemplo, suponha uma regressão de Poisson com duas variáveis explicativas categóricas X_1 e X_2 que podem assumir os valores 0 ou 1. Nosso modelo então fica:

$$\ln \lambda = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Se quisermos encontrar um intervalo de confiança para λ quando $X_1 = 1$ e $X_2 = 1$, teremos

$$\hat{\lambda} = e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2},$$

de tal forma que $\mathbf{T}_n = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$, $g(\mathbf{T}_n) = e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2}$,

$$\boldsymbol{\phi}_{\hat{\boldsymbol{\beta}}} = \begin{pmatrix} \frac{\partial g}{\partial \beta_0} \\ \frac{\partial g}{\partial \beta_1} \\ \frac{\partial g}{\partial \beta_2} \end{pmatrix} = \begin{pmatrix} e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2} \\ e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2} \\ e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2} \end{pmatrix} \text{ e}$$

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & Cov(\hat{\beta}_0, \hat{\beta}_2) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_0, \hat{\beta}_2) & Cov(\hat{\beta}_1, \hat{\beta}_2) & Var(\hat{\beta}_2) \end{pmatrix}.$$

Assim, intervalo de confiança para λ é dado por

$$IC_{100(1-\alpha)\%} : e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2} \pm z_\alpha \sqrt{\boldsymbol{\phi}'_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \boldsymbol{\phi}_{\hat{\boldsymbol{\beta}}}} \ .$$

Capítulo 3

Modelando Dados de Contagem

3.1 Introdução

Ao lidarmos com dados de contagem, a distribuição de probabilidade muitas vezes é a de Poisson. Se além disso estivermos lidando com dados categorizados, sempre podemos construir tabelas de contigência separando as contagens pelos diferentes fatores possíveis.

Neste caso é importante conhecer como a amostra foi obtida, pois isto nos dará informação sobre o modelo probabilístico que se encontra por trás da tabela de contingência. Na próxima seção discutiremos as estratégias de amostragem possíveis e os modelos probabilísticos associados. Verificaremos que no caso de dados categorizados encontramos uma relação entre a distribuição Poisson e multinomial.

Nas seções seguintes apresentaremos modelos com regressão de Poisson, bem como os problemas que comumente aparecem.

3.2 Modelos Probabilísticos

Em Paulino e Singer (2006) há uma discussão sobre os diferentes modelos probabilísticos para analisar dados categorizados. Basicamente, as suposições dependem do planejamento do experimento. Há 3 estratégias, que serão ilustradas através de um exemplo prático. Queremos saber para duas sub-populações, ou estratos de uma população, qual é a opinião das pessoas a respeito de um assunto. As diferentes sub-populações podem estar divididas por faixa etária, nível de renda, de escolaridade, etc.

Estratégia 1. O entrevistador faz a pergunta às pessoas que passarem num determinado período de tempo pré fixado (ele não sabe quantas n pessoas ele vai conseguir entrevistar). Terminado o período de entrevistas, as respostas obtidas serão classificadas de acordo com a sub-população e a resposta dada (cada n_{ij}), como ilustra a Tabela 3.1.

Tabela 3.1: Estratégia 1 - Período fixo, número total de entrevistas indeterminado.

Opinião	favor	contra	Total
sub-população 1	n_{11}	n_{12}	$n_{1.}$
sub-população 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Se X_{11} é o número de entrevistados da sub-população 1 que são a favor, podemos fazer algumas suposições sobre a variável:

1. Incrementos estacionários: a distribuição do número de pessoas no intervalo de tempo $(s, s+t]$ só depende da extensão t e não de s. Esta hipótese é simplificadora, pois despreza períodos de tempo em que existe maior movimento de transeuntes, mas continua sendo uma boa aproximação em períodos curtos de tempo;

2. Incrementos independentes: o número de pessoas que passa num determinado período de tempo $(s, s+t]$ é independente do número de pessoas que passam em outro intervalo disjunto de tempo $(u, u+v]$, ou seja, $(s, s+t] \cap (u, u+v] = \emptyset$;

3. Chegadas não coincidentes: a probabilidade de duas pessoas passarem simultaneamente num intervalo de tempo muito pequeno é desprezível.

Estas suposições, utilizadas por James (1996), permitem deduzir o modelo probabilístico para o processo de Poisson. Desta forma, o número de pessoas n_{11} da sub-população 1, que são favoráveis à proposta e que passam no intervalo de tempo de comprimento t, é uma variável aleatória, X_{11}, com distribuição de Poisson com média $\mu_{11} = t\lambda_1$. Estas mesmas suposições podem ser aplicadas aos n_{ij} restantes. Admitindo a independência das variáveis aleatórias X_{ij}, chegamos ao modelo de produto de distribuições de Poisson, representado por

$$f(\mathbf{n}|\mu) = \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{e^{-\mu_{ij}}\mu_{ij}^{n_{ij}}}{n_{ij}!},$$

em que $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})'$ e $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})'$, com $n_{ij} \in \mathbb{N}_0$ e $\mu_{ij} \in \mathbb{R}^+$, $i, j = 1, 2$. Podemos testar a hipótese de interesse como sendo a de que a proporção de pessoas a favor é a mesma para as duas sub-populações:

$$H_0 : \frac{\mu_{11}}{\mu_{1.}} = \frac{\mu_{21}}{\mu_{2.}} \left(= \frac{\mu_{.1}}{\mu_{..}} \right),$$

onde $\mu_{.j} = \sum_i \mu_{ij}$, $\mu_{i.} = \sum_j \mu_{ij}$ e $\mu_{..} = \sum_{i,j} \mu_{ij}$. Podemos expressar H_0 também da

forma alternativa:

$$H_0 : \mu_{ij} = \frac{\mu_{i.}\mu_{.j}}{\mu_{..}} \quad , \quad \text{para} \quad i, j = 1, 2.$$

Estatégia 2: Neste caso fixa-se o total N de entrevistas que devem ser feitas antes de iniciar o processo de pesquisa. Ao final, classificam-se as respostas entre as sub-populações como na Tabela 3.2.

Tabela 3.2: Estratégia 2 - Número total de entrevistas a serem feitas é fixado de antemão.

Opinião	favor	contra	Total
sub-população 1	θ_{11}	θ_{12}	$\theta_{1.}$
sub-população 2	θ_{21}	θ_{22}	$\theta_{2.}$
Total	$\theta_{.1}$	$\theta_{.2}$	1

Definimos a probabilidade do indivíduo apresentar a característica (i, j) como $\theta_{ij} = P(X_1 = i, X_2 = j)$. Ou seja, cada indivíduo vai ser classificado como pertencendo a uma sub-população X_1 e apresentando uma das possíveis categorias de resposta X_2. Então, dos N entrevistados, n_{ij} deles pertencerão à sub-população i e darão a resposta j. Temos então combinações possíveis de valores para n_{ij}, sendo que $\sum_{i,j} n_{ij} = N$.

Vamos falar agora da frequência relativa de cada resposta $\hat{\theta}_{ij} = n_{ij}/N$. O vetor de probabilidades associado às proporções de cada categoria de resposta por estrato é dado por $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})'$, com $\sum_{i,j} \theta_{ij} = 1$. O vetor aleatório $\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22})$ têm distribuição Multinomial

$$P(X_{11} = n_{11}, X_{12} = n_{12}, X_{21} = n_{21}, X_{22} = n_{22}) = N! \prod_{i,j=1}^{2} \frac{\theta_{ij}^{n_{ij}}}{n_{ij}!} \quad , \quad \sum_{i,j} n_{ij} = N.$$

Se estamos interessados em saber se existe indepêndencia entre as respostas dadas no estrato 1 e 2 da população, fazemos um teste de independência,

$$H_0 : \theta_{ij} = \theta_{i.} \times \theta_{.j}.$$

Estratégia 3: Fixam-se, de antemão, o número N_1 de pessoas que serão entrevistadas no estrato 1 e o número N_2 de quantas no estrato 2. Aqui, os totais marginais das linhas estão fixados. Temos em cada linha uma variável aleatória X_i com distribuição

Binomial $(n_{i.},\theta_i)$. Mas, podemos tornar este exemplo mais geral e pensar que, ao invés de somente duas categorias de resposta, temos r categorias. Assim, em cada linha haverá uma distribuição Multinomial. Agora a probabilidade de um indivíduo apresentar a categoria de resposta j é condicional à linha em que ele se encontra: $\theta_{(i)j} = P(X_2 = j|X_1 = i)$.

Tabela 3.3: Estratégia 3 - Tamanho das subpopulações fixados de antemão.

Opinião	resposta 1	resposta 2	...	resposta r	Total
sub-população 1	n_{11}	n_{12}	...	n_{1r}	N_1
sub-população 2	n_{21}	n_{22}	...	n_{2r}	N_2
Total	$n_{.1}$	$n_{.2}$	...	$n_{.r}$	$N_1 + N_2$

$$P(X_{i1} = n_{i1}, X_{i2} = n_{i2}, ..., X_{ir} = n_{ir}) = N_i! \prod_{k=1}^{r} \frac{\theta_{(i)k}^{n_{ik}}}{n_{ik}!} \ , \ \sum_{j=1}^{r} n_{ij} = N_i.$$

Como as linhas são independentes, ficamos com o modelo de produto de distribuições Multinomiais dado por

$$f(\mathbf{n}|N, \pi) = \prod_{i=1}^{2} \left(N_i! \prod_{k=1}^{r} \frac{\theta_{(i)k}^{n_{ik}}}{n_{ik}!} \right).$$

Estamos trabalhando com o exemplo em que $i = 1, 2$, mas assim como aumentamos as categorias de resposta, podemos também pensar em s sub-populações independentes e estender para $i = 1, ..., s$.

3.2.1 Relação entre os modelos Poisson e Multinomial

Palmgren (1981) discute modelos para tabelas de contigência em que cada casela é uma variável aleatória Poisson independente. As variáveis do modelo pertencem a dois conjuntos: um de variáveis resposta e outro de variáveis explicativas. Para estudar como as variáveis resposta dependem das variáveis explicativas, condiciona-se os totais marginais observados nas variáveis explicativas.

Consideremos uma tabela de contigência com r categorias de resposta e s categorias para a variável explicativa. As caselas Y_{ij} da tabela são tratadas como variáveis independentes tendo distribuição Poisson, ou seja,

$$Y_{ij} \sim Poisson(\mu_{ij}), \quad i = 1, \ldots, r \quad i = 1, \ldots, s,$$

$$\sum Y_{ij} \sim Poisson(\sum \mu_{ij}) \ \text{ onde } \ \sum \mu_{ij} = \mu_{..} \ \text{ e}$$

$$P((Y_{11}, Y_{12}, ..., Y_{rs}) = (a_{11}, a_{12}, ..., a_{rs})| \sum Y_{ij} = n) = \frac{P(\mathbf{Y} = \mathbf{a} \bigcap \sum Y_{ij} = n)}{P(\sum Y_{ij} = n)},$$

de tal forma que estamos sujeitos à restrição $a_{rs} = n - a_{11} - a_{12} - ... - a_{ij} - ... - a_{rs}$ com $i = 1, ..., r$ e $j = 1, ..., s$. Uma vez que os Y_{ij} são independentes,:

$$\begin{aligned}
\frac{P(\mathbf{Y} = \mathbf{a} \bigcap \sum Y_{ij} = n)}{P(\sum Y_{ij} = n)} &= \frac{P(Y_{11} = a_{11})P(Y_{12} = a_{12})...P(Y_{rs} = n - a_{11} - a_{12} - ...)}{P(\sum Y_{ij} = n)} \\
&= \frac{e^{-\mu_{11}} \frac{(\mu_{11})^{a_{11}}}{a_{11}!} e^{-\mu_{12}} \frac{(\mu_{12})^{a12}}{a_{12}!} ... e^{-\mu_{rs}} \frac{(\mu_{rs})^{(n-a_{11}-...)}}{(n-a_{11}-...)!}}{e^{-\mu_{..}} \frac{(\mu_{..})^n}{n!}} \\
&= \frac{n!}{a_{11}!a_{12}!...a_{rs}!} \left(\frac{\mu_{11}}{\mu_{..}}\right)^{a_{11}} \left(\frac{\mu_{12}}{\mu_{..}}\right)^{a_{12}} ... \left(\frac{\mu_{rs}}{\mu_{..}}\right)^{a_{rs}}, \quad \text{tal que } \frac{\mu_{ij}}{\mu_{..}} = \theta_{ij}
\end{aligned}$$

Assim, temos que a distribuição de $\mathbf{Y} | \sum Y_{ij} = n$ é Multinomial com parâmetros n e $(\theta_{11}, \theta_{12}, ...\theta_{rs})$.

3.3 Regressão de Poisson e Superdispersão

A primeira abordagem possível para dados de contagem é fazer uma regressão de Poisson. Estimamos um MLG com distribuição Poisson e função de ligação canônica (o logaritmo da média). Freqüentemente, alguns problemas podem surgir, como superdispersão ou um número excessivo de zeros.

Existem vários artigos sugerindo soluções para estes problemas comuns. Em relação à superdispersão, Hinde e Demétrio (1998) apresentam um artigo bastante abrangente, com vários exemplos e técnicas alternativas de estimação. Para o caso do número excessivo de zeros, temos o artigo de Lambert (1992) para a chamada regressão de Poisson inflacionada de zeros (ZIP - *Zero Inflated Poisson*). Além disso, é bem provável estarmos diante de dados que misturem estes dois problemas clássicos.

No caso da distribuição de Poisson, a superdispersão é um fenômeno que aparece quando estamos diante de uma variabilidade dos dados maior do que a média. A hipótese inicial de $\phi = 1$ não mais se verifica e precisamos encontrar formas possíveis de lidar com a variabilidade extra. Basicamente temos três soluções possíveis:

1. abordagem bayesiana, assumindo que o parâmetro do modelo possui uma distribuição de probabilidade;
2. estimação por Quase-verossimilhança, incluindo um fator de dispersão diferente da unidade ou uma função de variância diversa;
3. modelos mistos com a separação de efeitos fixos e efeitos aleatórios para as variáveis explicativas. Este modelo é apresentado no próximo capítulo.

3.3.1 Modelo Binomial Negativa

Verificando que temos a presença da superdispersão, podemos levantar agora a hipótese de que, condicionada nos parâmetros, a variável resposta tem distribuição de Poisson. Ou seja, os parâmetros da Poisson são considerados como variáveis aleatórias com distribuição conhecida. Se a distribuição a priori dos parâmetros for Gama, chegamos à solução da Binomial Negativa para a distribuição marginal de Y.

A abordagem bayesiana também lida com tipos diferentes de variâncias: uma função linear da média ou uma função quadrática, por exemplo. Vamos assumir primeiro a hipótese de que estamos diante de variáveis $Y|Z \sim Poisson(Z)$ e que a distribuição a priori de Z é Gama com parâmetros $\phi\mu$ e ϕ, ou seja, $E(Z) = \mu$ e $Var(Z) = \mu/\phi$. Assim,

$$Y|Z \sim Poisson(Z) \ \text{, i.e., } \ f(y|z) = e^{-z}\frac{z^y}{y!}\mathbf{1}_{\{0,1,\ldots\}}(y),$$

$$Z \sim Gama(\phi\mu, \phi) \ \text{, i.e., } \ g(z) = \frac{\phi^{\phi\mu}}{\Gamma(\phi\mu)}z^{\phi\mu-1}e^{-\phi z}\mathbf{1}_{(0,\infty)}(z).$$

Para encontrar a distribuição marginal de Y integramos a distribuição conjunta de Ye Z em relação a Z, produzindo

$$\begin{aligned} f(y) &= \int_0^\infty f(y|z)g(z)dz = \int_0^\infty e^{-z}\frac{z^y}{y!}\frac{\phi^{\phi\mu}}{\Gamma(\phi\mu)}z^{\phi\mu-1}e^{-\phi z}dz \\ &= \frac{\phi^{\phi\mu}}{y!\Gamma(\phi\mu)}\int_0^\infty z^{y+\phi\mu-1}e^{-(\phi+1)z}dz \\ &= \frac{\phi^{\phi\mu}}{y!\Gamma(\phi\mu)}\frac{\Gamma(y+\phi\mu)}{(\phi+1)^{\phi\mu+y}}\int_0^\infty \frac{(\phi+1)^{\phi\mu+y}}{\Gamma(y+\phi\mu)}z^{y+\phi\mu-1}e^{-(\phi+1)z}dz. \end{aligned} \tag{3.1}$$

Temos que em (3.1) a integral é um, pois estamos integrando uma variável com distribuição Gama com parâmetros $y + \phi\mu$ e $\phi + 1$. Rearranjando os termos que se encontram à esquerda da integral, chegamos a uma distribuição Binomial Negativa com parâmetros $\phi\mu$ e $\phi/(\phi + 1)$, i.e.,

$$f(y) = \frac{(y+\phi\mu-1)!}{y!(\phi\mu-1)!}\left(\frac{\phi}{\phi+1}\right)^{\phi\mu}\left(\frac{1}{\phi+1}\right)^{y}\mathbf{1}_{\{0,1,\ldots\}}(y).$$

Desta forma,

$$E(Y) = \mu \ \ e \ \ Var(Y) = \mu\frac{(\phi+1)}{\phi}. \tag{3.2}$$

Outra possibilidade considera $Y|Z \sim Poisson(Z)$ com Z tendo distribuição a priori Gama, com parâmetros ν e μ/ν, tal que, $E(Z) = \mu$ e $Var(Z) = \mu^2/\nu$. A distribuição marginal de Y será novamente binomial negativa, mas com uma função de variância quadrática, i.e.,

$$Y|Z \sim Poisson(Z) \ \text{, i.e., } \ f(y|z) = e^{-z}\frac{z^y}{y!}\mathbf{1}_{\{0,1,\ldots\}}(y)$$

$$Z \sim Gama(\nu, \mu/\nu) \ \text{, i.e., } \ g(z) = \frac{1}{\Gamma(\nu)(\frac{\mu}{\nu})^{\nu}} z^{\nu-1} e^{-z/\frac{\mu}{\nu}} \mathbf{1}_{(0,\infty)}(z).$$

Procedemos da mesma maneira para encontrar a distribuição marginal de Y,

$$\begin{aligned} f(y) &= \int_0^{\infty} e^{-z} \frac{z^y}{y!} \frac{1}{\Gamma(\nu)(\frac{\mu}{\nu})^{\nu}} z^{\nu-1} e^{-z/\frac{\mu}{\nu}} dz \\ &= \frac{(\frac{\nu}{\mu})^{\nu}}{y!\Gamma(\nu)} \int_0^{\infty} z^{y+\nu-1} e^{-z\frac{\nu}{\mu}} dz \\ &= \frac{(\frac{\nu}{\mu})^{\nu}}{y!\Gamma(\nu)} \frac{\Gamma(y+\nu)}{(\frac{\nu}{\mu}+1)^{y+\nu}} \int_0^{\infty} \frac{(\frac{\nu}{\mu}+1)^{y+\nu}}{\Gamma(y+\nu)} z^{y+\nu-1} e^{-z(\frac{\nu}{\mu}+1)} dz. \end{aligned} \tag{3.3}$$

Da mesma forma, temos que em (3.3) o valor da integral de uma variável com distribuição Gama de parâmetros $y+\nu$ e $(\nu/\mu+1)$ é um. Rearranjando os termos à esquerda da integral, chegamos a uma outra distribuição Binomial Negativa, com parâmetros $\left(\nu, \frac{\frac{\nu}{\mu}}{\frac{\nu}{\mu}+1}\right)$:

$$f(y) = \frac{(y+\nu-1)!}{y!(\nu-1)!} \left(\frac{\frac{\nu}{\mu}}{\frac{\nu}{\mu}+1}\right)^{\nu} \left(\frac{1}{\frac{\nu}{\mu}+1}\right)^{y} \mathbf{1}_{\{0,1,\ldots\}}(y).$$

Neste caso,

$$E(Y) = \mu \ \ e \ \ Var(Y) = \mu + \frac{1}{\nu}\mu^2. \tag{3.4}$$

Podemos notar tanto em (3.2) quanto em (3.4) que temos a variância maior do que a média. Assim, uma solução alternativa seria estimar via máxima verossimilhança, um MLG usando a Binomial Negativa como distribuição. Questões mais detalhadas, referentes aos modelos de regressão da Binomial Negativa, aparecem em Lawless (1987).

3.3.2 Quase-verossimilhança

Outra abordagem para lidar com a presença da superdispersão nos dados são os modelos de Quase-verossimilhança, uma generalização dos MLG's. Agora não precisamos conhecer a distribuição da variável resposta, mas postular uma forma de relação entre a média e a variância como é tratada por Wedderburn (1974):

1. a média da variável é função dos parâmetros do modelo linear,i.e.: $E(Y) = \mu(\beta)$;

2. a variância da variável é função da média:$Var(Y) = \phi V(\mu)$.

