Filipa Margarida de Almeida Januário

Modelos de regressão para dados de contagem e estimação da abundância de aves na cidade do Porto

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
setembro de 2012

Filipa Margarida de Almeida Januário

Modelos de regressão para dados de contagem e estimação da abundância de aves na cidade do Porto

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Engenharia Matemática

Orientadora: Prof.ª Doutora Ana Rita Gaio

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
setembro de 2012

Agradecimentos

Apesar do processo solitário a que qualquer investigador está destinado na elaboração de uma tese, a realização deste estudo só foi possível com o apoio, incentivo e colaboração de várias pessoas, às quais gostaria de exprimir algumas palavras de agradecimento e profundo reconhecimento.
Em particular, gostaria de agradecer:

À Prof.ª Doutora Ana Rita Gaio pela disponibilidade manifestada para orientar este trabalho, pela exigência de método e rigor e pela confiança e apoio que sempre me concedeu. Um agradecimento especial por tudo o que me ensinou e que em muito enriqueceu a minha formação.

À equipa do CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos da Universidade do Porto, em particular, à Filipa Guilherme, à Prof.ª Doutora Cláudia Fernandes e ao Prof. Doutor Paulo Farinha-Marques pela cedência dos dados em análise nesta dissertação. Um agradecimento especial à Filipa Guilherme pela disponibilidade e ajuda constante no estudo das aves.

Ao CMUP, Centro de Matemática da Universidade do Porto, pelo financiamento concedido aos estudos desenvolvidos nesta tese e apresentação em conferências.

À família e aos amigos, pela benevolência por todos os momentos em que não pude estar presente, pelo amparo e constante incentivo.

À avó Mimi, por tudo o que me ensinou, por todo o carinho que ainda me transmite e porque continua a ser uma referência.

Aos meus pais, pelo apoio e compreensão inestimáveis, pelos diversos sacrifícios suportados e pelo carinho inigualável ao longo desta caminhada.

À minha irmã Ana, pelo encorajamento, pela amizade e pela presença assídua com que me presenteou desde sempre.

A todos quantos não citados individualmente, mas que se sabem credores de desmedido reconhecimento, o meu sincero agradecimento.

Resumo

Os modelos de regressão para dados de contagem têm vindo a ser amplamente utilizados nas mais díspares áreas de estudo para a modelação de fenómenos e integram um quadro metodológico especial decorrente do facto de a variável resposta tomar apenas valores inteiros não negativos. A distribuição de Poisson é a mais conhecida e, porventura, a mais utilizada para modelar dados de contagem embora, sempre que existe sobredispersão, o pressuposto de que a média é igual à variância não seja verificado. Nesta situação, torna-se necessário modelar os dados recorrendo à distribuição Quasi-Poisson, caso a média e a variância se relacionem linearmente, ou à distribuição Binomial Negativa, caso a variância apresente uma relação quadrática com a média.

Este trabalho cumpre dois objetivos principais: o estudo de modelos de regressão para dados de contagem e a sua aplicação a dados ornitológicos da cidade do Porto. A estimação da abundância de aves surge na sequência do projeto "Estrutura verde urbana: Estudo da Relação entre a Morfologia do Espaço Público e a Diversidade de Flora e Fauna na cidade do Porto"(http://bio-diver-city.fc.up.pt) desenvolvido pelo CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, da Universidade do Porto. Em particular, analisamos o efeito de diferentes variáveis ambientais sobre a abundância de aves das ordens *Columbiformes* e *Passeriformes*.

Metodologicamente, recorremos a modelos lineares generalizados para a análise da abundância de aves da ordem *Passeriformes* e aos modelos de regressão com zeros inflacionados e com barreira para a análise da abundância de aves da ordem *Columbiformes*, face à presença de um número excessivo de zeros nos dados. No estudo longitudinal, e cingidos apenas às aves da ordem *Passeriformes*, utilizamos modelos lineares generalizados mistos e modelos marginais. Dado tratar-se de dados de contagem, as distribuições de Poisson, Binomial Negativa e Quasi-Poisson foram as distribuições consideradas. A implementação dos modelos foi efetuada com recurso a bibliotecas adequadas no software R versão 2.14.2 (R Development Core Team, 2012).