Em seguida, define-se a função de Quase-verossimilhança por

$$Q(\mu; y) = \int_y^{\mu} \frac{y-t}{V(t)} dt,$$

de tal forma que

$$\frac{\partial Q(\mu;y)}{\partial \mu} = \frac{y-\mu}{V(\mu)}.$$

Por esta definição constatamos que a função de Quase-verossimilhança vai apresentar as mesmas propriedades de uma função de verossimilhança. Estas propriedades são apresentadas por Wedderburn como um teorema:

Teorema 1:

1. $E\left(\frac{\partial Q(\mu;y)}{\partial \mu}\right) = 0$;
2. $E\left(\frac{\partial Q(\mu;y)}{\partial \boldsymbol{\beta}}\right) = \mathbf{0}$;
3. $E\left(\frac{\partial Q(\mu;y)}{\partial \mu}\right)^2 = -E\left(\frac{\partial Q^2(\mu;y)}{\partial \mu^2}\right) = \frac{1}{V(\mu)}$;
4. $E\left(\frac{\partial Q(\mu;y)}{\partial \beta_i \beta_j}\right)^2 = -E\left(\frac{\partial Q^2(\mu;y)}{\partial \beta_i \beta_j}\right) = \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_i}\frac{\partial \mu}{\partial \beta_j}$;

A propriedade 1 do Teorema segue diretamente da definição:

$$E\left(\frac{\partial Q(\mu;y)}{\partial \mu}\right) = E\left(\frac{y-\mu}{V(\mu)}\right) = 0.$$

Para mostrar a propriedade 2 utilizamos o fato que μ é função de β, então pela regra da cadeia $\partial Q/\partial \beta_i = (\partial Q/\partial \mu)(\partial \mu/\partial \beta_i)$:

$$\begin{aligned} E\left(\frac{\partial Q}{\partial \beta_i}\right) &= E\left(\frac{\partial Q}{\partial \mu}\frac{\partial \mu}{\partial \beta_i}\right) \\ &= \frac{\partial \mu}{\partial \beta_i} E\left(\frac{\partial Q}{\partial \mu}\right) = 0. \end{aligned}$$

A propriedade 3 é um caso especial da propriedade 4. Para provar esta última, note que:

$$\begin{aligned} E\left(\frac{\partial Q}{\partial \beta_i}\frac{\partial Q}{\partial \beta_j}\right) &= \frac{\partial \mu}{\partial \beta_i}\frac{\partial \mu}{\partial \beta_j} E\left(\frac{\partial Q}{\partial \mu}\right)^2 \\ &= \frac{\partial \mu}{\partial \beta_i}\frac{\partial \mu}{\partial \beta_j} E\left(\frac{(y-\mu)^2}{(V(\mu))^2}\right) \\ &= \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_i}\frac{\partial \mu}{\partial \beta_j}. \end{aligned}$$

Também temos

$$\begin{aligned} E\left(\frac{\partial^2 Q}{\partial\beta_i\partial\beta_j}\right) &= E\left(\frac{\partial}{\partial\beta_j}\left[\frac{\partial Q}{\partial\beta_i}\right]\right) \\ &= E\left(\frac{\partial}{\partial\beta_j}\left[\frac{\partial Q}{\partial\mu}\frac{\partial\mu}{\partial\beta_i}\right]\right) \\ &= E\left(\frac{\partial}{\partial\beta_j}\left[\frac{(y-\mu)}{V(\mu)}\frac{\partial\mu}{\partial\beta_i}\right]\right) \\ &= \frac{1}{V(\mu)}E\left(\frac{\partial}{\partial\beta_j}(y-\mu)\frac{\partial\mu}{\partial\beta_i}+(y-\mu)\frac{\partial}{\partial\beta_j}\frac{\partial\mu}{\partial\beta_i}\right) \\ &= \frac{1}{V(\mu)}E\left(-\frac{\partial\mu}{\partial\beta_j}\frac{\partial\mu}{\partial\beta_i}+(y-\mu)\frac{\partial^2\mu}{\partial\beta_i\partial\beta_j}\right). \end{aligned}$$

Ao aplicar o operador esperança temos a igualdade procurada

$$-E\left(\frac{\partial^2 Q}{\partial\beta_i\partial\beta_j}\right)=\frac{1}{V(\mu)}\frac{\partial\mu}{\partial\beta_j}\frac{\partial\mu}{\partial\beta_i}.$$

Em seguida, Wedderburn apresenta um corolário que mostra que se sabemos qual é a distribuição da variável, e portanto conhecemos a verossimilhança $L(\mu;y)$, então a informação dada pela verossimilhança é maior do que a informação dada pela quase-verossimilhança.

$$-E\left(\frac{\partial Q^2(\mu;y)}{\partial\mu^2}\right)\leq -E\left(\frac{\partial L^2(\mu;y)}{\partial\mu^2}\right).$$

Da propriedade 1 do Teorema 1 temos que:

$$E\left(\frac{\partial Q(\mu;y)}{\partial\mu}\right)=E\left(\frac{Y-\mu}{V(\mu)}\right)=0 \quad \text{ou} \quad E(Y)=\mu.$$

Da propriedade 3 temos que:

$$E\left(\frac{\partial Q(\mu;y)}{\partial\mu}\right)^2=E\left(\frac{Y-\mu}{V(\mu)}\right)^2=\frac{E(Y-\mu)^2}{V(\mu)^2}=\frac{Var(Y)}{V(\mu)^2}=\frac{1}{V(\mu)} \quad \text{ou} \quad Var(Y)=V(\mu).$$

Como

$$\frac{\partial E(Y)}{\partial\mu}=1,$$

pela desigualdade de Cramer-Rao temos

$$Var(Y)\geq\frac{1}{E(\frac{\partial L(\mu;y)}{\partial\mu})^2},$$

e sabemos que

$$E\left(\frac{\partial L(\mu;y)}{\partial\mu}\right)^2=-E\left(\frac{\partial L^2(\mu;y)}{\partial\mu^2}\right).$$

Assim,

$$Var(Y) \geq \frac{-1}{E(\frac{\partial L^2(\mu;y)}{\partial \mu^2})} \quad \text{ou} \quad \frac{1}{Var(Y)} \leq -E\left(\frac{\partial L^2(\mu;y)}{\partial \mu^2}\right)$$

$$\text{mas,} \quad Var(Y) = V(\mu) \quad \text{e} \quad -E\left(\frac{\partial Q^2(\mu;y)}{\partial \mu^2}\right) = \frac{1}{V(\mu)},$$

e desta forma mostra-se que

$$-E\left(\frac{\partial Q^2(\mu;y)}{\partial \mu^2}\right) \leq -E\left(\frac{\partial L^2(\mu;y)}{\partial \mu^2}\right).$$

Wedderburn (1974) também destaca o fato de que quando estamos diante de distribuições que pertencem à família exponencial e com um único parâmetro, a quase-verossimilhança é idêntica à verossimilhança. Para a distribuição Binomial temos que $Var(Y) = \mu(1-\mu)$ então,

$$\begin{aligned}
Q(\mu;y) &= \int_y^\mu \frac{y-t}{t(1-t)}dt = \int_y^\mu (y-t)\left(\frac{1}{t} - \frac{1}{(1-t)}\right)dt \\
&= \left[y\ln(t) - t + t - (y-1)\ln(1-t)\right]_y^\mu = \left[y\ln\left(\frac{t}{1-t}\right) + \ln(1-t)\right]_y^\mu \\
&= y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu) - y\ln\left(\frac{y}{1-y}\right) - \ln(1-y) \\
&\propto y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu).
\end{aligned}$$

Para a distribuição Poisson temos $Var(Y) = \mu$ e,

$$\begin{aligned}
Q(\mu;y) &= \int_y^\mu \frac{y-t}{t}dt = \left[y\ln(t) - t\right]_y^\mu \\
&= y\ln(\mu) - \mu - y\ln(y) + y \\
&\propto y\ln(\mu) - \mu.
\end{aligned}$$

Não são todas as funções de quase-verossimilhança que correspondem a distribuições conhecidas. Se tivermos uma função de variância dada por $Var(Y) = \mu^2(1-\mu)^2$

$$\begin{aligned}
Q(\mu;y) &= \int_y^\mu \frac{y-t}{t^2(1-t)^2}dt = \int_y^\mu \left(\frac{2y-1}{t} + \frac{y}{t^2} + \frac{2y-1}{1-t} + \frac{y-1}{(1-t)^2}\right)dt \\
&= \left[(2y-1)\ln(t) - \frac{y}{t} - (2y-1)\ln(1-t) + \frac{y-1}{1-t}\right]_y^\mu \\
&= \left[(2y-1)\ln\left(\frac{t}{1-t}\right) - \frac{y}{t} - \frac{1-y}{1-t}\right]_y^\mu \\
&\propto (2y-1)\ln\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}.
\end{aligned}$$

Neste caso a função de quase-verossimilhança não corresponde a nenhuma distribuição conhecida da família exponencial.

3.3.3 Número excessivo de zeros - ZIP

Lambert (1982) em seu artigo discute a existência de defeitos em peças fabricadas. O número de peças defeituosas tem uma distribuição de Poisson com média λ. Assim, em uma amostra de n peças esperaríamos encontrar $ne^{-\lambda}$ peças que não apresentassem nenhum defeito. Porém, podemos encontrar um número ainda maior de peças não-defeituosas, pois estaríamos diante de um "estado perfeito" em que há itens extremamente resistentes a defeitos.

Ridout et al. (1998) referem-se aos "zeros estruturais", que são inevitáveis e aos "zeros amostrais" que podem ocorrer ao acaso. Num estudo que conte lesões em plantas, uma planta pode não apresentar a lesão por que é resistente àquela doença ("zero estrutural") ou simplesmente por que o agente causador da doença não recaiu sobre aquela planta ("zero amostral").

Quando há um número excessivo de zeros, a solução mais freqüente é a de estimarmos um modelo que misture a Poisson com uma distribuição que leve em conta o excesso de zeros. No caso, tem-se a hipótese que, com probabilidade p nossa variável resposta assume o valor zero e com probabilidade $(1-p)$ assume o valor de uma variável aleatória com distribuição de Poisson com média λ. Assim a distribuição de Y é dada por

$$P[Y=y] = \begin{cases} p + (1-p)e^{-\lambda}, & \text{se } \ y = 0, \\ \\ (1-p)e^{-\lambda}\frac{\lambda^y}{y!}, & \text{se } \ y > 0. \end{cases} \tag{3.5}$$

Para esta densidade, a esperança é

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} yP[Y=y] = \sum_{y=1}^{\infty} yP[Y=y] \\ &= \sum_{y=1}^{\infty} y(1-p)e^{-\lambda}\frac{\lambda^y}{y!} \\ &= (1-p)e^{-\lambda}\lambda \sum_{(y-1)=0}^{\infty} \frac{\lambda^{(y-1)}}{(y-1)!} \\ &= (1-p)e^{-\lambda}\lambda e^{\lambda} = (1-p)\lambda = \mu. \end{aligned}$$

Para encontrarmos a variância precisamos calcular $E(Y^2)$:

$$\begin{aligned}
E(Y^2) &= \sum_{y=0}^{\infty} y^2 P[Y=y] = \sum_{y=1}^{\infty} y^2 P[Y=y] = \sum_{y=1}^{\infty} y^2 (1-p) e^{-\lambda} \frac{\lambda^y}{y!} \\
&= (1-p) \sum_{y-1=0}^{\infty} y e^{-\lambda} \frac{\lambda^y}{(y-1)!} \\
&\overset{(y-1=m)}{=} (1-p) \sum_{m=0}^{\infty} (m+1) e^{-\lambda} \frac{\lambda^{m+1}}{m!} \\
&= (1-p) \left(\sum_{m=0}^{\infty} m e^{-\lambda} \lambda \frac{\lambda^m}{m!} + \sum_{m=0}^{\infty} e^{-\lambda} \lambda \frac{\lambda^m}{m!} \right) \\
&= (1-p) \left(\lambda \sum_{m=0}^{\infty} m e^{-\lambda} \frac{\lambda^m}{m!} + \lambda \sum_{m=0}^{\infty} e^{-\lambda} \frac{\lambda^m}{m!} \right) \\
&= (1-p)(\lambda^2 + \lambda).
\end{aligned}$$

De tal forma que

$$\begin{aligned}
Var(Y) &= E(Y^2) - (E(Y))^2 = (1-p)(\lambda^2+\lambda) - ((1-p)\lambda)^2 \\
&= (1-p)\lambda + \frac{p}{1-p}((1-p)\lambda)^2 \\
&= \mu + \frac{p}{1-p}\mu^2.
\end{aligned}$$

Note que a variância da mistura é maior que a média da distribuição. Quanto maior a probabilidade do excesso de zeros, maior a variância da variável. À medida que p se aproxima de zero, a variância se aproxima de μ, ou seja, voltamos a lidar somente com uma distribruição Poisson padrão.

Vamos apresentar os modelos ZIP como descritos por Lambert (1982). Sejam as variáveis aleatórias independentes $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ com

$$Y_i = 0 \ \text{ com probabilidade } \ p_i \ \text{ e}$$

$$Y_i \sim Poisson(\lambda_i) \ \text{ com probabilidade } \ 1 - p_i.$$

Além disso, Os parâmetros $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)$ e $\mathbf{p} = (p_1, p_2, ..., p_n)$ satisfazem

$$\log(\boldsymbol{\lambda}_i) = \mathbf{X}\boldsymbol{\beta} \ \text{ e } \ \log\left(\frac{\mathbf{p}_i}{1-\mathbf{p}_i}\right) = \mathbf{Z}\boldsymbol{\gamma}, \tag{3.6}$$

e $\mathbf{X}$ e $\mathbf{Z}$ são as matrizes de variáveis explicaticas. Estas variáveis podem ou não ser idênticas. Se forem, pode-se pensar em $\mathbf{p}$ como função de $\boldsymbol{\lambda}$. Esta expressão vai estar condicionada à função de ligação escolhida para a probabilidade de zeros. O chamado

modelo ZIP(τ) padrão é obtido quando tem-se como função de ligação o logaritmo para a média da Poisson e o logito para a probabilidade de zeros:

$$\log(\boldsymbol{\lambda}_i) = \mathbf{X}\boldsymbol{\beta} \quad \text{e} \quad \log\left(\frac{\mathbf{p}_i}{1-\mathbf{p}_i}\right) = -\tau\mathbf{X}\boldsymbol{\beta}. \tag{3.7}$$

Há um outro parâmetro desconhecido a estimar (τ), mas em contrapartida, em relação às covariáveis, temos somente um vetor $\boldsymbol{\beta}$. A expressão de p_i como função de λ_i é dada por

$$\begin{aligned}
\log(\lambda_i) &= -\frac{1}{\tau}\log\left(\frac{p_i}{1-p_i}\right), \\
\log(\lambda_i^{-\tau}) &= \log\left(\frac{p_i}{1-p_i}\right), \\
\lambda_i^{-\tau} &= \left(\frac{p_i}{1-p_i}\right), \\
p_i &= \frac{\lambda_i^{-\tau}}{1+\lambda_i^{-\tau}} = \frac{1}{1+\lambda_i^{\tau}}, \text{ para todo } i = 1..., n.
\end{aligned}$$

Como a função logito é simétrica em torno de 0,5, Lambert (1982) sugere outras funções de ligação possíveis que lidam com a assimetria de $\mathbf{p}$. Temos por exemplo a função log-log:

$$\begin{aligned}
\log(-\log(p_i)) &= \tau\mathbf{X}_i\boldsymbol{\beta}, \\
\log(-\log(p_i)) &= \tau\log(\lambda_i), \\
-\log(p_i) &= \lambda_i^{\tau} \quad \text{ou} \\
p_i &= e^{-\lambda_i^{\tau}}.
\end{aligned}$$

Ou a função log-log complementar

$$\begin{aligned}
\log(-\log(1-p_i) &= -\tau\mathbf{X}_i\boldsymbol{\beta} = -\tau\log(\lambda_i), \\
-\log(1-p_i) &= \lambda_i^{-\tau}, \\
1-p_i &= e^{-\lambda_i^{-\tau}} \quad \text{ou} \\
p_i &= 1 - e^{-\lambda_i^{-\tau}}.
\end{aligned}$$

A Tabela 3.4 ilustra o que acontece com a probabilidade de zeros de acordo com as possíveis combinações de τ e λ_i.

No caso em que $\tau < 0$, a média da Poisson aumenta à medida que o excesso de zeros fica mais provável.

A função de verossimilhança do modelo ZIP é dada por

$$L(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{i=1}^{n}\left\{\left(p_i + (1-p_i)e^{-\lambda_i}\right)\mathbf{1}_{y_i=0} + \left((1-p_i)e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}\right)\mathbf{1}_{y_i>0}\right\}.$$

Tabela 3.4: Comportamento de τ e λ para diferentes funções de ligação.

	logito	log-log	log-log compl.
p_i	$\frac{1}{1+\lambda_i^{\tau}}$	$e^{-\lambda_i^{\tau}}$	$1-e^{-\lambda_i^{-\tau}}$
$\tau>0 \quad \lambda_i \to \infty$	$p_i \to 0$	$p_i \to 0$	$p_i \to 0$
$\tau \to \infty \quad \lambda_i$ fixo	$p_i \to 0$	$p_i \to 0$	$p_i \to 0$
$\tau \to -\infty$	$p_i \to 1$	$p_i \to 1$	$p_i \to 1$

Arranjando os termos em relação às funções indicadoras, o logaritmo da função de verossimilhança fica

$$\begin{aligned}\ell(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) &= \log\left\{\prod_{y_i=0}(p_i+(1-p_i)e^{-\lambda_i})\prod_{y_i>0}(1-p_i)e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}\right\} \\ &= \log\left\{\prod_{i=1}^{n}(1-p_i)\prod_{y_i=0}(\frac{p_i}{1-p_i}+e^{-\lambda_i})\prod_{y_i>0}e^{-\lambda_i}\lambda_i^{y_i}\prod_{y_i>0}\frac{1}{y_i!}\right\}.\end{aligned}$$

Sabemos que

$$\log(\lambda_i)=\mathbf{X}_i\boldsymbol{\beta} \quad \text{e} \quad \log\left(\frac{p_i}{1-p_i}\right)=\mathbf{Z}_i\boldsymbol{\gamma},$$

$$\lambda_i=e^{\mathbf{X}_i\boldsymbol{\beta}} \quad , \quad \frac{p_i}{1-p_i}=e^{\mathbf{Z}_i\boldsymbol{\gamma}} \quad \text{e} \quad 1-p_i=\frac{1}{1+\mathbf{Z}_i\boldsymbol{\gamma}},$$

$$= -\log(\prod_{i=1}^{n}(1+e^{\mathbf{Z}_i\boldsymbol{\gamma}}))+\log(\prod_{y_i=0}(e^{\mathbf{Z}_i\boldsymbol{\gamma}}+e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}))+\log(\prod_{y_i>0}e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}e^{\mathbf{X}_i\boldsymbol{\beta}y_i})-\log(\prod_{y_i>0}y_i!)$$

$$= -\sum_{i=1}^{n}\log(1+e^{\mathbf{Z}_i\boldsymbol{\gamma}})+\sum_{y_i=0}\log(e^{\mathbf{Z}_i\boldsymbol{\gamma}}+e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}))+\sum_{y_i>0}(-e^{\mathbf{X}_i\boldsymbol{\beta}}+\mathbf{X}_i\boldsymbol{\beta}y_i)-\sum_{y_i>0}\log(y_i!).$$

A maximização da função de verossimilhança fica bastante complicada, uma vez que temos o termo com o logaritmo da soma de exponenciais. Lambert (1982) apresenta então uma solução que envolve o algoritmo EM, que costuma ser usado na maximização de verossimilhanças incompletas. Supõe-se poder identificar qual contagem de valor zero vem da Poisson e qual vem da probabilidade de zeros, de tal forma que chegamos a uma verossimilhança mais simples com dois termos que podem ser maximizados separadamente.

Além dos modelos sugeridos por Lambert (1982), trabalhamos com um outro modelo em que a probabilidade de zeros é vista como um parâmetro a ser estimado como apresentado em Böhning et al.(1999). Ou seja, há um p comum a todas as observações e um modelo linear para o logaritmo da média Poisson. O modelo é dado por (3.5) com

$$\log(\lambda)=\mathbf{X}\boldsymbol{\beta}. \tag{3.8}$$

Neste caso a verossimilhança é dada por

$$L(\mathbf{y}; p, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ (p + (1-p)e^{-\lambda_i})\mathbf{1}_{y_i=0} + ((1-p)e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!})\mathbf{1}_{y_i>0} \right\}.$$

Tomando-se o logaritmo da verossimilhança, temos que

$$\begin{aligned}
\ell(\mathbf{y}_i; p, \boldsymbol{\beta}) &= \log\left\{ \prod_{y_i=0} (p + (1-p)e^{-\lambda_i}) \prod_{y_i>0} (1-p)e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!} \right\} \\
&= \log\left\{ \prod_{i=1}^{n}(1-p) \prod_{y_i=0} (\frac{p}{1-p} + e^{-\lambda_i}) \prod_{y_i>0} e^{-\lambda_i}\lambda_i^{y_i} \prod_{y_i>0} \frac{1}{y_i!} \right\} \\
&= n\log(1-p) + \sum_{y_i=0} \log(\frac{p}{1-p} + e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}) + \log(\prod_{y_i>0} e^{-e^{X_i\boldsymbol{\beta}}} e^{\mathbf{X}_i\boldsymbol{\beta} y_i}) - \sum_{y_i>0} \log y_i! \\
&= n\log(1-p) + \sum_{y_i=0} \log(\frac{p}{1-p} + e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}) + \sum_{y_i>0} (-e^{\mathbf{X}_i\boldsymbol{\beta}} + \mathbf{X}_i\boldsymbol{\beta} y_i) - \sum_{y_i>0} \log y_i!.
\end{aligned}$$

Para encontrar o vetor de parâmetros $(p, \boldsymbol{\beta})$ maximiza-se o logaritmo de verossimilhança, igualhando suas primeiras derivadas à zero, i.e.,

$$\begin{aligned}
\frac{\partial \ell(p, \boldsymbol{\beta}; \mathbf{y})}{\partial p} &= \frac{n}{p-1} + \frac{1}{1-p} \sum_{y_i=0} \frac{1}{p + e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}(1-p)} = 0 \\
\frac{\partial \ell(p, \boldsymbol{\beta}; \mathbf{y})}{\partial \beta_j} &= \sum_{y_i=0} \frac{e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}(-e^{\mathbf{X}_i\boldsymbol{\beta}})x_{ij}}{\frac{p}{1-p} + e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}} + \sum_{y_i>0} (-e^{\mathbf{X}_i\boldsymbol{\beta}}x_{ij} + \frac{y_i}{\mathbf{X}_i\boldsymbol{\beta}}x_{ij}) = 0.
\end{aligned}$$

Uma vez que estamos diante de funções não lineares, pode-se, na estimação, recorrer ao uso dos vários algoritmos de maximização existentes. Segundo o manual do SAS, não existe um algoritmo que sempre encontre o valor ótimo para o problema de maximizar funções não lineares, por isto o programa oferece várias opções de escolha da técnica de maximização.

A função escore, que corresponde às primeiras derivadas parciais do logaritmo da verossimilhança é dada por

$$U(p, \boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial \ell(p, \boldsymbol{\beta}; y)}{\partial p} \\ \frac{\partial \ell(p, \boldsymbol{\beta}; y)}{\partial \boldsymbol{\beta}_j} \end{pmatrix}.$$

Para escrevermos a matriz de Informação de Fisher necessitamos calcular todas as deri-

vadas de segunda ordem, e temos nesse caso,

$$\frac{\partial \ell^2(p,\boldsymbol{\beta};\mathbf{y})}{\partial p^2} = \frac{-n}{(p-1)^2} - \frac{1}{1-p}\sum_{y_i=0}\frac{1-2(p+(1-p)e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}})}{(p+e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}(1-p))^2}\frac{1}{(1-p)^2},$$

$$\begin{aligned}\frac{\partial \ell^2(p,\boldsymbol{\beta};\mathbf{y})}{\partial \beta_j^2} &= \sum_{y_i=0}\frac{(e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}e^{2\mathbf{X}_i\boldsymbol{\beta}}x_{ij}^2 - e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}e^{\mathbf{X}_i\boldsymbol{\beta}}x_{ij}^2)(\frac{p}{1-p}+e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}) + e^{-2e^{\mathbf{X}_i\boldsymbol{\beta}}}e^{2\mathbf{X}_i\boldsymbol{\beta}}x_{ij}^2}{(\frac{p}{1-p}+e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}})^2} \\ &+ \sum_{y_i>0}(-e^{\mathbf{X}_i\boldsymbol{\beta}}x_{ij}^2 + \frac{y_i}{(\mathbf{X}_i\boldsymbol{\beta})^2}x_{ij}^2),\end{aligned}$$

$$\frac{\partial \ell^2(p,\boldsymbol{\beta};\mathbf{y})}{\partial p\partial \beta_j} = \sum_{y_i=0}\frac{e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}e^{\mathbf{X}_i\boldsymbol{\beta}}x_{ij}}{(p+(1-p)e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}})^2},$$

$$\begin{aligned}\frac{\partial \ell^2(p,\boldsymbol{\beta};\mathbf{y})}{\partial \beta_j\partial \beta_k} &= \sum_{y_i=0}\frac{e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}e^{\mathbf{X}_i\boldsymbol{\beta}}x_{ij}x_{ik}(e^{\mathbf{X}_i\boldsymbol{\beta}}-1)(\frac{p}{1-p}+e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}}) + e^{-2e^{\mathbf{X}_i\boldsymbol{\beta}}}e^{2\mathbf{X}_i\boldsymbol{\beta}}x_{ij}x_{ik}}{(\frac{p}{1-p}+e^{-e^{\mathbf{X}_i\boldsymbol{\beta}}})^2} \\ &+ \sum_{y_i>0}(-e^{\mathbf{X}_i\boldsymbol{\beta}}x_{ij}x_{ik} + \frac{y_i}{(\mathbf{X}_i\boldsymbol{\beta})^2}x_{ij}x_{ik}).\end{aligned}$$

A matriz de Informação de Fisher será portanto

$$-E\left(\begin{pmatrix} \frac{\partial \ell^2(p,\boldsymbol{\beta};y)}{\partial p^2} & \frac{\partial \ell^2(p,\boldsymbol{\beta};y)}{\partial p\partial \boldsymbol{\beta}_j} \\ \frac{\partial \ell^2(p,\boldsymbol{\beta};y)}{\partial p\partial \boldsymbol{\beta}_j} & \frac{\partial \ell^2(p,\boldsymbol{\beta};y)}{\partial \beta_j\partial \boldsymbol{\beta}_k} \end{pmatrix}\right).$$

Böhning et al. (1999) apresentam a verossimilhança p homogêneo. Considere a densidade da variável aleatória com ZIP p homogêneo,

$$P[Y=y] = (p+(1-p)e^{-\lambda})\mathbf{1}_{\{y=0\}} + ((1-p)e^{-\lambda}\frac{\lambda^y}{y!})\mathbf{1}_{\{y>0\}}.$$

No intuito de facilitar as contas, usaremos $q = 1-p$ e chamaremos de n_0 o número de todos os valores $y = 0$ que se encontram na amostra. Os valores diferentes de zero são chamados de n_y para $y = 1, ..., m$. Aqui m é o maior valor não-nulo que y assume na amostra de n indivíduos e $\sum_{y=1}^m n_y = n - n_0$. A verossimilhança pode ser escrita como

$$\begin{aligned}L(\mathbf{y};q,\lambda) &= \prod_{i=1}^n\left\{(1-q+qe^{-\lambda})\mathbf{1}_{y_i=0} + (qe^{-\lambda}\frac{\lambda^{y_i}}{y_i!})\mathbf{1}_{y_i>0}\right\} \\ &= \prod_{y_i=0}(1-q+qe^{-\lambda})\prod_{y_i>0}qe^{-\lambda}\frac{\lambda^{y_i}}{y_i!}.\end{aligned}$$

Tomando-se o logaritmo da verossimilhança, obtém-se

$$\ell(\mathbf{y}; q, \lambda) = n_0 \ln(1 - q + qe^{-\lambda}) + \sum_{y=1}^{m} n_y \ln\left(qe^{-\lambda}\frac{\lambda^{y_i}}{y_i!}\right).$$

Derivando o logaritmo da verossimilhança temos a função escore

$$U(q, \lambda) = \begin{pmatrix} \frac{\partial \ell(\mathbf{y};q,\lambda)}{\partial q} \\ \frac{\partial \ell(\mathbf{y};q,\lambda)}{\partial \lambda} \end{pmatrix}.$$

Para maximizar a verossimilhança tomamos

$$\frac{\partial \ell(\mathbf{y}; q, \lambda)}{\partial q} = n_0 \frac{(e^{-\lambda} - 1)}{1 - q + qe^{-\lambda}} + \frac{n - n_0}{q} = 0, \tag{3.9}$$

$$\frac{\partial \ell(\mathbf{y}; q, \lambda)}{\partial \lambda} = n_0 \frac{(-qe^{-\lambda})}{1 - q + qe^{-\lambda}} - (n - n_0) + \frac{n\bar{y}}{\lambda} = 0. \tag{3.10}$$

Na equação (3.9) encontramos $\hat{q}$ resolvendo

$$n_0 \frac{(e^{-\lambda} - 1)}{1 - q + qe^{-\lambda}} + \frac{n - n_0}{q} = 0,$$

de tal forma que

$$qn_0(e^{-\lambda} - 1) + (n - n_0)(1 - q + qe^{-\lambda}) = 0$$

$$qn(1 - e^{-\lambda}) = n - n_0 \quad \text{ou} \quad \hat{q} = \frac{1 - \frac{n_0}{n}}{1 - e^{-\hat{\lambda}}}.$$

Na equação (3.10) encontramos $\hat{\lambda}$, deduzindo da equação anterior que

$$q - qe^{-\lambda} = \frac{n - n_0}{n} \quad \text{e} \quad -qe^{-\lambda} = \frac{n - n_0 - nq}{n}.$$

Substituindo estas igualdades somente no primeiro termo de (3.10) ficamos com

$$n_0 \frac{(-qe^{-\lambda})}{1 - q + qe^{-\lambda}} = \frac{n_0 \frac{n - n_0 - nq}{n}}{1 - \frac{n - n_0}{n}} = n - n_0 - nq.$$

Voltando a (3.10) ficamos com

$$n - n_0 - nq - (n - n_0) + \frac{n\bar{y}}{\lambda} = 0 \quad \text{ou} \quad \hat{\lambda} = \frac{\bar{y}}{\hat{q}}$$

Chegamos então a um sistema não-linear com duas equações:

$$\hat{q} = \frac{1 - \frac{n_0}{n}}{1 - e^{-\hat{\lambda}}} \quad \text{e} \quad \hat{\lambda} = \frac{\bar{y}}{\hat{q}}$$

Sem nos esquecer que $\hat{\lambda} = e^{\mathbf{x}\hat{\boldsymbol{\beta}}}$.

Os diferentes métodos de otimização disponíveis no programa SAS se diferenciam basicamente por aqueles que calculam diretamente as derivadas de primeira e/ou segunda ordem e aqueles que usam aproximações. Assim como cada método possui critérios próprios para determinar a convergência. A escolha do melhor método geralmente envolve tentativa e erro, mas, no caso de conjuntos de dados menos complicados, os resultados obtidos pelos diferentes métodos são muito próximos.

Capítulo 4

Modelos Lineares Generalizados Mistos

Os modelos lineares discutidos no capítulo anterior são modelos de efeitos fixos, ou seja, conhecemos a matriz $\mathbf{X}$ e estimamos os efeitos fixos representados pelo vetor $\boldsymbol{\beta}$. Os efeitos fixos modelam a média de $\mathbf{Y}$ no caso dos modelos lineares normais e modelam uma função da média de $\mathbf{Y}$ no caso dos MLGs. Neste caso, supõe-se que não há correlação entre as respostas. Mas, na presença de uma estrutura de correlação entre as respostas, pode-se trabalhar com os modelos mistos, que incluem os chamados efeitos aleatórios.

A definição de quais efeitos serão tratados como fixos ou aleatórios depende do problema com o qual estamos lidando e, principalmente, da amostragem dos dados. Por exemplo, queremos testar o efeito de diferentes procedimentos cirúrgicos, os quais chamaremos de tratamentos. Além disso, dispomos destas informações para três hospitais determinados. Podemos testar a eficiência dos diversos tratamentos, assim como o efeito dos hospitais e possivelmente, de uma interação entre tratamentos e hospitais. Neste caso, estamos diante de dois efeitos fixos: tratamentos e hospitais. Agora, suponha que ao invés de determinar os hospitais, faz-se uma escolha aleatória de alguns deles na cidade. Neste caso, os tratamentos continuam sendo considerados como efeitos fixos, mas agora os hospitais serão efeitos aleatórios.

McCullogh e Searle (2001) propõem uma "árvore de decisão" para decidir se um efeito deve ser considerado como fixo ou aleatório. Se for razoável assumir que os níveis de um efeito vêm de uma distribuição conhecida, então este deve ser tratado como aleatório. Caso contrário, o efeito deve ser considerado fixo. Para os efeitos aleatórios existem dois focos de interesse possíveis. Se houver interesse somente na distribuição dos efeitos aleatórios, então estima-se os parâmetros de sua distribuição. Se além disso houver interesse nos valores observados dos efeitos aleatórios, então calcula-se seus preditores.

Iniciaremos com modelos mistos com variável resposta com distribuição normal, e depois introduziremos os modelos mistos generalizados.

4.1 Modelos Mistos Lineares

Considere o modelo de efeitos fixos adicionado de efeitos aleatórios, isto é,

$$E[\mathbf{Y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \quad \text{e} \quad Var[\mathbf{Y}|\mathbf{u}] = \mathbf{R}.$$

Note que adicionamos a matriz $\mathbf{Z}$, que é uma matriz conhecida, e o vetor $\mathbf{u}$ de efeitos aleatórios. Além disso, note que estamos diante de uma esperança condicional no vetor $\mathbf{u}$ observado. Precisamos também especificar que

$$\mathbf{u} \sim (\mathbf{0}, \mathbf{D}),$$

ou seja, $E[\mathbf{u}] = \mathbf{0}$ e $Var[\mathbf{u}] = \mathbf{D}$. Dessa forma a esperança e a variância de $\mathbf{Y}$ são dadas por:

$$E[\mathbf{Y}] = E\{E[\mathbf{Y}|\mathbf{u}]\} = E\{\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\} = \mathbf{X}\boldsymbol{\beta},$$

$$\begin{aligned} Var[\mathbf{Y}] &= Var\{E[\mathbf{Y}|\mathbf{u}]\} + E\{Var[\mathbf{Y}|\mathbf{u}]\} \\ &= Var[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}] + E[\mathbf{R}] \\ &= \mathbf{Z}Var[\mathbf{u}]\mathbf{Z}' + \mathbf{R} \\ &= \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R} = \mathbf{V}. \end{aligned}$$

Note que os efeitos fixos entram somente na média de $\mathbf{Y}$, enquanto a matriz de efeitos aleatórios e a variância aparecem somente na expressão da variância de $\mathbf{Y}$.

No emprego de modelos de efeitos mistos existem várias hipóteses que podem ser feitas quanto à estrutura de variância-covariância, ver McCullogh e Searle (2001).

4.2 Modelos Lineares Generalizados Mistos

A especificação do modelo generalizado misto também começa com a distribuição condicional de $\mathbf{Y}$ dado $\mathbf{u}$. Dado $\mathbf{u}$, o vetor de respostas $\mathbf{Y}$ consiste de elementos independentes com distribuição pertencente à família exponencial:

$$Y_i|\mathbf{u} \sim f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}),\ i = 1, ..., n, \quad \text{onde}$$

$$f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) = \exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}.$$

Aqui estamos modelando uma função (função de ligação) da média da distribuição, $\boldsymbol{\mu}$, ou seja, $g(\mu_i) = x_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}$, onde $E[Y_i|\mathbf{u}] = \mu_i$.

Necessitamos também definir uma distribuição para os efeitos aleatórios ($\mathbf{u} \sim f_{\mathbf{U}}(\mathbf{u})$). Da mesma forma como nos modelos lineares normais, podemos encontrar a distribuição

marginal de $\mathbf{Y}$. Porém, não estamos mais diante de modelos puramente lineares, o que pode gerar funções mais complexas, como por exemplo no caso $E[Y_i]$, onde

$$E[Y_i] = E[E[Y_i|\mathbf{u}]] = E[\mu_i] = E[g^{-1}(x_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})].$$

Note que para o caso da distribuição de Poisson, em que $g(\mu) = \log(\mu)$ e $g^{-1}(x) = \exp(x)$,

$$\begin{aligned} E[Y_i] &= E[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] \\ &= \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}E[\exp\{\mathbf{z}_i'\mathbf{u}\}] \\ &= \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}M_u(\mathbf{z}_i), \end{aligned}$$

em que $M_u(\mathbf{z}_i)$ é a função geratriz de momentos de $\mathbf{u}$ calculada em $\mathbf{z}_i$. Simplificando, podemos supor que $u_i \sim N(0, \sigma_u^2)$ e que cada linha da matriz $\mathbf{Z}$ tem uma única entrada unitária, com todo o restante igual a zero. Sabemos que a função geratriz de momentos para uma normal com média zero é $\exp\{\sigma_u^2/2\}$, e portanto temos

$$E[Y_i] = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\exp\{\frac{\sigma_u^2}{2}\} = e^{\mathbf{x}_i'\boldsymbol{\beta}+\frac{\sigma_u^2}{2}}, \text{ ou}$$

$$\log E[Y_i] = \mathbf{x}_i'\boldsymbol{\beta} + \frac{\sigma_u^2}{2}.$$

Uma outra maneira de encontrar a esperança marginal é a partir do modelo misto. Sabemos que $Y_{it}|u_i \sim Poisson(\mu_{it})$. Estamos nos referindo à observação t no grupo i em que $t = 1, ..., n$ e i engloba os *clusters* ou grupos de fatores distintos. Escolhendo $g(\mu_{it}) = \log(\mu_{it})$ o modelo misto é dado por

$$\log[E[Y_{it}|u_i]] = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i \quad \text{ou} \quad E[Y_{it}|u_i] = e^{\mathbf{x}_{it}'\boldsymbol{\beta}+u_i}.$$

Usando a propriedade da esperança condicional,

$$E[Y_{it}] = E[E[Y_{it}|u_i]] = E[e^{\mathbf{x}_{it}'\boldsymbol{\beta}+u_i}] = e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\frac{\sigma_u^2}{2}}$$

e

$$\begin{aligned} Var[Y_{it}] &= E[Var[Y_{it}|\mathbf{u}]] + Var[E[Y_{it}|\mathbf{u}]] \\ &= E[\exp\{\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] + Var[\exp\{\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] \\ &= E[\exp\{\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] + E[[\exp\{\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}]^2] - [E[\exp\{\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}]]^2 \\ &= \exp\{-\mathbf{x}_{it}'\boldsymbol{\beta}\}M_u(\mathbf{z}_i)) + \exp\{2\mathbf{x}_{it}'\boldsymbol{\beta}\}(M_u(2\mathbf{z}_i) - (M_u(\mathbf{z}_i))^2. \end{aligned}$$

Continuando com a ilustração da distribuição de Poisson,

$$\begin{aligned} Var[Y_{it}] &= E[Var[Y_{it}|u_i]] + Var[E[Y_{it}|u_i]] = E[e^{\mathbf{x}_{it}'\boldsymbol{\beta}+u_i}] + Var[e^{\mathbf{x}_{it}'\boldsymbol{\beta}+u_i}] \\ &= e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\frac{\sigma_u^2}{2}} + e^{2\mathbf{x}_{it}'\boldsymbol{\beta}}Var[e^{u_i}] \\ &= e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\frac{\sigma_u^2}{2}} + e^{2\mathbf{x}_{it}'\boldsymbol{\beta}}(E(e^{2u_i}) - [E(e^{u_i})]^2) \\ &= e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\frac{\sigma_u^2}{2}} + e^{2\mathbf{x}_{it}'\boldsymbol{\beta}}(e^{2\sigma^2} - e^{\sigma^2}) \\ &= E[Y_{it}] + E[Y_{it}]^2(e^{\sigma^2} - 1). \end{aligned}$$

Embora a distribuição de Y_{it} dado u_i seja Poisson, a distribuição marginal de Y_{it} pode não ser. Neste exemplo, podemos verificar que a variância marginal de Y_{it} é somada a um termo positivo, o que indica a presença de superdispersão (a variância é maior que a média). Neste sentido, podemos pensar nos efeitos aleatórios como uma maneira alternativa de modelar a superdispersão.

Podemos tentar escrever a função densidade marginal de Y, ou seja,

$$\begin{aligned} f(Y) &= \int_u f(Y,u)du = \int_u f(Y|u)f(u)du \\ &= \int_{-\infty}^{\infty} e^{-\mu}\frac{\mu^y}{y!}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{u^2}{2\sigma^2}}du \\ &= \int_{-\infty}^{\infty} e^{-e^{(\mathbf{x}'\boldsymbol{\beta}+u)}}\frac{(e^{\mathbf{x}'\boldsymbol{\beta}+u})^y}{y!}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{u^2}{2\sigma^2}}du \\ &= \frac{e^{\mathbf{x}'\boldsymbol{\beta}y}}{y!\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-e^{(\mathbf{x}'\boldsymbol{\beta}+u)}}e^{yu}e^{-\frac{u^2}{2\sigma^2}}du. \end{aligned} \tag{4.1}$$

Vamos nos concentrar somente na integral. Podemos escrever:

$$e^{-e^{(\mathbf{x}'\boldsymbol{\beta}+u)}} = \sum_{n=0}^{\infty}\frac{(-1)^n e^{n(\mathbf{x}'\boldsymbol{\beta}+u)}}{n!}. \tag{4.2}$$

Substituindo (4.2) na integral (4.1) temos que

$$\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\int_{-\infty}^{\infty} e^{n\mathbf{x}'\boldsymbol{\beta}+(n+y)u-\frac{u^2}{2\sigma^2}}du$$

Completando com os termos que faltam para o quadrado e rearranjando a equação temos que

$$f(Y) = \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}e^{n\mathbf{x}'\boldsymbol{\beta}+(n+y)^2\frac{\sigma^2}{2}}\int_{-\infty}^{\infty} e^{-(\frac{u}{\sqrt{2}\sigma}-\frac{(n+y)\sigma}{\sqrt{2}})^2}du. \tag{4.3}$$

Podemos fazer uma mudança de variável na integral assumindo

$$v = \frac{u}{\sqrt{2}\sigma} - \frac{(n+y)\sigma}{\sqrt{2}} \quad \text{e} \quad dv = \frac{du}{\sqrt{2}\sigma},$$

de tal forma que a equação (4.3) fica

$$f(Y) = \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}e^{n\mathbf{x}'\boldsymbol{\beta}+(n+y)^2\frac{\sigma^2}{2}}\int_{-\infty}^{\infty} e^{-v^2}\sqrt{2}\sigma dv. \tag{4.4}$$

Sabendo que

$$\int_{-\infty}^{\infty} e^{-v^2}\sqrt{2}\sigma dv = \sigma\sqrt{2\pi}.$$

temos que

$$f(Y) = \sigma\sqrt{2\pi}e^{\frac{y^2\sigma^2}{2}} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} e^{n(\mathbf{x}'\boldsymbol{\beta}+y\sigma^2+\frac{n\sigma^2}{2})}$$

$$= \sigma\sqrt{2\pi}e^{\frac{y^2\sigma^2}{2}} \sum_{n=0}^{\infty} c_n z^{m_n}, \tag{4.5}$$

em que $c_n = \frac{(-1)^n}{n!}e^{n(\mathbf{x}'\boldsymbol{\beta}+y\sigma^2)} \cdot 3^{\frac{(n\sigma)^2}{2}}$, $z = e/3$ e $m_n = (n\sigma)^2/2$.

A somatória em (4.5) não pode ser expressa por um número finito de operações elementares $(+, -, *, /, \widehat{\ }, \dots)$ envolvendo somente funções elementares $(\log, \cos, \dots)$, pois segundo Weiss e Weiss (1963) vale o seguinte resultado:

Se $\sum_{n=0}^{\infty} c_n z^{m_n}$, $|z| < 1$, $m_{n+1}/m_n > cte > 0$ satisfazendo $\sum_{n=0}^{\infty} |c_n| = \infty$, então a série assume cada valor complexo w para infinitos pontos z no disco unitário.

Tal propriedade não se verifica para uma expressão finita envolvendo somente operações e funções elementares. Neste caso (4.1) é calculada numericamente pelo MatLab.