No estudo dos *Passeriformes*, o modelo que melhor se ajustou aos dados foi o modelo Quasi-Poisson, que revelou que a *área de coberto vegetal* influencia positivamente o número esperado de *Passeriformes*. Na análise das aves da ordem *Columbiformes*, observou-se uma distribuição de frequências com um excesso de zeros e, quer o modelo de regressão de Poisson com zeros inflacionados, quer o modelo de regressão com barreira, apresentaram resultados satisfatórios. Ambos os modelos evidenciaram que o número esperado de *Columbiformes* decresce com o aumento da *área de coberto vegetal*, contrariamente aos jardins com água, e o *odds* para a existência de um número (estritamente) positivo também aumenta caso a *área de coberto fanerófito* aumente. No estudo longitudinal, escolheu-se um modelo de Poisson com um efeito aleatório no *tempo* e efeitos fixos nas variáveis *área de coberto fanerófito* e *presença de água*; no modelo marginal, a escolha recaíu sobre uma matriz de correlação autorregressiva de ordem 1, AR(1), e com variáveis explicativas *tempo* e *presença de água* com uma associação negativa com o número esperado de *Passeriformes*, e *área de coberto fanerófito* com uma associação positiva.

Palavras-chave: MODELO COM BARREIRA, MODELO COM ZEROS INFLACIONADOS, MODELO LINEAR GENERALIZADO MISTO, MODELO MARGINAL, REGRESSÃO BINOMIAL NEGATIVA, REGRESSÃO DE POISSON, REGRESSÃO QUASI-POISSON.

Abstract

Regression models for count data have been widely used in several research fields and they require special methodological frameworks as the response variable only takes non-negative integer values. The Poisson distribution has probably been the most used distribution to model count data although, every time there is overdispersion, the assumption that the average is equal to the variance is not satisfied. In this situation, it is necessary to use the QuasiPoisson distribution, if average and variance are linearly related, or the Negative Binomial distribution, if the variance presents a quadratic relation with the average.
This work has two main objectives: the detailed study of regression models for count data and its application to ornithological data from the city of Porto. Modelling the abundance of birds in Porto arised from the project: "Estrutura verde urbana: Estudo da Relação entre a Morfologia do Espaço Público e a Diversidade de Flora e Fauna na cidade do Porto"(http://bio-diver-city.fc.up.pt) developed by CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto. In particular, we analyze the effect of different environmental variables on the abundance of birds of the *Columbiform* and *Passeriform* orders.
Methodologically, we use generalized linear models for the analysis of the abundance of birds from the *Passeriform* order, and zero-inflated and hurdle models for the analysis of the abundance of birds from the *Columbiform* order, given an excessive number of zeros in the data.
In the longitudinal study, and restricted just to birds of the *Passeriform* order, we apply generalized linear mixed models and marginal models. Since we are dealing with count data, the Poisson, Negative Binomial, and QuasiPoisson distributions are considered. Models' implementation was made using appropriate libraries on the R software version 2.14.2 (R Development Core Team, 2012).
The number of birds from the *Passeriform* order is modelled by the QuasiPoisson distribution and its mean is shown to be significantly and positively associated with the *area of vegetation* of the green spaces. The frequency distribution of birds from the *Columbiform* order exhibit an excess of zeros and, both zero-inflated Poisson and hurdle models presents satisfactory results. The two models show that the expected number of *Columbiform* birds decreases with *absence of water* and increasing *area of vegetation* while the odds to the existence of a (strictly) positive number increases with *area covered with phanerophyte*. For the longitudinal data on birds from the *Passeriform* order, a Poisson model with a random effect for the variable *time* and fixed effects for the variables *area covered with phanerophyte* and *presence of water* are chosen. Regarding marginal models, the choice falls on a type (1)-autoregressive correlation matrix and on *time* and an *indicator of presence of water* as statistically significant explanatory variables.