O uso de efeitos aleatórios também leva em consideração a correlação entre as observações que tenham algum efeito aleatório em comum. Assim, Y_{it} e Y_{is} são independentes dado u_i, mas são marginalmente correlacionadas. Para $t \neq s$,

$$\begin{aligned} Cov(Y_{it}, Y_{is}) &= E[Cov(Y_{it}, Y_{is}|u_i)] + Cov[E(Y_{it}|u_i), E(Y_{is}|u_i)] \\ &= 0 + Cov[e^{\mathbf{x}'_{it}\boldsymbol{\beta}+u_i}, e^{\mathbf{x}'_{is}\boldsymbol{\beta}+u_i}]. \end{aligned}$$

Como os termos da covariância são ambos funções monótonas crescentes de u_i, temos que Y_{it} e Y_{is} são correlacionados não-negativamente.

O capítulo 10 de McCullogh e Searle (2001) traz uma discussão bastante abrangente dos métodos possíveis para a estimação de modelos generalizados mistos. Uma vez que estamos diante de modelos não lineares, várias dificuldades de cálculo aparecem. Atualmente o programa SAS já possui uma rotina para modelos mistos não lineares (NLMIXED), que engloba os modelos mistos generalizados. Porém, tal rotina tem melhor desempenho quando limitada a um único efeito aleatório com distribuição normal.

Capítulo 5

Aplicações de Regressão de Poisson

5.1 Dados de câncer de pulmão

Analisaremos os dados de Hirayama (1981) sobre os efeitos do fumo passivo na incidência de câncer de pulmão para mulheres japonesas não fumantes. Em um estudo prospectivo em 29 centros de saúde do Japão, um dos objetivos do estudo era verificar o efeito do fumo passivo na incidência de câncer de pulmão. Durante 14 anos (1966-1979), entre 91.540 mulheres casadas, não-fumantes e que tinham 40 anos ou menos no início do estudo, ocorreram 174 mortes por câncer de pulmão. Além do status do marido quanto ao fumo, temos outras variáveis como o tipo de trabalho do marido (se na agricultura ou não) e a faixa de idade do marido (de 40 a 59 anos ou com 60 anos ou mais). Na Tabela 5.1 temos o número de casos para cada sub-população, assim como o número de pessoas expostas.

Neste caso específico, para $i = 1, ..., 12$, temos que Y_i representa o número de casos de câncer e n_i o número de pessoas expostas em cada grupo. A incidência de câncer é dada por λ_i, escreve-se o modelo:

$$Y_i \sim Poisson(n_i\lambda_i),$$

$$\log\left(\frac{E(Y_i)}{n_i}\right) = x_i'\boldsymbol{\beta} \quad \text{ou} \quad \log(E(Y_i)) = x_i'\boldsymbol{\beta} + \log(n_i).$$

A quantidade $\log(n_i)$ é chamada de *offset*. As variáveis explicativas são a ocupação do marido (*ocup*), o status do marido quanto ao fumo ($fumo$) e o grupo de idade (*idade*). Os efeitos das diferentes variáveis foram codificados conforme a Tabela 5.2. Um modelo com efeitos principais e todas as possíveis interações (modelo saturado) foi ajustado. Os resultados mostraram que as interações não são significativas. O melhor modelo foi aquele ajustado somente com as três variáveis principais. Na Tabela 5.3 temos os valores para a *deviance* e o critério AIC dos vários modelos de efeitos principais possíveis.

Tabela 5.1: Estudo sobre efeito de fumo passivo em câncer de pulmão.

ocupação	idade	status quanto ao fumo	casos	pessoas expostas	mortes por 1000
agricultura	40-59 anos	não fumante	3	5.999	0,5
		ex ou 1-19 cig/dia	20	12.753	1,6
		mais de 20 cig/dia	16	7.150	2,2
	mais de 60 anos	não fumante	14	4.407	3,2
		ex ou 1-19 cig/dia	32	7.291	4,4
		mais de 20 cig/dia	8	2.241	3,6
outros	40-59 anos	não fumante	8	8.021	1,0
		ex ou 1-19 cig/dia	20	17.923	1,1
		mais de 20 cig/dia	20	13.434	1,5
	mais de 60 anos	não fumante	7	3.468	2,0
		ex ou 1-19 cig/dia	14	6.217	2,3
		mais de 20 cig/dia	12	2.636	4,6

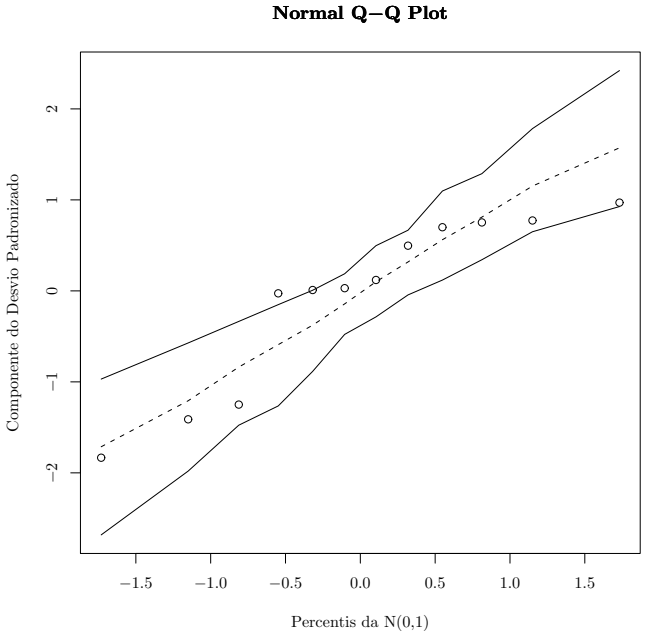

Figura 5.1: Gráfico do envelope para dados de câncer de pulmão utilizando a regressão de Poisson.

Tabela 5.2: Codificação dos Efeitos.

Variável	Fatores	Codificação
ocup	agricultura	0
	outros	1
idade	40-59 anos	0
	mais de 60 anos	1
fumo	não fumante	0
	ex ou 1-19 cig/dia	1
	mais de 20 cig/dia	1

Tabela 5.3: Comparação dos possíveis modelos, segundo o critério de *deviance* e AIC.

Modelo	*deviance*	Critério AIC
$ocup$	$45,8$	102,2
$fumo$	$49,1$	107,5
$idade$	$17,8$	74,2
$ocup, fumo$	$41,4$	101,9
$ocup, idade$	$14,2$	72,7
$fumo, idade$	$10,3$	70,8
$ocup, fumo, idade$	$6,1$	68,5

Tabela 5.4: Estimativa dos parâmetros, respectivos erros padrão e p-valor.

Parâmetro	Valor	erro padrão	p-valor
intercepto	$-6,8263$	0,2130	<0,0001
ocupação1	$-0,3166$	0,1538	0,0395
fumo1	0,3492	0,2073	0,0920
fumo2	0,6285	0,2245	0,0051
idade1	0,9364	0,1549	<0,0001

Os modelos que melhor se ajustam aos dados são aqueles com a menor *deviance* e também o menor critério AIC, no caso o modelo com todos os efeitos principais. Agora, uma pergunta que pode surgir é se a redução da *deviance* devido à introdução de uma nova variável é significativa ou não. Por exemplo, no modelo ajustado com as variáveis $fumo$ e $idade$ temos uma *deviance* de $10,3395$. Será que ao introduzirmos a variável *ocup*, a redução na *deviance* é significativa? Note que

$$D(ocup|fumo, idade) = D(fumo, idade) - D(ocup, fumo, idade) =$$

$$= 10,3395 - 6,0882 = 4,2513.$$

O valor de $4,2513$ é significativo ao nível de $0,0392$, de acordo com uma distribuição $\chi^2_{(1)}$. Portanto devemos incluir a variável *ocup* no modelo. Para o modelo com menor da *deviance*, a Tabela 5.4 apresenta as estimativas dos parâmetros, os erros padrão e os p-valores para os testes $H_0 : \beta_i = 0, i = 0, ..., 4$. A interpretação dos parâmetros é dada a seguir, em que

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{\beta}}_3 \\ \hat{\boldsymbol{\beta}}_4 \end{pmatrix} = \begin{cases} \beta_0 = & \log(\lambda_i) \text{ para casela de referência,} \\ \beta_1 = & \text{incremento em } \log(\lambda_i) \text{ devido às outras ocupações,} \\ \beta_2 = & \text{incremento em } \log(\lambda_i) \text{ devido ao efeito de fumo 1,} \\ \beta_3 = & \text{incremento em } \log(\lambda_i) \text{ devido ao efeito de fumo 2,} \\ \beta_4 = & \text{incremento em } \log(\lambda_i) \text{ devido ao efeito de idade.} \end{cases}$$

Como podemos verificar pela interpretação dos parâmetros, β_0, a casela de referência, refere-se ao logaritmo da quantidade das mulheres cujo marido trabalha na agricultura, está na faixa de idade de 40 a 59 anos e não é fumante.

Uma vez que não estamos diante de superdispersão, pois

$$deviance/g.l. = 6,0882/7 = 0,8697,$$

parece razoável aceitar uma regressão de Poisson para estes dados. Na Figura 5.1 temos o gráfico do envelope que indica que o modelo está bem ajustado, apesar de haver um ponto fora e outro no limite da banda de confiança.

Tabela 5.5: Dados de leucemia para diferentes doses de radiação.

número de casos	pessoas expostas	dose de radiação
13	26523	0
45	55134	5
19	14407	30
7	3896	75
13	2906	150
42	2770	300

5.2 Dados de morte por leucemia

Para ilustrar o aparecimento da superdispersão, usaremos dados de morte por leucemia por diferentes doses de radiação. Também neste caso estamos lidando com incidência, e o modelo tem um $offset$ com o logaritmo do número de pessoas expostas. Diferente do modelo anterior, a variável explicativa não é categórica, mas refere-se às diferentes doses de radiação. Estamos diante de seis grupos diferentes de pessoas expostas em que foram anotados os casos de leucemia conforme mostra a Tabela 5.5. Para $i = 1, ..., 6$ temos que Y_i é o número de casos de leucemia e n_i é o número de pessoas expostas em cada grupo. Como a incidência de leucemia é λ_i, escreve-se o modelo:

$$Y_i \sim Poisson(n_i\lambda_i),$$

$$\log\left(\frac{E(Y_i)}{n_i}\right) = x_i'\boldsymbol{\beta} \quad \text{ou} \quad \log(E(Y_i)) = x_i'\boldsymbol{\beta} + \log(n_i).$$

O modelo de Poisson apresenta superdispersão, pois o valor da *deviance* dividido pelos graus de liberdade é

$$4921,9/4 = 1230,5,$$

que é muito maior que 1. Nesse caso, os seis pontos do conjunto de dados estão completamente fora da banda de confiança. Já para o modelo com Binomial Negativa, notamos uma melhora sensível: somente um ponto se encontra ligeiramente fora do envelope. A Figura 5.2 nos mostra claramente como o modelo de Poisson com superdispersão apresenta resíduos muito grandes, e portanto fora da banda de confiança. Quando aplicamos a solução da Binomial Negativa percebemos um ajuste bem melhor para os resíduos.

Estimamos também dois modelos de quase-verossimilhança, o primeiro com $V(\mu) = \mu^2$ e o outro com $V(\mu) = \mu^3$. Estes também não apresentaram uma solução melhor pois a Tabela 5.6 compara os valores observados com os valores preditos e também nos mostra o fraco poder de todos os modelos para este pequeno conjunto de dados.

Tabela 5.6: Valores preditos pelos diferentes modelos.

Valor obs.	Poisson	Bin.Negativa	Quase-vero μ^2	Quase-vero μ^3
13	20.0	20.6	20.6	20.3
45	43.8	45.0	45.0	44.7
19	14.8	15.3	15.3	15.5
7	6.3	6.6	6.6	6.9
13	10.1	10.8	10.8	12.0
42	44.1	49.0	49.0	61.9

Este exemplo é somente ilustrativo, uma vez que estamos diante de um conjunto de dados muito pequeno. Como todos os testes de hipóteses são baseados em propriedades assintóticas dos estimadores de máxima verossimilhaça, é sempre mais adequado trabalhar com amostras maiores. A Tabela 5.6 compara os valores observados com os valores preditos.

Para o modelo de Poisson o critério AIC é de 4995,9 e para o modelo Binomial Negativa, é 123,7. Não temos esta informação para os modelos de quase-verossimilhança uma vez que nestes modelos não temos uma função de distribuição associada como foi exposto no capítulo 3. Neste caso, o modelo Binomial Negativa é bem melhor do que o modelo Poisson.

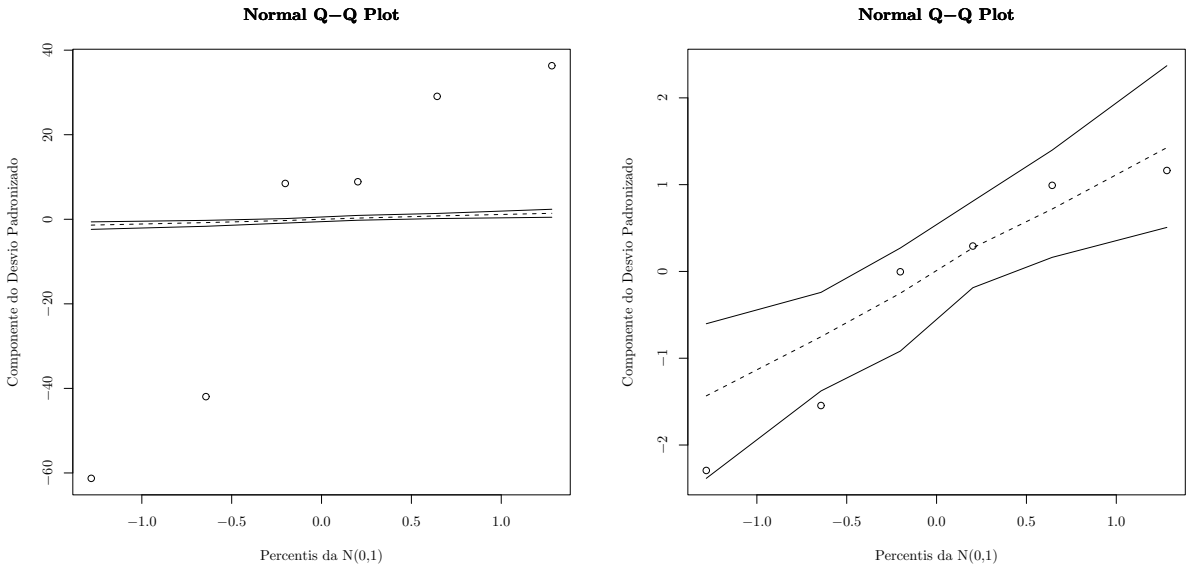

Figura 5.2: Gráfico envelope para regressão de Poisson e Binomial Negativa.

Capítulo 6

Aplicações a dados reais

Em um estudo referente à alimentação de rãs da espécie *Adenomera*, foi anotado para cada um dos 102 indivíduos capturados, o conteúdo do estômago. Este poderia incluir contagens de 9 tipos de comidas diferentes. Entre as nove categorias de comida possíveis temos: *Blattodea*, *Isoptera Alates*, *Isoptera non-Alates*, *Heteroptera*, *Hemiptera*, *Larvae*, *Coleoptera*, *Hymenoptera-Formicidae* e *Aranae*.

As rãs foram coletadas entre outubro de 1999 e março de 2001 pelo Museu de Biodiversidade do Cerrado da Universidade Federal de Uberlândia, Minas Gerais. Os dados foram coletados para ambos os sexos e na estação seca (meses de abril a agosto) e úmida (meses de setembro a março). Os indivíduos foram classificados por sexo e segundo a estação em que foram observados, como mostra a Tabela 6.1. Como variáveis explicativas temos a estação do ano em que o animal foi capturado (seca ou úmida) e o sexo de cada animal. Podemos verificar tanto os efeitos principais como a interação entre estes. Os efeitos foram codificados de acordo com a Tabela 6.2.

Vamos ilustrar para alguns indivíduos como é a apresentação dos dados. Cada rã foi identificada com um número. No caso da Tabela 6.3 temos para efeito de ilustração somente uma parte dos dados coletados. Ilustramos assim para 8 fêmeas que foram coletadas na estação úmida como os dados se apresentam. Na primeira coluna temos o número identificador de cada indivíduo e nas colunas subseqüentes o quanto foi ingerido

Tabela 6.1: *Adenomera* - Distribuição dos indivíduos por estação e sexo.

	Machos	Fêmeas	Total
Seca	10	4	14
Úmida	49	39	88
Total	59	43	102

Tabela 6.2: Codificação dos Efeitos do sexo e da estação.

Variável		valor
Sexo	Fêmea	0
	Macho	1
Estação	Seca	0
	Úmida	1

Tabela 6.3: *Adenomera* - Quantidade ingerida por cada indivíduo.

Ind	Blatt	Isopal	Isonon	Hete	Hemi	Larv	Cole	Hyme	Aran
1341	0	0	0	0	0	2	2	0	0
1300	0	0	4	0	1	0	0	0	0
1298	1	0	0	2	0	0	1	3	1
1291	1	0	0	0	0	0	0	0	0
1310	0	0	0	0	0	0	0	1	1
1309	0	0	0	0	1	0	0	4	1
1301	0	0	0	0	0	0	0	0	2
1288	0	4	0	0	0	0	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

de cada tipo de comida. Por exemplo, a fêmea identificada com o número 1341 ingeriu 2 unidades da comida *Larvae* e duas unidades da comida *Coleoptera*.

A particularidade destes dados, é que para cada animal estudado, podia-se encontrar tanto poucas categorias de comidas representadas quanto poucas unidades de cada tipo de comida ingerida. Os dados, da forma como são apresentados, sugerem um modelo multimonial para cada indivíduo. Uma vez que estes são independentes, cada linha seria um modelo multinomial com o vetor $\mathbf{p} = (p_1, p_2, ..., p_9)$ representando as probabilidades da ingestão de cada comida.

Agresti (2002) apresenta um exemplo de alimentação de jacarés em que ele testa modelos logito nominal para respostas multinomiais. Para cada par de categoria de resposta ele apresenta a razão entre as proporcções ingeridas de comida, sempre comparando com uma categoria fixa. A diferença entre este exemplo e os nossos dados é que para os 219 jacarés que compõem a amostra foi anotado somente a escolha principal de alimento. Assim, um mesmo animal poderia ter comido peixes, invertebrados e répteis, mas somente o volume maior de comida ingerida era anotado. Para o conjunto das rãs, não foi anotado somente o alimento principal. Assim, tal modelo apresentava muitas dificuldades de estimação

devido à peculiaridade dos dados. Além de existirem 9 categorias diversas, os dados são esparsos e apresentam muitos zeros. Podemos postular que o fato de um indivíduo não ter ingerido determinada comida se deu por que ele não a encontrou no ambiente ("zero amostral"), ou, ele encontrou, mas preferiu não comer ("zero estrutural").

Optamos então por analisar cada categoria de comida separadamente. Para cada tipo de comida estamos interessados em verificar se há variação no hábito alimentar de acordo com sexo e estação. Desta forma teremos a oportunidade de testar os diversos modelos lineares existentes para dados de contagem, inclusive modelos mais complexos como os modelos mistos e os modelos com excesso de zeros.

Para cada comida analisamos os modelos possíveis partindo do mais simples, que é a regressão de Poisson. Em seguida verificaremos se o modelo apresenta ou não superdispersão, ou seja, se a *deviance* dividida pelos graus de liberdade é maior que 1. Na presença da superdispersão podemos buscar a alternativa do modelo Binomial Negativa ou modelo misto, com efeito aleatório. Outra abordagem é verificar se para a comida em questão encontramos um modelo ZIP com probabilidade extra de zero (p) significativa.

Assim, definimos as seguintes variáveis explicativas X_1, X_2 e X_3:

$$\begin{aligned} X_1 &= \begin{cases} 1 = & \text{se macho,} \\ 0 = & \text{se fêmea.} \end{cases} \\ X_2 &= \begin{cases} 1 = & \text{se estação úmida,} \\ 0 = & \text{se estação seca.} \end{cases} \\ X_3 &= \begin{cases} 1 = & \text{se estação úmida e macho,} \\ 0 = & \text{caso contrário.} \end{cases} \end{aligned} \tag{6.1}$$

O primeiro modelo testado é a regressão de Poisson. No caso, Y_{it} é a quantidade da comida que foi ingerida por cada indivíduo i que pertencente à subpopulação ou ao grupo t. Temos que $Y_{it} \sim Poisson(\lambda_{it})$, $i = 1, ..., n$ e t depende de quantos grupos diferentes tivermos em um modelo.

Por exemplo, em um modelo em que os efeitos do sexo e da estação sejam significativos teremos t=4, pois teremos os grupos de fêmeas na estação seca, fêmeas na estação úmida, machos na estação seca e machos na estação úmida. Para cada um destes grupos teremos n_t indivíduos na amostra de tal forma que $\sum_t n_t = n$.

A regressão de Poisson fica expressa por

$$\log(\lambda_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta},$$

em que $\mathbf{x}'_{it}$ é a i-ésima linha da matriz $\mathbf{X}$ de efeitos fixos conhecidos. Teremos um $\hat{\lambda}_t$ para cada grupo:

$$\hat{\lambda}_t = e^{\mathbf{x}'_t\hat{\boldsymbol{\beta}}}.$$

A primeira opção de modelo a ser testado é o modelo saturado:

$$\log(\lambda_{it}) = \beta_0 + X_{1_{it}}\beta_1 + X_{2_{it}}\beta_2 + X_{3_{it}}\beta_3. \tag{6.2}$$

O vetor de parâmetros estimados neste modelo é:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{\beta}}_3 \end{pmatrix} \begin{cases} \beta_0 = & \text{casela de referência,} \\ \beta_1 = & \text{efeito do sexo,} \\ \beta_2 = & \text{efeito da estação} \\ \beta_3 = & \text{interação sexo e estação.} \end{cases}$$

Se tivermos todos efeitos significativos teremos a seguinte interpretação para as quatro sub-populações possíveis:

β_0 representa o logaritmo da quantidade de comida ingerida pelo animal do sexo feminino na estação seca;
β_1 representa o efeito do sexo masculino;
β_2 representa o efeito da estação úmida;
β_3 representa o efeito conjunto do sexo masculino e da estação úmida.

Para encontrarmos as médias ingeridas por cada sub-população temos:

e^{β_0} = média de comida ingerida pelo animal do sexo feminino na estação seca;
$e^{\beta_0+\beta_2}$ = média de comida ingerida pelo animal do sexo feminino na estação úmida;
$e^{\beta_0+\beta_1}$ = média de comida ingerida pelo animal do sexo masculino na estação seca;
$e^{\beta_0+\beta_1+\beta_2+\beta_3}$ = média de comida ingerida pelo animal do sexo masculino na estação úmida.

Outro modelo a ser testado é o modelo de efeitos principais:

$$\log(\lambda_{it}) = \beta_0 + X_{1_{it}}\beta_1 + X_{2_{it}}\beta_2. \tag{6.3}$$

O vetor de parâmetros estimados neste modelo é:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} \begin{cases} \beta_0 = & \text{casela de referência,} \\ \beta_1 = & \text{efeito do sexo,} \\ \beta_2 = & \text{efeito da estação.} \end{cases}$$

Se tivermos todos efeitos significativos teremos a seguinte interpretação para as quatro sub-populações possíveis:

β_0 representa o logaritmo da quantidade de comida ingerida pelo animal do sexo feminino na estação seca;
β_1 representa o efeito do sexo masculino;

β_2 representa o efeito da estação úmida.

Para encontrarmos as médias ingeridas por cada sub-população temos:

e^{β_0} = média de comida ingerida pelo animal do sexo feminino na estação seca;
$e^{\beta_0+\beta_2}$ = média de comida ingerida pelo animal do sexo feminino na estação úmida;
$e^{\beta_0+\beta_1}$ = média de comida ingerida pelo animal do sexo masculino na estação seca;
$e^{\beta_0+\beta_1+\beta_2}$ = média de comida ingerida pelo animal do sexo masculino na estação úmida.

Depois será testado o modelo somente com efeito do sexo:

$$\log(\lambda_{it}) = \beta_0 + X_{1_{it}}\beta_1. \tag{6.4}$$

O vetor de parâmetros estimados neste modelo é:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} \begin{cases} \beta_0 = & \text{casela de referência,} \\ \beta_1 = & \text{efeito do sexo.} \end{cases}$$

Agora temos somente duas sub-populações e a interpretação dada aos parâmetros é a seguinte:

β_0 representa o logaritmo da quantidade de comida ingerida pelo animal do sexo feminino;
β_1 representa o efeito do sexo masculino.

Agora não temos mais o efeito da estação, então para encontrarmos as médias ingeridas por cada sub-população temos:

e^{β_0} = média de comida ingerida pelo animal do sexo feminino;
$e^{\beta_0+\beta_1}$ = média de comida ingerida pelo animal do sexo masculino.

Outro modelo a ser testado é o modelo somente com a estação:

$$\log(\lambda_{it}) = \beta_0 + X_{2_{it}}\beta_2. \tag{6.5}$$

O vetor de parâmetros estimados neste modelo é:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} \begin{cases} \beta_0 = & \text{casela de referência,} \\ \beta_2 = & \text{efeito da estação.} \end{cases}$$

A interpretação dada aos parâmetros é a seguinte:

β_0 representa o logaritmo da quantidade de comida ingerida pelo animal na estação seca;
β_2 representa o efeito da estação úmida.

Para encontrarmos as médias ingeridas por cada sub-população temos:

e^{β_0} = média de comida ingerida pelo animal na estação seca;
$e^{\beta_0+\beta_2}$ = média de comida ingerida pelo animal na estação úmida.

Finalmente estimaremos o efeito somento com o β_0, que neste caso significa uma média geral, sem distinção de sub-populações. A quantidade média de comida ingerida vai ser dada por e^{β_0}. Este modelo será chamado de modelo com a média geral e será dado por:

$$\log(\lambda_{it}) = \beta_0. \tag{6.6}$$

Para encontrar as quantidades preditas pelos modelos Poisson iremos calcular as probabilidades para cada quantidade observada na comida em questão e multiplicar pelo número de animais no grupo. Por exemplo, a previsão do número de indivíduos no grupo t que não ingeriram nenhum item da comida em estudo é:

$$\widehat{P(Y=0)}n_t = e^{-\hat{\lambda}_t}n_t.$$

O valor predito para o modelo dos que ingeriram y itens no grupo t, para $y = 1, 2, ...$, é:

$$\widehat{P(Y=y)}n_t = e^{-\hat{\lambda}_t}\frac{\hat{\lambda}_t^y}{y!}n_t.$$

Outro modelo a ser testado, principalmente no caso da regressão de Poisson apresentar superdispersão, é o modelo Binomial Negativa. Também estimaremos os modelos saturado, de efeitos principais, efeito do sexo, efeito da estação e média geral. Temos agora que $Y|Z \sim Poisson(Z)$ e $Z \sim Gama(\nu, \lambda/\nu)$ como mostrado no capítulo 3. A distribuição marginal de Y é Binomial Negativa com média λ e parâmetro de dispersão ν. O modelo é também

$$\log(\lambda_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta},$$

e temos para cada grupo t o valor estimado de λ,

$$\hat{\lambda}_t = e^{\mathbf{x}'_t\hat{\boldsymbol{\beta}}}.$$

Para os modelos Binomial Negativa também valem os modelos (6.2) a (6.6) para a média ingerida, a diferença deste modelo é que temos uma dispersão estimada, ou seja, além do $\hat{\lambda}$ teremos um valor $\hat{\nu}$, de tal forma que os valores preditos pelo modelo para o grupo t ficam:

$$\widehat{P(Y=y)}n_t = \frac{(y+\hat{\nu}-1)!}{y!(\hat{\nu}-1)!}\left(\frac{\frac{\hat{\nu}}{\hat{\lambda}}}{\frac{\hat{\nu}}{\hat{\lambda}}+1}\right)^{\hat{\nu}}\left(\frac{1}{\frac{\hat{\nu}}{\hat{\lambda}}+1}\right)^{y}n_t.$$

O modelo misto ou de efeitos aleatórios foi discutido no capítulo 4. Temos $Y_{it}|u_i \sim Poisson(\lambda_{it})$ e $u_i \sim Normal(0, \sigma^2)$ e

$$\log[E[Y_{it}|u_i]] = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i \quad \text{com}$$

$$\hat{\lambda}_t = e^{\mathbf{x}'_t\hat{\boldsymbol{\beta}} + \frac{\hat{\sigma}^2}{2}}. \tag{6.7}$$

No caso dos modelos estimados, para calcular as médias para cada sub-população temos que considerar a equação (6.7). Neste caso o $\hat{\sigma}$ vai ser considerado na média, mas a interpretação dos parâmetros $\boldsymbol{\beta}$ permanece a mesma dada para os modelos Poisson e Binomial Negativa.

Os valores preditos são obtidos pela distribuição marginal de Y:

$$\widehat{P(Y=y)}n_t = \left(\frac{e^{\mathbf{x}'_t\hat{\boldsymbol{\beta}}y}}{y!\sqrt{2\pi}\hat{\sigma}}\int_{-\infty}^{\infty} e^{-e^{(\mathbf{x}'_t\hat{\boldsymbol{\beta}}+u)}}e^{yu}e^{-\frac{u^2}{2\hat{\sigma}^2}}du\right)n_t.$$

Como foi mostrado no capítulo 4, a densidade explícita não pode ser escrita, mas podemos calcular estas probabilidades no MatLab.

Finalmente podemos testar se existe um número excessivo de zeros na amostra. Estimamos o modelo ZIP e verificamos se a estimativa de p é significativa. Como foi mostrado no capítulo 3, p é a probabilidade extra de zeros que não é levada em conta pela distribuição Poisson.

O modelo testado é da forma

$$\log(\lambda_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta},$$

com $Y_{it} \sim ZIP(p, \lambda_{it})$, neste caso, além de estimar λ_{it}, estimamos p. Vale lembrar que Lambert (1982) mostra outros modelos para estimar p, como mostrado no Capítulo 3 em (3.6) e (3.7). No nosso caso estimaremos p separadamente como mostrado em (3.8). Uma vez que temos em nossos dados poucas varíaveis explicativas, estes modelos se equivalem. Um exemplo ilustrativo de modelo ZIP em que o p é estimado com logito pode ser visto em Slymen (2006).

Assim como nos modelos de efeitos aleatórios, para calcular as médias consumidas teremos uma pequena modificação em relação às médias calculadas nos modelos Poisson e Binomial Negativa. Temos que utilizar o fato que estamos lidando com uma distribuição que é a mistura de uma Poisson com a probabilidade extra de zeros, e como foi mostrado no capítulo 3, a média consumida por cada população fica dada por $(1-p)\lambda_t$. Para calcularmos as médias consumidas usamos $(1-\hat{p})\hat{\lambda}_t$ em que $\hat{\lambda}_t = e^{\mathbf{x}'_t\hat{\boldsymbol{\beta}}}$.

Uma vez que temos os valores para $\hat{\lambda}_t$ e $\hat{p}$, calculamos os valores preditos pelo modelo:

$$\begin{aligned} \widehat{P(Y=0)}n_t &= (\hat{p}+(1-\hat{p})e^{-\hat{\lambda}_t})n_t \quad \text{e} \\ \widehat{P(Y=y)}n_t &= ((1-\hat{p})e^{-\hat{\lambda}}\frac{\hat{\lambda}^y}{y!})n_t, \quad \text{se} \quad y>0. \end{aligned}$$

Todos os modelos foram estimados utilizando-se o programa SAS. Os modelos Poisson e Binomial Negativa foram estimados pelo *procedure* GENMOD. Para os modelos com efeitos aleatórios e ZIP utilizamos o *procedure* NLMIXED. Para comparar os modelos usamos o critério AIC, procurando o modelo com o menor valor, ou seja,

$$AIC = -2\ell(\mu, y) + 2k,$$

em que $\ell(\mu, y)$ é o logaritmo da verossimilhança do modelo em questão e k é o número de parâmetros estimados. O programa SAS somente fornece os valores do critério AIC para o *procedure* NLMIXED, então, os modelos estimados pelo *procedure* GENMOD são também estimados usando o programa R.

Na literatura, o critério AIC é utilizado para comparar modelos encaixados, ou seja, para uma configuração de modelo, por exemplo, na regressão de Poisson, escolhemos aquele com o menor valor do critério AIC entre (6.2) a (6.6). Em nossa aplicação estamos comparando quatro possibilidades diferentes: regressão de Poisson, Binomial Negativa, Modelo de Efeitos Aleatórios e Modelo ZIP com p homogêneo. Desta forma, procuraremos o menor valor de AIC entre as configurações possíveis, uma vez que as funções de verossimilhança são parecidas.

Embora McCullagh e Nelder (1989) postulem que nem sempre é possível estabelecer o "melhor" modelo entre os analisados, escolhemos o "melhor" dentre eles para cada tipo de comida, a título de ilustração. Por exemplo, Agresti (2002) compara modelos diferentes para o mesmo conjunto de dados e comenta a pertinência de preferir aquele que apresenta o menor valor da *deviance* entre eles.

Também faremos a inspeção de ajuste do modelo comparando valores observados com valores estimados. A título de ilustração, para os modelos Poisson e Binomial Negativa, construiremos os gráficos de envelope mencionados no capítulo 2. Estes gráficos são construídos no programa R, conforme Paula (2004). Com este programa construiremos também os gráficos de diagnóstico para o afastamento da verossimilhança e o gráfico de resíduos de *deviance* contra os valores lineares preditos.

6.1 Modelos para alimentação da espécie *Adenomera*

Nosso interesse é explicar o consumo de cada tipo de comida pelas rãs da espécie *Adenomera* de acordo com sexo e estação.

Nas próximas subseções apresentaremos para cada tipo de comida ingerida os modelos de contagem discutidos anteriormente. Ou seja, para cada categoria de comida estimaremos o modelo de Poisson, Binomial Negativa, Efeitos Aleatórios e ZIP. E para cada modelo em questão estimaremos o modelo saturado, o de efeitos principais, o de efeito somente do sexo, o de efeito da estação e o de média geral.

Em seguida compararemos os modelos em termos de estimativa dos parâmetros, valores observados e preditos, valor de AIC e gráficos de diagnóstico, para buscar para cada tipo de comida um modelo que apresente o melhor ajuste. No Apêndice A apresentamos todos os valores do critério AIC para os modelos estimados em cada subseção. No Apêndice B apresentamos gráficos de diagnóstico para alguns dos modelos estimados para ilustrar como estes gráficos podem nos orientar na busca de modelos mais adequados e na identificação de observações influentes.

6.1.1 Modelos para *Blattodea*

Em termos gerais, para esta comida não há muita diferença no valor do critério AIC dos diferentes modelos, como atesta a Tabela A.1. O menor valor é para a regressão de Poisson somente com o sexo, em que este efeito é significativo a 9,44%.

Não há evidência de superdispersão no modelo Poisson, pois no modelo saturado

$$deviance/g.l. = 93,5077/98 = 0,9542.$$

E no modelo somente com o efeito do sexo,

$$deviance/g.l. = 97,0024/98 = 0,9700.$$

Na Figura 6.1 temos os gráficos de envelope para os modelos saturado e somente com o efeito de sexo. Por este gráfico não notamos muita diferença entre os modelos, eles parecem satisfatórios. Porém quando comparamos os valores observados com os preditos para os dois modelos, encontramos maiores discrepâncias dos valores preditos no modelo saturado. Elegemos como melhor modelo a regressão de Poisson somente com o efeito do sexo.

Modelo Poisson para *Blattodea*:

$$\log(\hat{\lambda}_{it}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1_{it}}.$$

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -1,1962 + 0,5527 x_{1_{it}}.$$

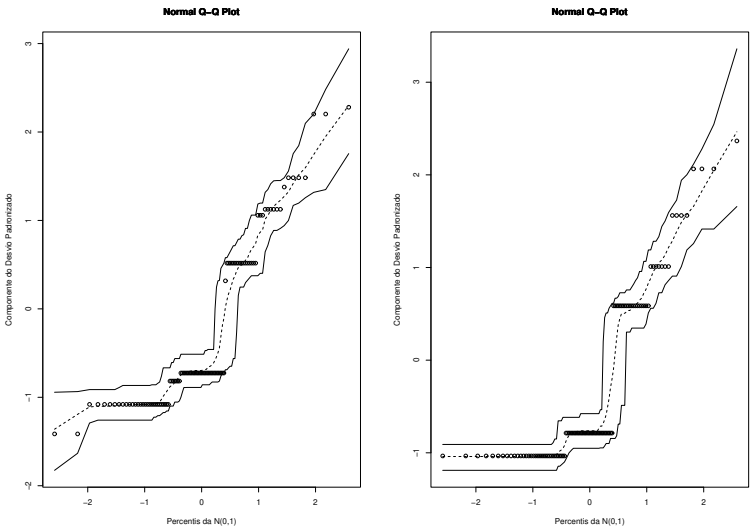

Figura 6.1: *Blattodea* - Gráfico envelope para comparação entre os modelos (a) Poisson - saturado; (b) Poisson - sexo.

Tabela 6.4: *Blattodea* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

parâmetro	estimativa	erro padrão	χ^2	p-valor
$\hat{\beta}_0$	$-1,1962$	0,2773	18,6	$<$0,0001
$\hat{\beta}_1$	0,5527	0,3304	2,80	0,0944

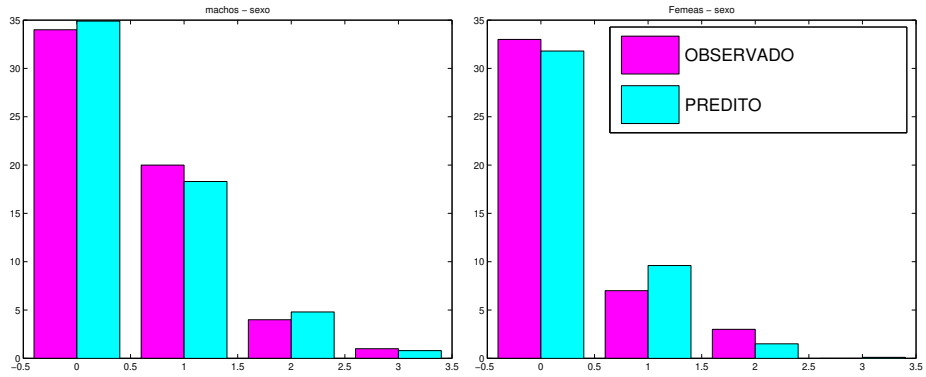

Figura 6.2: *Blattodea* - Comparação de valores observados e preditos - Regressão de Poisson com efeito do sexo (a) machos; (b) fêmeas.

Para cada sub-população do modelo temos os valores preditos das quantidades médias ingeridas pelas fêmeas e machos:

$$\text{Fêmeas: } \hat{\lambda}_f = e^{\hat{\beta}_0} = 0,3023.$$
$$\text{Machos: } \hat{\lambda}_m = e^{\hat{\beta}_0+\hat{\beta}_1} = 0,5255.$$

Razão de médias:

Comparando fêmeas (f) e machos (m):

$$\frac{\hat{\lambda}_f}{\hat{\lambda}_m} = \frac{e^{\hat{\beta}_0}}{e^{\hat{\beta}_0+\hat{\beta}_1}} = \frac{1}{e^{\hat{\beta}_1}} = \frac{1}{e^{0,5527}} = 0,5754.$$

Como os machos comem em média mais do que as fêmeas, consideraremos o inverso da razão de médias calculado acima. Temos então que os machos comem 1,7379 vezes mais *Blattodea* do que as fêmeas. $\text{IC}_{90\%} : [0,7933; 2,6826]$.

O efeito do sexo é significativo a 9%, desta forma, calculamos os intervalos de confiança com 90% de nível de confiança. Os intervalos de confiança são construídos a partir do método Delta descrito na seção 2.4.

6.1.2 Modelos para *Isoptera Alates*

Para esta comida não houve consumo na estação seca, por isso os modelos foram testados somente na estação úmida, verificando se havia o efeito de sexo. Ficamos assim com um

Tabela 6.5: *Blattodea* - Valores observados e preditos pelo modelo.

	Valores Observados		Poisson	
Quant.	Machos	Fêmeas	Machos	Fêmeas
0	34	33	34,9	31,8
1	20	7	18,3	9,6
2	4	3	4,8	1,5
3	1	0	0,8	0,1

total de 88 indivíduos: 59 machos e 49 fêmeas. Como não temos o efeito da estação, os modelos testados para esta comida serão somente os modelos com efeito de sexo e o média geral.

Para esta comida há evidências para excesso de zeros e o modelo ZIP média geral é o melhor entre todos pelo critério AIC.

Modelo ZIP para *Isoptera Alates*

$$\log(\hat{\lambda}_{it}) = \hat{\beta}_0 \quad \text{e} \quad \hat{p}.$$

Tabela 6.6: *Isoptera-Alates* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

parâmetro	estimativa	erro padrão	valor t	p-valor
$\hat{p}$	0,9033	0,0328	27,53	<0,0001
$\hat{\beta}_0$	1,0372	0,2252	4,61	<0,0001

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = 1,0372 \quad \text{e} \quad \hat{p} = 0,9033.$$

O valor médio ingerido:

$$(1-\hat{p})\hat{\lambda} = (1-\hat{p})e^{\hat{\beta}_0} = 0,2728.$$

Tabela 6.7: *Isoptera Alates* - Valores observados e preditos pelo modelo.

Quant.	Observado	Modelo ZIP
0	80	80,0
1	1	1,4
2	2	2,0
3	3	1,9
4	1	1,3
5	0	0,8
6	1	0,4

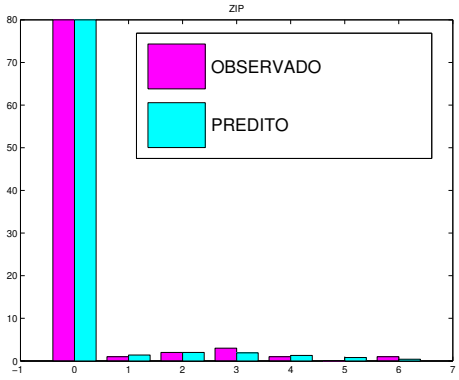

Figura 6.3: *Isoptera Alates* - Comparação de valores observados e preditos - Modelo ZIP média geral.

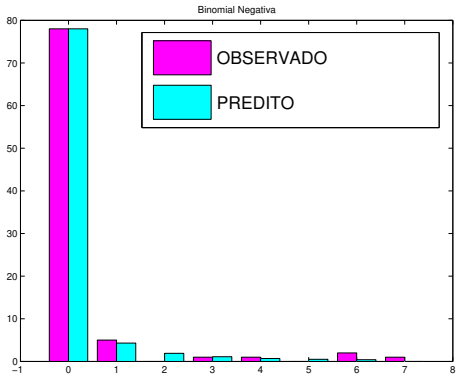

Figura 6.4: *Isoptera Non-Alates* - Comparação de valores observados e preditos - Modelo Binomial Negativa.

6.1.3 Modelos para *Isoptera Non-Alates*

Para *Isoptera Non-Alates* temos o mesmo fenômeno que para *Isoptera Alates*, somente foi consumida na estação úmida. Portanto, só testaremos se há ou não efeito do sexo. Pelo critério AIC o melhor modelo é Binomial Negativa média geral. Há evidências para excesso de zeros na amostra, porém o modelo ZIP tem um ajuste pior.

Modelo Binomial Negativa para *Isoptera Non-Alates*

$$\log(\hat{\lambda}_{it}) = \hat{\beta}_0 \quad \text{e} \quad \hat{\nu}.$$

Tabela 6.8: *Isoptera Non-Alates* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

parâmetro	estimativa	erro padrão	χ^2	p-valor
$\hat{\beta}_0$	-1,0433	0,4547	5,26	0,0218
$\hat{\nu}$	15,3578	6,7408	—	—

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -1,0433 \quad \text{e} \quad \hat{\nu} = 15,3578.$$

O valor médio ingerido:

$$\hat{\lambda} = e^{\hat{\beta}_0} = 0,3523.$$

Tabela 6.9: *Isoptera Non-Alates* - Valores observados e preditos pelo modelo.

Quant.	Observado	Modelo BinNeg
0	78	78,0
1	5	4,3
2	0	1,9
3	1	1,1
4	1	0,7
5	0	0,5
6	2	0,4
7	1	0,0

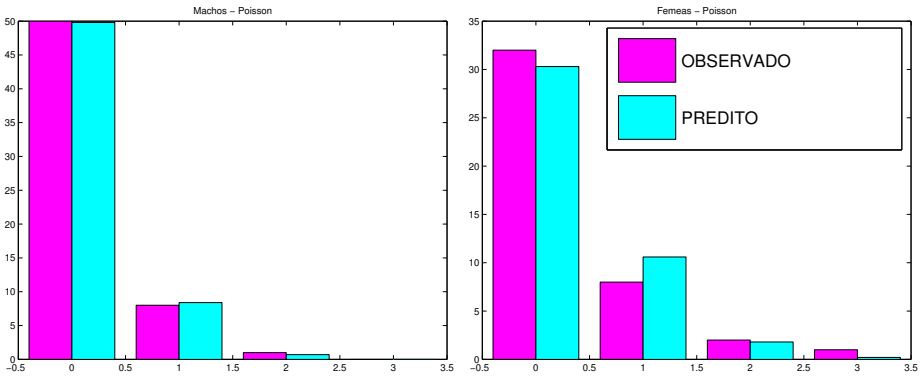

Figura 6.5: *Heteroptera* - Comparação de valores observados e preditos - Regressão de Poisson com efeito do sexo (a) Machos; (b) Fêmeas.

6.1.4 Modelos para *Heteroptera*

Para o tipo de comida *Heteroptera* temos resultados bastante surpreendentes. Os modelos não se diferenciam entre si. Parece haver uma pequena diferença entre os sexos, o que dá um efeito negativo para o sexo ao nível de 10%, ou seja, as fêmeas comem uma quantidade um pouco maior do que os machos. O modelo Poisson não apresenta superdispersão, pois a *deviance* dividida pelos graus de liberdade é 0,82. Ao mesmo tempo, há indícios de que estamos diante de um número excessivo de zeros pois a probabilidade extra de zeros estimada no modelo ZIP é significativa ao nível de 6%. Quanto ao gráfico que compara as previsões dos modelos também não percebemos diferenças significativas. Levando em conta o critério AIC e a parcimônia, ficamos com o modelo Poisson com efeito de sexo.

Modelo Poisson para *Heteroptera*:

$$\log(\lambda_{it}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1_{it}}.$$

Tabela 6.10: *Heteroptera* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

parâmetro	estimativa	erro padrão	χ^2	p-valor
$\hat{\beta}_0$	-1,0531	0,2582	16,64	<0,0001
$\hat{\beta}_1$	-0,7218	0,4082	3,13	0,0771

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -1,0531 - 0,7218x_{1_{it}}.$$

Para cada sub-população do modelo temos os valores preditos das quantidades médias ingeridas pelas fêmeas e machos:

$$\text{Fêmeas: } \hat{\lambda}_f = e^{\hat{\beta}_0} = 0,3489.$$
$$\text{Machos: } \hat{\lambda}_m = e^{\hat{\beta}_0+\hat{\beta}_1} = 0,1695.$$

Razão de médias:

Comparando fêmeas (f) e machos (m):

$$\frac{\hat{\lambda}_f}{\hat{\lambda}_m} = \frac{e^{\hat{\beta}_0}}{e^{\hat{\beta}_0+\hat{\beta}_1}} = \frac{1}{e^{\hat{\beta}_1}} = \frac{1}{e^{-0,7218}} = 2,0581 \quad \text{IC}_{90\%} : [0,6759; 3,4403]$$

Em média as fêmeas comem 2,0581 vezes mais *Heteroptera* do que os machos.

Tabela 6.11: *Heteroptera* - Valores observados e preditos pelo modelo

	Valores Observados		Modelo Poisson	
Quant.	Machos	Fêmeas	Machos	Fêmeas
0	50	32	49,8	30,3
1	8	8	8,4	10,6
2	1	2	0,7	1,8
3	0	1	0,0	0,2

6.1.5 Modelos para *Hemiptera*

Para esta comida o modelo escolhido é a regressão de Poisson somente com a média geral, ou seja, o modelo mais simples possível. O valor da *deviance* dividido pelos graus de liberdade é 0,8511. A Tabela A.5 do Apêndice A que apresenta a comparação entre os modelos nos mostra que não há diferença significativa entre estes. A probabilidade de zero no modelo ZIP não é significativa assim como não há efeitos significativos de sexo e estação.

Modelo Poisson para *Hemiptera*:

$$\log(\hat{\lambda}_{it}) = \hat{\beta}_0.$$

Tabela 6.12: *Hemiptera* - Estimativa do parâmetro com seu respectivo erro padrão, estatística do teste e p-valor para o modelo escolhido.

parâmetro	estimativa	erro padrão	χ^2	p-valor
$\hat{\beta}_0$	$-1,0696$	0,1690	40,04	<0,0001

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -1,0696.$$

O valor médio ingerido:

$$\hat{\lambda} = e^{\hat{\beta}_0} = 0,3431.$$

6.1.6 Modelos para *Larvae*

Começando pela regressão de Poisson para o modelo saturado encontramos

$$deviance/g.l. = 236,9755/98 = 2,4181,$$

que indica a presença de superdispersão. Os parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_3$ são significativos, mostrando efeito do sexo e da interação deste com a estação. Inspecionando os gráficos de envelope na Figura 6.7 para o modelo Poisson notamos a presença de um valor aberrante

Tabela 6.13: *Hemiptera* - Valores observados e preditos pelo modelo.

Quant.	Observado	Modelo Poisson
0	71	72,4
1	27	24,8
2	4	4,3

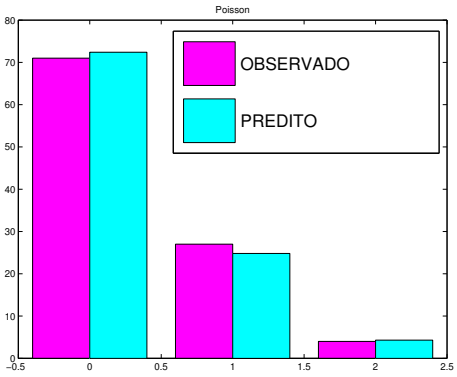

Figura 6.6: *Hemiptera* - Comparação de valores observados e preditos - Regressão de Poisson - média geral.

muito alto, correspondente ao indivíduo número 511 que ingeriu 31 unidades da comida em questão. A alternativa do modelo Binomial Negativa acomoda melhor os resíduos, apesar de ainda termos pontos fora do envelope.

Para todas as opções de modelos, ao retirarmos o indivíduo 511 da amostra, não teremos mais os efeitos sexo e interação significativos, o que indica que a presença do valor aberrante influenciou muito o modelo.

Comparando os modelos, notamos que aquele com o menor valor para o critério AIC é o modelo com efeitos aleatórios somente com a média geral. Ou seja, não há efeito de sexo e/ou estação influenciando na escolha da comida do tipo *Larvae*. Para efeito de ilustração, refizemos todos os modelos retirando o valor aberrante consumido pelo indivíduo 511 e percebemos a contribuição que esta observação traz ao valor da verossimilhança do modelo. Novamente, o menor valor de AIC é para o modelo com efeitos aleatórios somente com a média geral.

Outro aspecto importante em relação a esta comida, é que existe um número excessivo de zeros na amostra, pois temos um valor significativo para o p do modelo ZIP. Mas o melhor modelo para esta amostra é o modelo de efeitos aleatórios somente com a média geral, conforme Tabela A.6 do Apêndice A.

Modelo de Efeitos Aleatórios para *Larvae*:

$$\log(\hat{\lambda}_{it}) = \hat{\beta}_0 + u.$$

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -1,4786 \quad \text{e} \quad \hat{\sigma} = 1,3575.$$

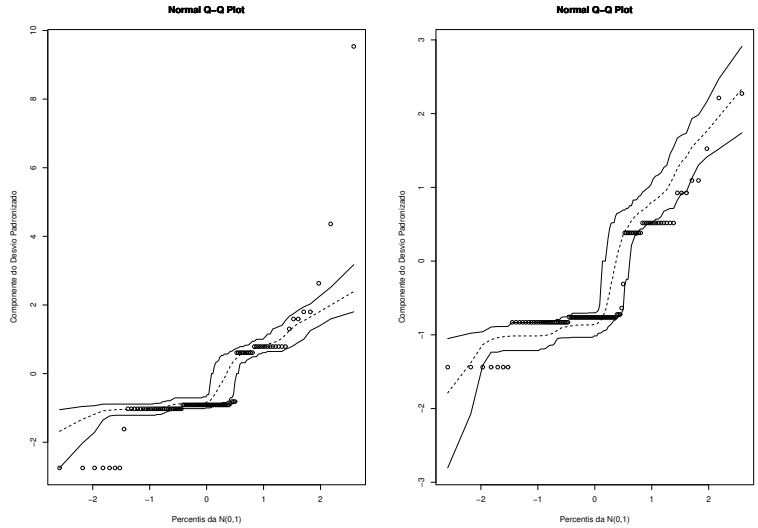

Figura 6.7: *Larvae* - Gráficos envelope para os modelos: (a) Poisson; (b) Binomial Negativa.

Tabela 6.14: *Larvae* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

Parâmetro	Estimativa	erro padrão	valor t	p-valor
$\hat{\beta}_0$	-1,4786	0,2884	-5,13	<,0001
$\hat{\sigma}$	1,3575	0,2134	6,36	<,0001

O valor médio ingerido:

$$\hat{\lambda} = e^{\hat{\beta}_0 + \frac{\hat{\sigma}^2}{2}} = 0,5728.$$

A Figura 6.8 mostra a comparação dos valores observados para a ingestão de *Larvae* e os valores preditos pelo modelo de efeitos aleatórios. O valor aberrante do indivíduo que consumiu 31 unidades não aparece no gráfico, mas o modelo foi feito para a amostra total. Os valores observados e preditos são mostrados na Tabela 6.15.

Estamos diante de uma amostra peculiar, pois a presença ou ausência nos cálculos do valor aberrante muda bastante os modelos obtidos. Apesar de termos selecionado um modelo pouco informativo, pois só temos a média geral, o modelo de efeitos aleatórios parece ser a melhor escolha neste caso.

Tabela 6.15: *Larvae* - Valores observados e preditos pelo modelo.

Quant.	Observado	Modelo Ef.Aleat.
0	69	71,9
1	25	18,4
2	5	6,0
3	1	2,5
4	0	1,2
5	0	0,6
6	1	0,4
31	1	0

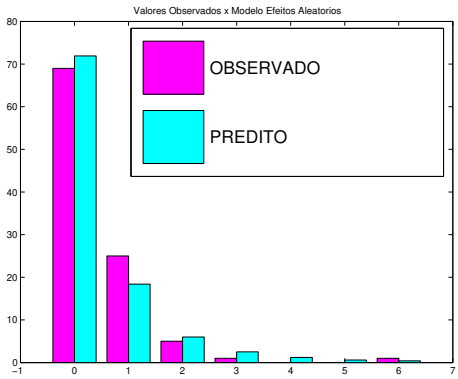

Figura 6.8: *Larvae* - Comparação de valores observados e preditos - Modelo de efeitos aleatórios com média geral.

6.1.7 Modelos para *Coleoptera*

Para a comida *Coleoptera* não temos evidência de superdispersão, nem de excesso de zeros. O melhor modelo neste caso é a regressão de Poisson com o efeito de sexo significativo ao nível de 10%.

Modelo Poisson para *Coleoptera*:

$$\log(\hat{\lambda}_{it}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1_{it}}.$$

Tabela 6.16: *Coleoptera* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

parâmetro	estimativa	erro padrão	χ^2	p-valor
$\hat{\beta}_0$	$-0,7655$	0,2236	11,72	0,0006
$\hat{\beta}_1$	0,4491	0,2707	2,75	0,0970

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -0,7655 + 0,4491 x_{1_{it}}.$$

Para cada sub-população do modelo temos os valores preditos das quantidades médias ingeridas pelas fêmeas e machos:

$$\text{Fêmeas: } \hat{\lambda}_f = e^{\hat{\beta}_0} = 0,4651.$$
$$\text{Machos: } \hat{\lambda}_m = e^{\hat{\beta}_0+\hat{\beta}_1} = 0,7288.$$

Razão de médias:

Comparando fêmeas (f) e machos (m):

$$\frac{\hat{\lambda}_f}{\hat{\lambda}_m} = \frac{e^{\hat{\beta}_0}}{e^{\hat{\beta}_0+\hat{\beta}_1}} = \frac{1}{e^{\hat{\beta}_1}} = 0,6382.$$

Em média os machos comem 1,5669 (1/0,6382) vezes mais *Coleoptera* do que as fêmeas. $\text{IC}_{90\%}$: $[0,8692; 2,2646]$

Tabela 6.17: *Coleoptera* - Valores observados e preditos pelo modelo.

	Valores Observados		Poisson	
Quant.	Machos	Fêmeas	Machos	Fêmeas
0	29	28	28,5	27,0
1	21	12	20,7	12,6
2	6	2	7,6	2,9
3	2	0	1,8	0,5
4	1	1	0,3	0,1

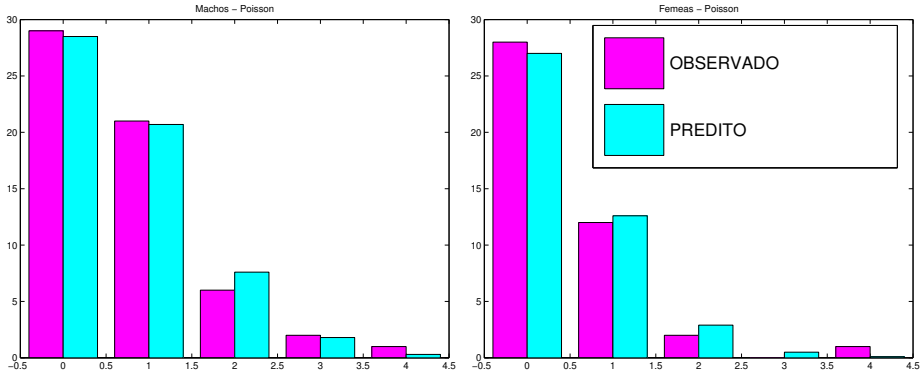

Figura 6.9: *Coleoptera* - Comparação de valores observados e preditos - Regressão de Poisson com efeito do sexo (a) Machos; (b) Fêmeas.

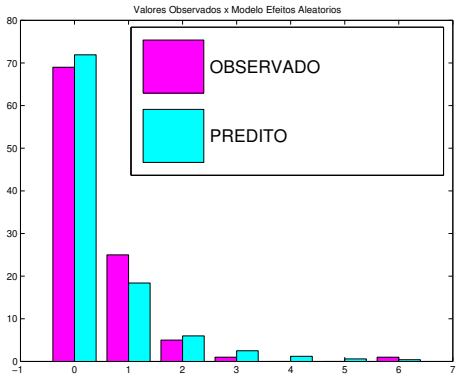

Figura 6.10: *Hymenoptera* - Comparação de valores observados e preditos - Modelo de efeitos aleatórios com média geral.

6.1.8 Modelos para *Hymenoptera*

Não há efeito de sexo e estação em nenhum modelo. Temos evidência de excesso de zeros mas o modelo que se ajusta melhor pelo critério AIC é o modelo de efeitos aleatórios com média geral.

Modelo de Efeitos Aleatórios para *Hymenoptera*:

$$\log(\hat{\lambda}_{it}) = \hat{\beta}_0 + u.$$

Tabela 6.18: *Hymenoptera* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

Parâmetro	Estimativa	erro padrão	valor t	p-valor
$\hat{\beta}_0$	-0,0662	0,1588	-0,42	0,6778
$\hat{\sigma}$	0,9444	0,1386	6,81	<0,0001

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -0,0662 \quad \text{e} \quad \hat{\sigma} = 0,9444.$$

O valor médio ingerido:

$$\hat{\lambda} = e^{\hat{\beta}_0 + \frac{\hat{\sigma}^2}{2}} = 1,4620.$$

Tabela 6.19: *Hymenoptera* - Valores observados e preditos pelo modelo

Quant.	Observado	Modelo EfAleat
0	39	40,6
1	32	27,3
2	11	14,8
3	7	7,9
4	8	4,4
5	1	2,5
6	0	1,5
7	1	0,9
8	1	0,6
9	0	0,4
10	1	0,3
11	0	0,2
12	1	0,1

6.1.9 Modelos para *Aranae*

Não há indício de superdisperão, pois o valor da *deviance* dividida pelos graus de liberdade no modelo Poisson saturado é de 1,1084. Em relação aos efeitos, somente o efeito para sexo deu significativo. Para modelo ZIP a probabilidade extra de zeros não deu significativa. Como a regressão de Poisson não apresentou problemas de superdispersão, era pouco provável que houvesse problemas de excesso de zeros.

Comparando os modelos Poisson, Binomial Negativa e Efeitos Aleatórios em termos de previsão e de valores do critério AIC não encontramos diferenças significativas. Somente para efeito de ilustração, no apêndice C apresentamos para a comida *Aranae* uma comparação entre os diferentes modelos possíveis, uma vez que estes não parecem se diferenciar tanto assim. Escolheremos o modelo Poisson como melhor modelo por ser o mais parcimonioso.

Modelo Poisson para *Aranae*:

$$\log(\lambda_{it}) = \beta_0 + \beta_1 x_{1_{it}}.$$

Modelo ajustado:

$$\log(\hat{\lambda}_{it}) = -0,1236 - 0,7350 x_{1_{it}}.$$

Tabela 6.20: *Aranae* - Estimativa dos parâmetros com seus respectivos erro padrão, estatística do teste e p-valor para o modelo escolhido.

parâmetro	estimativa	erro padrão	χ^2	p-valor
$\hat{\beta}_0$	$-0,1236$	0,1622	0,58	0,4461
$\hat{\beta}_1$	$-0,7350$	0,2575	8,15	0,0043

Para cada sub-população do modelo temos os valores preditos das quantidades médias ingeridas pelas fêmeas e machos:

$$\text{Fêmeas: } \hat{\lambda}_f = e^{\hat{\beta}_0} = 0,8837.$$
$$\text{Machos: } \hat{\lambda}_m = e^{\hat{\beta}_0+\hat{\beta}_1} = 0,4238.$$

Razão de médias:

Comparando fêmeas (f) e machos (m):

$$\frac{\hat{\lambda}_f}{\hat{\lambda}_m} = \frac{e^{\hat{\beta}_0}}{e^{\hat{\beta}_0+\hat{\beta}_1}} = \frac{1}{e^{\hat{\beta}_1}} = \frac{1}{e^{-0,7350}} = 2,0855 \quad \text{IC}_{95\%} : [1,0328; 3,1381].$$

Em média as fêmeas comem 2,0855 vezes mais aranhas do que os machos.

6.2 Discussão

Em termos de modelos lineares para dados de contagem, nossa aplicação aos dados reais foi bastante ilustrativa. Como efeitos significativos temos somente sexo, em 4 categorias de comida: *Blattodea*, *Heteroptera*, *Coleoptera* e *Aranae*. Somente para esta última o nível de significância do sexo foi menor que 5%. Para estas 4 comidas, o modelo Poisson com o efeito do sexo se apresentou como o melhor modelo em termos de critério AIC. Para *Hemiptera* temos como melhor modelo Poisson com a média geral.

Para as outras 4 comidas, tivemos a presença de uma superdispersão leve, de tal forma que para *Larvae* e *Hymenoptera-Formicidae* o modelo de Efeitos Aleatórios foi o melhor. Para *Isoptera non-Alates* foi o modelo Binomial Negativa e para *Isoptera Alates* o modelo ZIP, todos sem efeitos significativos, somente com a média geral.

A Tabela 6.21 resume os resultados obtidos, além de apresentar o valor da *deviance* dividida pelos graus de liberdade dos modelos Poisson citados. Para os quatros tipos de comida que não tiveram o modelo Poisson como melhor modelo, apresentamos o valor da

Tabela 6.21: *Adenomera* - Resumo comparativo dos modelos estimados.

Tipo de comida	Melhor modelo	efeito	*dev*/g.l.
Blattodea	Poisson	sexo	0,9700
Heteroptera	Poisson	sexo	0,8200
Coleoptera	Poisson	sexo	1,1537
Aranae	Poisson	sexo	1,1278
Hemiptera	Poisson	média geral	0,8511
Isoptera Alates	ZIP	média geral	1,3825
Isoptera non-Alates	Binomial Negativa	média geral	1,7542
Larvae	Efeitos Aleatórios	média geral	2,9800
Hymenoptera	Efeitos Aleatórios	média geral	2,3240

deviance dividida pelos graus de liberdade referente a um modelo Poisson com média geral.

Os dados para alimentação de *Adenomera* mostram somente uma pequena diferença entre os sexos para 4 tipos de comida sendo que para as demais não encontramos efeitos significativos. O modelo ZIP foi escolhido como o melhor modelo somente para um tipo de comida. O caso em que o melhor modelo foi o de Binomial Negativa é interessante por que a superdispersão não é tão grande assim. Nos dois casos em que a superdispersão foi um pouco mais alta, a solução de efeitos aleatórios mostrou-se a mais razoável.

Capítulo 7

Conclusão

Este trabalho procurou estudar os modelos lineares para dados de contagem. Na classe dos modelos lineares generalizados, começamos com a regressão de Poisson. Se esta apresentasse superdispersão, buscaríamos as soluções mais comuns: modelo Binomial Negativa, Efeitos Aleatórios ou ZIP.

A aplicação em estudo referente à alimentação dos sapos da espécie *Adenomera* foi bastante ilustrativa, inclusive apresentou alguns resultados curiosos. No caso das comidas em que os modelos Poisson apresentavam superdispersão, esta não era tão alta, mas mesmo assim, encontramos alternativas melhores.

No caso em que o melhor modelo foi o Poisson constatamos um resultado bastante uniforme em termos de critério AIC. Os valores dos diferentes modelos são marginalmente diferentes. O efeito do sexo, quando presente, só foi realmente significativo para *Aranae*. Para as outras comidas temos que o intervalo de confiança para a razão de médias inclui a unidade, o que torna a diferença entre fêmeas e machos pouco significativa. Devemos lembrar que os intervalos de confiança são calculados com métodos assintóticos, as aproximações provavelmente aumentam a amplitude destes.

Os métodos de diagnóstico mostraram-se bastante úteis para detectar o ponto influente em *Larvae*, principalmente o método de afastamento da verossimilhança. Apesar do valor da medida ser uma aproximação do verdadeiro valor, como foi discutido no Capítulo 2, ela é suficiente para identificar pontos influentes.

Para os modelos ZIP e mistos não encontramos medidas de diagnóstico tais como as que foram adaptadas para os MLGs. A verossimilhança para os modelos mais complexos envolvem misturas de variáveis ou funções de verossimilhança que não apresentam uma forma explícita, e assim, não temos valores para medidas clássicas como a matriz de

projeção **H**, os resíduos de *deviance* ou de Pearson. Um assunto bastante interessante para pesquisas futuras é estudar uma maneira de construir técnicas de diagnóstico para tais modelos.

Apêndice A

Critério AIC para os modelos testados

Tabela A.1: Modelos testados para *Blattodea*.

Modelo	Parâmetros	AIC
Poisson	saturado	176,8
	ef.principais	178,3
	sexo	**176,3**
	estação	179,3
	média geral	177,3
Binomial Negativa	saturado	178,8
	ef.principais	180,2
	sexo	178,3
	estação	181,2
	média geral	179,3
Efeitos Aleatórios	saturado	179,8
	ef.principais	180,3
	sexo	178,3
	estação	181,3
	média geral	179,2
ZIP	saturado	178,8
	ef.principais	180,2
	sexo	178,3
	estação	181,2
	média geral	179,2

Tabela A.2: Modelos testados para *Isoptera Alates*.

Modelo	Parâmetros	AIC
Poisson	sexo	143,5
	média geral	145,4
Binomial Negativa	sexo	92,4
	média geral	91,1
Efeitos Aleatórios	sexo	95,7
	média geral	94,7
ZIP	sexo	87,0
	média geral	**85,0**

Tabela A.3: Modelos testados para *Isoptera Non-Alates*.

Modelo	Parâmetros	AIC
Poisson	sexo	182,6
	média geral	181,9
Binomial Negativa	sexo	106,9
	média geral	**105,1**
Efeitos Aleatórios	sexo	109,2
	média geral	107,2
ZIP	sexo	109,3
	média geral	110,5

Tabela A.4: Modelos testados para *Heteroptera*.

Modelo	Parâmetros	AIC
Poisson	saturado	132,6
	ef.principais	130,8
	sexo	**128,8**
	estação	132,0
	média geral	130,1
Binomial Negativa	saturado	133,2
	ef.principais	131,4
	sexo	129,4
	estação	131,9
	média geral	130,0
Efeitos Aleatórios	saturado	134,6
	ef.principais	132,8
	sexo	129,5
	estação	132,1
	média geral	130,1
ZIP	saturado	133,0
	ef.principais	131,2
	sexo	129,2
	estação	131,9
	média geral	130,0

Tabela A.5: Modelos testados para *Hemiptera*.

Modelo	Parâmetros	AIC
Poisson	saturado	156,9
	ef.principais	155,3
	sexo	153,5
	estação	154,1
	média geral	**152,4**
Binomial Negativa	saturado	158,9
	ef.principais	157,3
	sexo	155,5
	estação	156,1
	média geral	154,4
Efeitos Aleatórios	saturado	158,9
	ef.principais	157,3
	sexo	155,5
	estação	156,1
	média geral	154,4
ZIP	saturado	158,0
	ef.principais	156,7
	sexo	154,8
	estação	155,6
	média geral	153,8

Tabela A.6: Modelos testados para *Larvae*.

Modelo	Parâmetros	AIC	AIC sem ind 511
Poisson	saturado	320,0	191,6
	ef.principais	332,0	189,9
	sexo	373,5	188,3
	estação	332,8	188,2
	média geral	378,0	186,8
Binomial Negativa	saturado	213,7	188,0
	ef.principais	215,5	185,2
	sexo	223,2	183,6
	estação	213,5	183,5
	média geral	223,0	181,9
Efeitos Aleatórios	saturado	210,6	186,0
	ef.principais	209,8	184,1
	sexo	208,7	182,5
	estação	207,8	182,3
	média geral	**206,7**	180,7
ZIP	saturado	253,3	190,9
	ef.principais	265,4	189,2
	sexo	323,7	187,5
	estação	265,7	187,6
	média geral	330,7	186,1

Tabela A.7: Modelos testados para *Coleoptera*.

Modelo	Parâmetros	AIC
Poisson	saturado	220,2
	ef.principais	219,8
	sexo	**218,8**
	estação	220,3
	média geral	219,7
Binomial Negativa	saturado	221,4
	ef.principais	220,8
	sexo	219,6
	estação	220,9
	média geral	220,1
Efeitos Aleatórios	saturado	221,3
	ef.principais	220,7
	sexo	220,8
	estação	220,9
	média geral	221,7
ZIP	saturado	222,0
	ef.principais	221,6
	sexo	220,5
	estação	221,8
	média geral	221,1

Tabela A.8: Modelos testados para *Hymenoptera*.

Modelo	Parâmetros	AIC
Poisson	saturado	399,4
	ef.principais	399,2
	sexo	398,3
	estação	397,2
	média geral	396,3
Binomial Negativa	saturado	348,0
	ef.principais	346,7
	sexo	345,1
	estação	344,7
	média geral	343,1
Efeitos Aleatórios	saturado	345,7
	ef.principais	344,0
	sexo	342,9
	estação	342,0
	média geral	**340,9**
ZIP	saturado	379,6
	ef.principais	379,5
	sexo	377,8
	estação	377,6
	média geral	375,9

Tabela A.9: Modelos testados para *Aranae*.

Modelo	Parâmetros	AIC
Poisson	saturado	214,7
	ef.principais	214,4
	sexo	**214,9**
	estação	219,8
	média geral	221,3
Binomial Negativa	saturado	215,9
	ef.principais	215,4
	sexo	215,8
	estação	219,9
	média geral	220,9
Efeitos Aleatórios	saturado	215,7
	ef.principais	215,2
	sexo	215,6
	estação	219,6
	média geral	220,6
ZIP	saturado	216,6
	ef.principais	216,2
	sexo	216,7
	estação	221,0
	média geral	222,0

Apêndice B

Gráficos de Diagnóstico

B.1 *Isoptera Non-alates*

Para *Isoptera Non-alates* escolhemos o modelo Binomial Negativa. Para este tipo de comida há um número bastante grande de zeros na amostra, porém em termos de critério AIC, o modelo Binomial Negativa apresenta o menor valor. Além disso, comparando os gráficos na Figura B.1 , percebemos que o modelo Binomial Negativa apresenta uma previsão um pouco melhor do que o modelo ZIP.

Na figura B.2 comparamos os gráficos de envelope para o modelo Poisson e Binomial Negativa. O modelo Poisson apresenta superdispersão leve (*deviance*/g.l.= 1,7542) e notamos vários pontos fora da banda de confiança. Para o modelo Binomial Negativa temos uma acomodação melhor dos desvios na banda de confiança, mas podemos perceber que na parte inferior do gráfico há uma grande concentração de pontos. Estes se referem aos animais que não ingeriram este tipo de comida.

Finalmente na figura B.3 temos os gráficos de resíduos de *deviance* contra preditor linear e afastamento da verossimilhança para cada observação. Como estamos diante de um modelo média geral, verificamos que os resíduos concentram-se numa mesma faixa, estando sobrepostos. No gráfico do afastamento da verossimilhança identificamos alguns pontos em destaque, acima da faixa em que se encontram a maioria dos valores aproximados da LD_i. Estes pontos referem-se aos animais que ingeriram de 3 a 7 unidades da comida. Mas, este pontos não são considerados influentes, pois a presença ou ausência destes na amostra não muda a interpretação dos parâmetros.

B.2 *Larvae*

Temos que para *Larvae* o ponto extremo identificado é o correspondente ao indivíduo 511, que ingeriu 31 unidades da comida em questão. Apresentamos os gráficos de diagnóstico

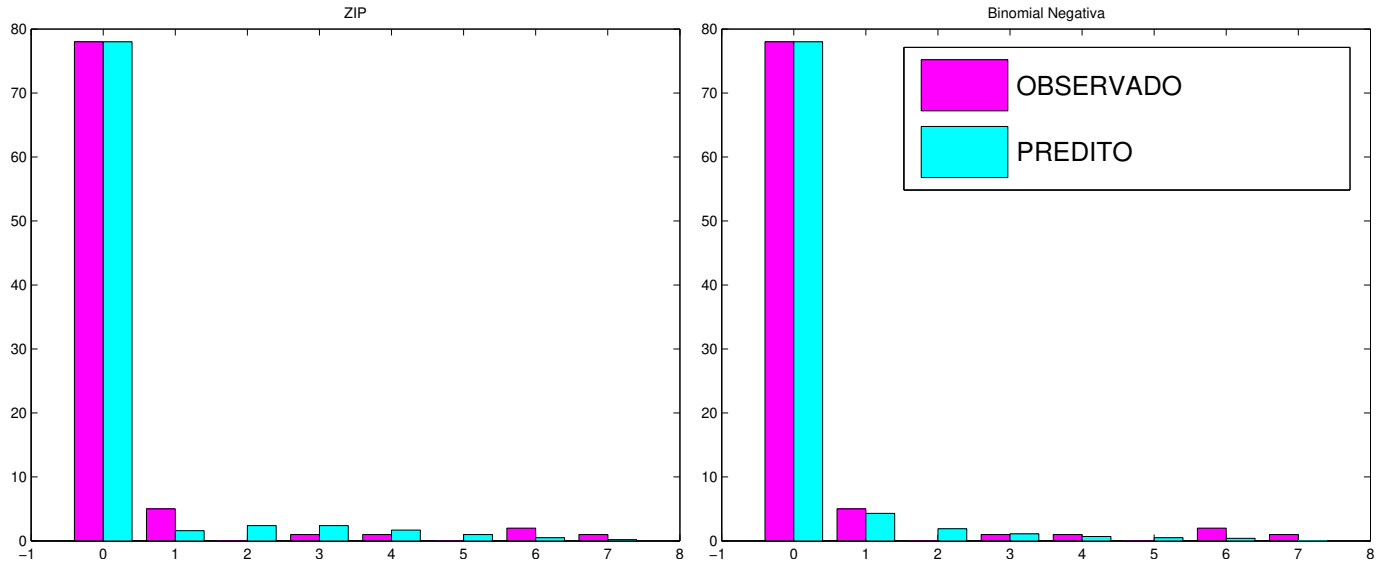

Figura B.1: *Isoptera Non-alates* - Valores preditos para (a) Modelo ZIP (b) Modelo Binomial Negativa.

para o modelo Poisson saturado, ou seja, com os efeitos principais e a interação.

No gráfico (a) da Figura B.4 em que temos o gráfico dos resíduos da *deviance* contra os preditores lineares podemos identificar o ponto extremo no canto superior direito.

No gráfico (b) da Figura B.4 com os valores aproximados do afastamento da verossimilhança para cada observação podemos identificar o ponto extremo no canto superior esquerdo. O distanciamento deste ponto é muito grande, pois os valores restantes se encontram num intervalo de 0,0086 a 1,5857, enquanto para o ponto aberrante é 27,6601.

Constatamos que se estimarmos uma regressão de Poisson saturada para *Larvae* incluindo todas as observações, os efeitos de sexo e interação sexo-estação são significativos. Se retirarmos a observação referente ao indivíduo 511 da amostra, não temos mais efeitos significativos. Portanto, este ponto é considerado influente por mudar a interpretação dos parâmetros.

O gráfico de envelope é bastante útil para identificar a superdispersão. Na figura B.5 encontramos vários pontos fora da banda de confiança e verificamos que os resíduos estão concentrados na parte inferior do gráfico, o que indica que este modelo apresenta superdispersão.

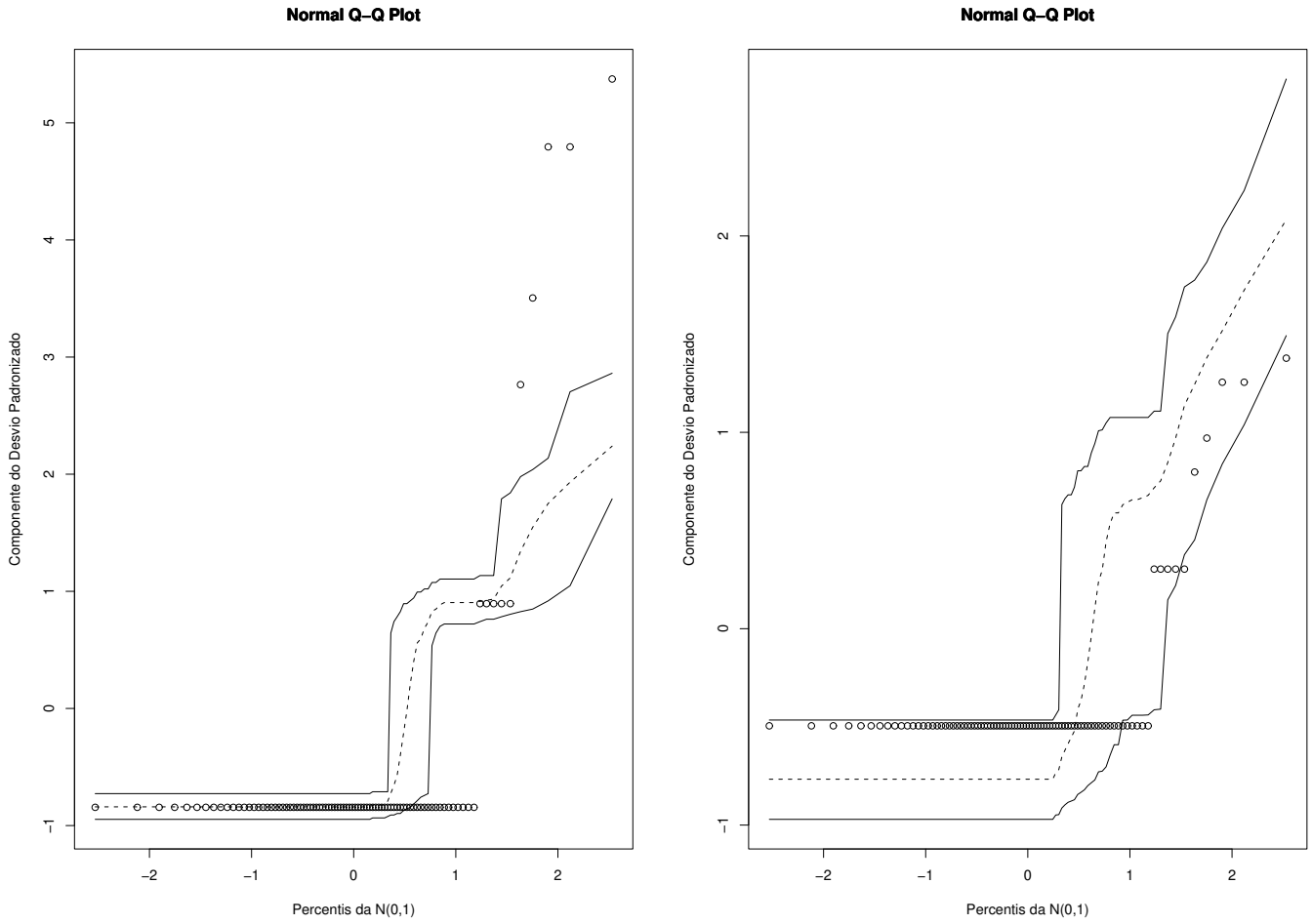

Figura B.2: *Isoptera Non-alates* - Gráfico envelope para (a) regressão de Poisson (b) Binomial Negativa.

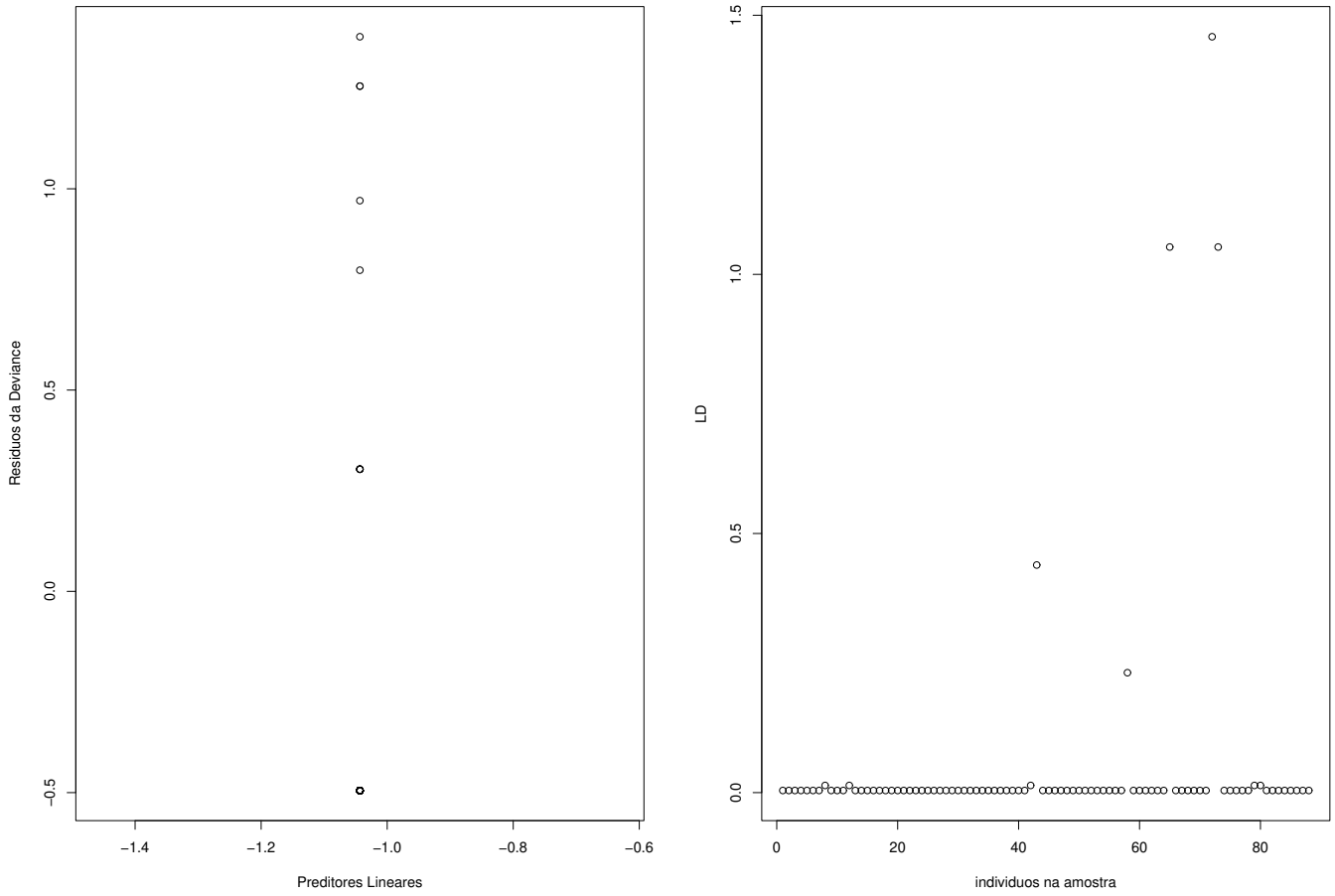

Figura B.3: *Isoptera Non-alates* - Modelo Binomial Negativa média geral (a) resíduos de *deviance* contra preditor linear; (b) afastamento da verossimilhança para cada observação.

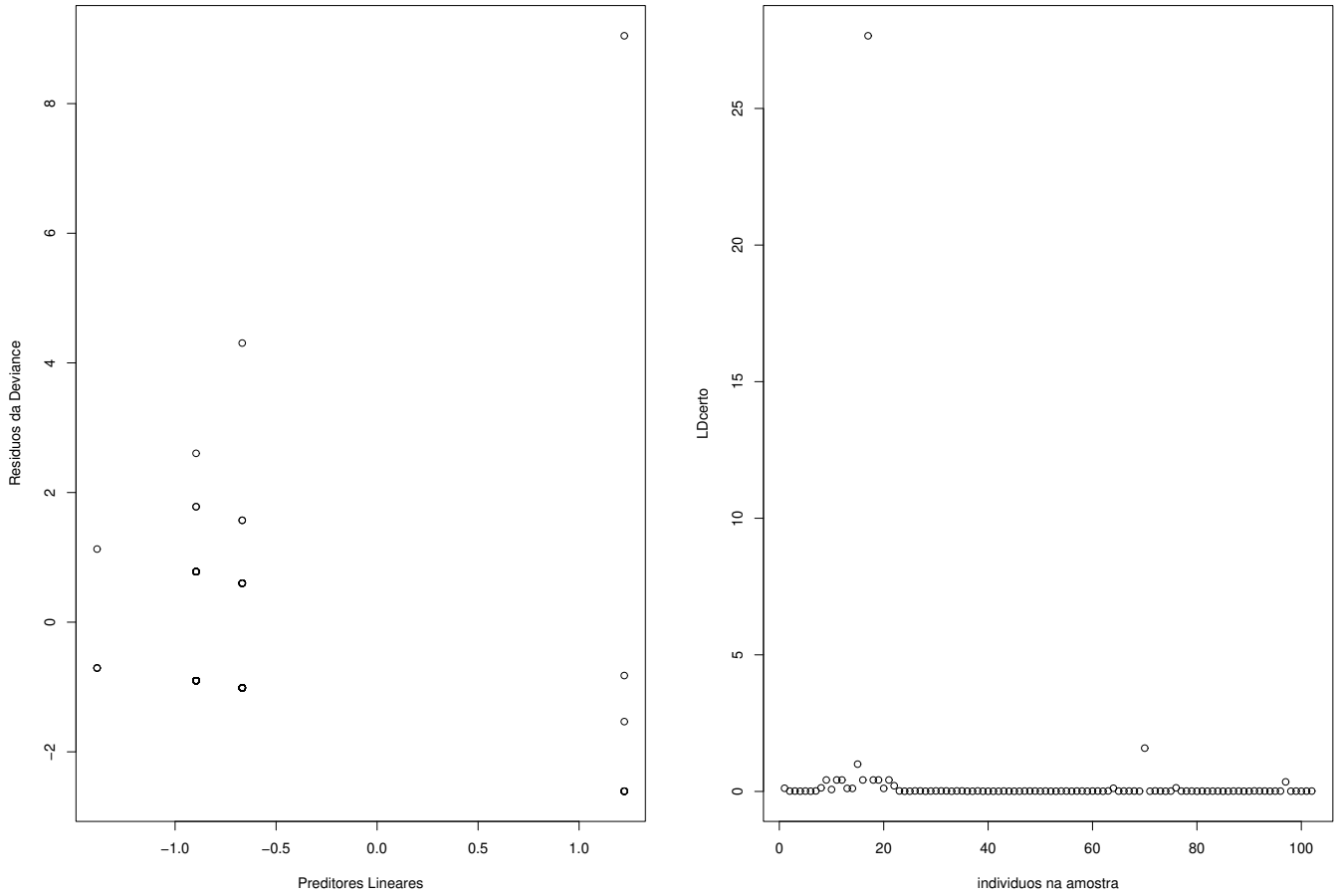

Figura B.4: *Larvae* - (a) resíduos de *deviance* contra preditor linear; (b) afastamento da verossimilhança para cada observação.

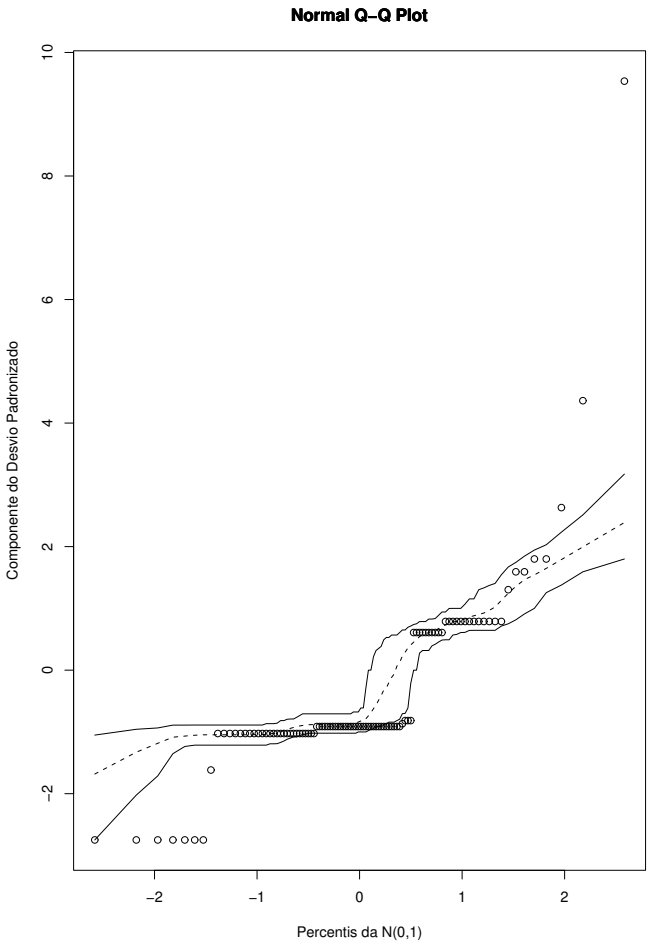

Figura B.5: *Larvae* - Gráfico envelope para regressão de Poisson - modelo saturado.

B.3 *Aranae*

O melhor modelo para a comida *Aranae* foi o modelo Poisson com efeito de sexo.

No gráfico (a) da Figura B.6, em que temos o gráfico dos resíduos da *deviance* contra os preditores lineares, não identificamos nenhum ponto extremo.

No gráfico (b) da Figura B.6, em que temos o gráfico do afastamento da verossimilhança para cada observação, podemos identificar dois pontos no canto superior esquerdo, porém neste caso o intervalo entre todos os valores de LD_i vai de 0,0004 a 0,5294, portanto não classificaríamos estes pontos como aberrantes. Constatamos que estes pontos se referem aos valores mais altos de ingestão de *Aranae*, ou seja, houve um indivíduo que ingeriu 4 unidades e outro que ingeriu 5 unidades deste tipo de comida.

Neste caso, se retirarmos da amostra um das duas observações em questão, ou mesmo se retirarmos as duas observações e reestimarmos as regressões de Poisson, a interpretação dos parâmetros é a mesma, ou seja, somente o efeito do sexo é significativo. Desta forma, estes pontos não foram considerados influentes.

O modelo de regressão de Poisson não apresenta superdispersão, como podemos constatar no gráfico envelope na figura B.7 No entanto, é interessante notar que para os valores de contagem mais altos, os resíduos padronizados da *deviance* encontram-se fora da banda de confiança de 95%, o que mostra que modelos lineares generalizados são sensíveis a pontos extremos.

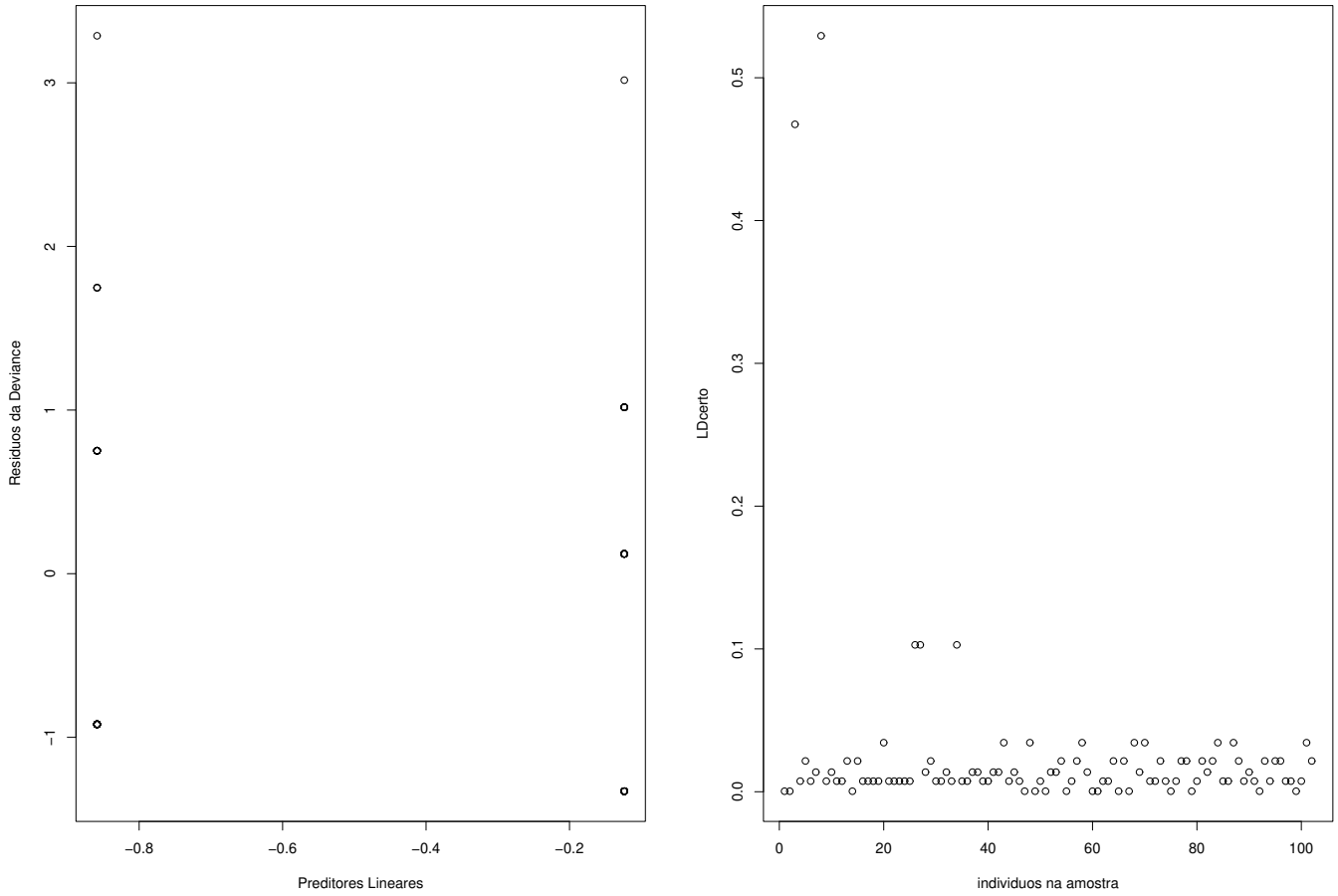

Figura B.6: *Aranae* - (a) resíduos de *deviance* contra preditor linear; (b) afastamento da verossimilhança para cada observação.

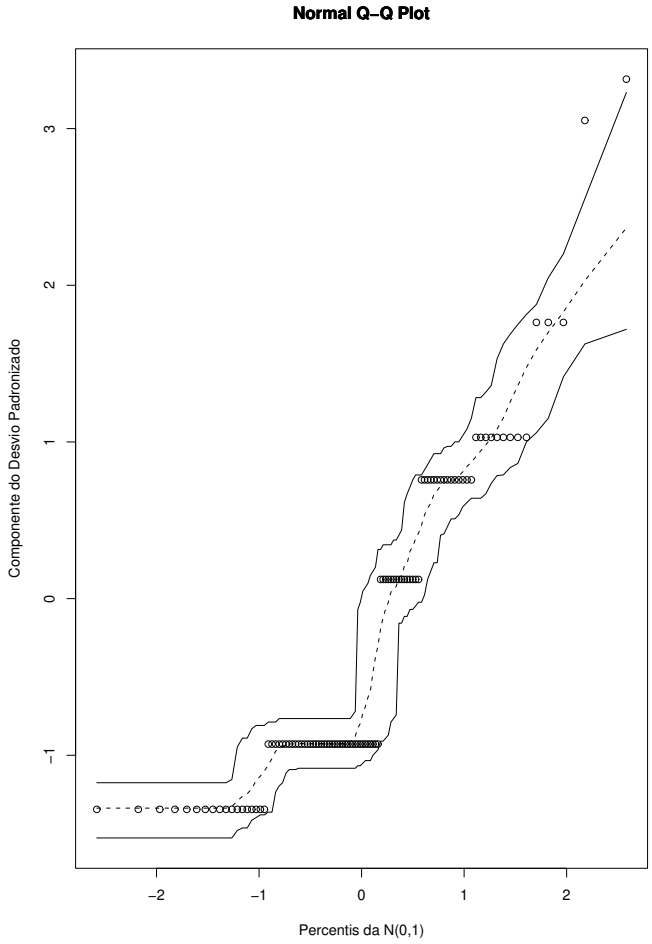

Figura B.7: *Aranae* - Gráfico envelope para regressão de Poisson - modelo com efeito de sexo.

Apêndice C

Comparação de modelos diferentes para *Aranae*

Como temos uma diferença muito pequena no valor do critério AIC para os modelos Poisson, Binomial Negativa e de Efeitos Aleatórios com o sexo como efeito, comparamos a previsão para os três modelos na tabela C.1. Nas figuras C.1 e C.2 temos a comparação entre os quantidades ingeridas observadas e previstas pelos três modelos.

Comparamos também os modelos Poisson e Binomial Negativa com o efeito do sexo significativo. Nota-se pela figura C.3 que não há praticamente diferença entre o gráfico envelope para os dois modelos, uma vez que não estamos diante de superdispersão. A banda de confiança para o gráfico envelope da Binomial Negativa é ligeiramente maior.

Tabela C.1: *Aranae* - Valores observados e preditos pelos modelos

	Valores Observados		Modelo Poisson		Binomial Negativa		Efeitos Aleatórios	
Quant.	Machos	Fêmeas	Machos	Fêmeas	Machos	Fêmeas	Machos	Fêmeas
0	40	18	38,6	17,8	39,4	19,2	39,5	19,3
1	15	15	16,4	15,7	15,2	14,1	15,1	14,1
2	3	9	3,5	6,9	3,6	6,4	3,6	6,3
3	0	0	0,5	2,0	0,7	2,3	0,7	2,3
4	1	0	0,1	0,5	0,1	0,7	0,1	0,7
5	0	1	0,0	0,1	0,0	0,2	0,2	0,2

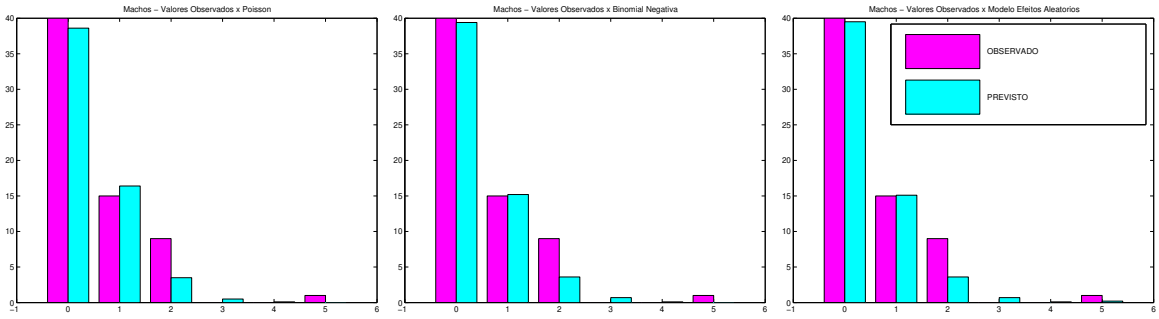

Figura C.1: *Aranae* Modelos com efeito de sexo - Valores observados e preditos para machos (a) Poisson; (b) Binomial Negativa (c) Efeitos Aleatórios.

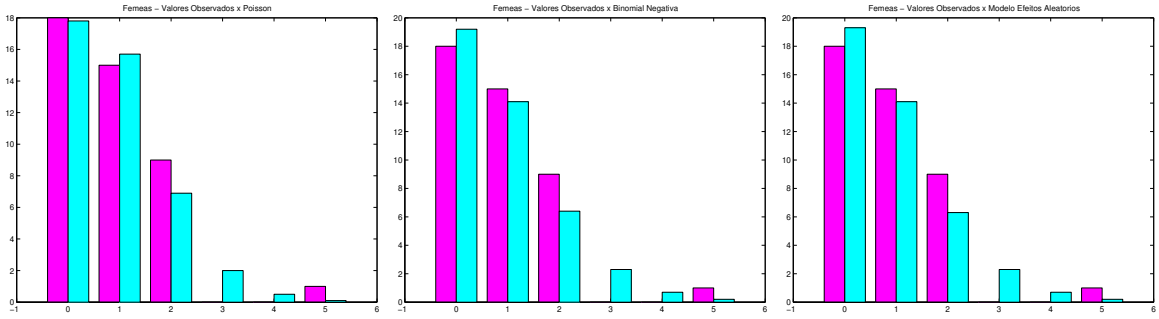

Figura C.2: *Aranae* - Modelos com efeito de sexo - Valores observados e preditos para fêmeas (a) Poisson; (b) Binomial Negativa (c) Efeitos Aleatórios.

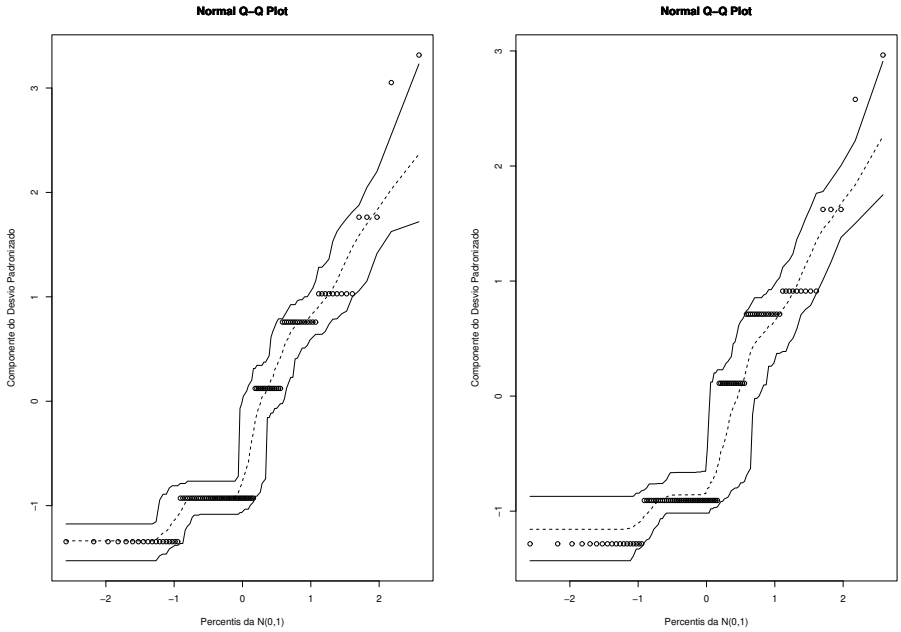

Figura C.3: Gráficos envelope para os modelos: (a) Poisson; (b) Binomial Negativa.

Referências Bibliográficas

Agresti, A. (2002). *Categorical Data Analysis, Second Edition.* Wiley Series in Probability and Statistics, John Wiley and Sons, Inc.

Bartlett, M. S. (1936). *The Square Root Transformation in Analysis of Variance.* Supplement to the Journal of the Royal Statistical Society, **3**, 68-78.

Böhning, D., Dietz, E., Schiattmann, P., Mendonça, L. e Kirchner, U. (1999). *The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology.* Journal of the Royal Statistical Society, **162**, 195-209.

Box, G. E. P., Cox, D. R. (1964). *An Analysis of Transformations.* Journal of the Royal Statistical Society, Série B, **26**, 211-252

Casella, G. e Berger, R. L. (2001). *Statistical Inference, Second Edition.* Duxbury, Thomson Learning.

Cláudio, D. M. e Marins, J. M. (1988). *Cálculo Numérico Computacional - Teoria e Prática.* Editora Atlas.

Draper, N. R., Smith, H. (1981). *Applied Regression Analysis, Second Edition.* Wiley Series in Probability and Statistics, John Wiley and Sons, Inc.

Durrett, R. (1991). *Probability: Theory and Examples.* Wadsworth and Brooks/Cole, Pacific Grove.

Finney, D. J. (1971). *Probit Analysis.* Cambridge University Press.

Hinde, J. e Demétrio, C. G. B. (1998). *Overdispersion model and estimation.* Computational Statistics and Data Analysis, **27**, 151-170.

Hirayama, T. (1981). *Non-smoking wives of heavy smokers have a higher risk of lung cancer: a study from Japan.* British Medical Journal, **282**, 183-185.

James, B. R. (1996) *Probabilidade: um curso em nível intermediário, Segunda Edição.* Projeto Euclides, Instituto de Matemática Pura e Aplicada.

Lambert, D. (1982). *Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing.* Technometrics, **34**, 1-14.

Lawless, J. F. (1987). *Negative Binomial and Mixed Poisson Regression.* The Canadian Journal of Statistics, **15**, 209-225.

McCullagh, N. e Nelder, J. A. (1989). *Generalized Linear Models, Second Edition.* Chapman and Hall, London.

McCullogh, C. E. e Searle, S. R. (2001). *Generalized, Linear and Mixed Models.* John Wiley and Sons, Inc.

Montgomery, D. C., Peck, E. A. (1991). *Introduction to Linear Regression Analysis, Second Edition.* Wiley Series in Probability and Statistics, John Wiley and Sons, Inc.

Myers, H. M., Montgomery, D. C., Vining, G. G. (2001). *Generalized Linear Models with Applications in Engineering and the Sciences.* Wiley Series in Probability and Statistics, John Wiley and Sons, Inc.

Nelder, J.A, Wedderburn, R.W.M. (1972). *Generalized Linear Models.* Journal of the Royal Statistical Society, Série A, **135**, 370-384.

Paula, G. A. (2004). *Modelos de Regressão com apoio computacional.* Instituto de Matemática e Estatística, Universidade de São Paulo.

Pregibon, D. (1981). *Logistic Regression Diagnostics.* The Annals of Statistics, **9**, 705-724.

Paulino, C. D., Singer, J. da M. (2006). *Análise de Dados Categorizados.* Editora Edgard Blücher.

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

Slymen, D. J., Ayala, G. X., Arredondo, E. M., Elder, J. P. (2006). *A demonstration of modelling count data with an application to physical activity.* Epidemiologic

Perspectives & Innovations, http://www.epi-perspectives.com/content/3/1/3.

Tukey, J. W. (1957). *On the Comparative Anatomy of Transformations.* The Annals of Mathematical Statistics, **28**, 602-632.

Venables, W. N. e Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus, Third Edition.* Springer, New York.

Wedderburn, R.W.M. (1974). *Quasi-likelihood functions, Generalized Linear Models and the Gauss-Newton Method.* Biometrika, **61**, 439-447.

Weiss, G., Weiss, M. (1963). *On the Picard property of lacunary power series.* Studia Mathematica, **22**, 221-245.