Keywords: HURDLE MODEL, ZERO INFLATED MODEL, GENARALIZED LINEAR MIXED MODEL, MARGINAL MODEL, NEGATIVE BINOMIAL REGRESSION, POISSON REGRESSION, QUASIPOISSON REGRESSION.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Este trabalho cumpre dois objetivos principais: o estudo de modelos de regressão para dados de contagem e a sua aplicação a dados ornitológicos da cidade do Porto. A estimação da abundância de aves surgiu na sequência do projeto "Estrutura verde urbana: Estudo da Relação entre a Morfologia do Espaço Público e a Diversidade de Flora e Fauna na cidade do Porto"(http://biodiver-city.fc.up.pt) desenvolvido pelo CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, da Universidade do Porto. Em particular, analisamos o efeito de diferentes variáveis ambientais sobre a abundância de aves das ordens[1] *Columbiformes* e *Passeriformes* nos espaços verdes de acesso público da cidade do Porto.

A biodiversidade e a abundância de espécies específicas nos espaços verdes têm vindo a ser alvo de estudo na área da ecologia. Num planeta onde a extinção de espécies é cada vez mais comum e os espaços verdes cada vez mais reduzidos, é imperativo perceber qual o efeito de diversos fatores ambientais sobre a abundância de determinadas espécies num local específico. Dados de abundância são usualmente apresentados como dados de contagem, isto é, como o número de indivíduos da espécie registados num determinado período de tempo.

Modelos de regressão para dados de contagem são amplamente utilizados nas mais diversas áreas de estudo para a modelação de diversos fenómenos e estão enquadrados num quadro metodológico especial decorrente do facto de a variável resposta tomar apenas valores inteiros não negativos. Nesta situação, o uso de modelos lineares clássicos não é, em geral, apropriado, pois os pressupostos do modelo dificilmente serão verificados.

Uma abordagem comum na análise de dados de contagem é a teoria de modelos lineares generalizados (GLM[2]) (Cameron and Trivedi, 1998; Dobson, 2002; McCullagh and Nelder, 1989). Estes modelos foram introduzidos por Nelder and Wedderburn (1972) e representam uma generalização da regressão linear clássica a respostas não-normais. Muito resumidamente, estes modelos apresentam uma estrutura linear nas variáveis explicativas, assumem que a resposta segue uma distribuição que pertence à família exponencial e modelam, não necessariamente a média da resposta, mas uma transformação desta última. No caso dos modelos de regressão para dados de contagem, o número esperado de ocorrências de um acontecimento num certo intervalo de tempo, ou numa determinada área é modelado em função de um conjunto de variáveis explicativas. A distribuição mais usada é a distribuição de Poisson, sendo o modelo de regressão de Poisson um caso particular dos GLM (Cameron and Trivedi, 1998). Este modelo pressupõe que a média e a variância da resposta, condicionadas pelos valores das variáveis explicativas, são iguais. Na prática, a variância tende a exceder a média e esta situação de sobredispersão condiciona a aplicação dos modelos de Poisson. Uma solução passa por considerar o modelo de regressão Quasi-

[1] Ordem - classificação taxonómica das espécies de aves (Catry et al., 2010).

[2] **G**eneralized **L**inear **M**odels.

Poisson ou o modelo de regressão Binomial Negativa, consoante a estrutura da variância (Hilbe, 2011; Hoef and Boveng, 2007). Enquanto que no modelo Quasi-Poisson a variável resposta segue condicionalmente uma distribuição de Poisson e a variância estabelece uma relação linear com a média, no modelo de regressão Binomial Negativa, a variável resposta segue condicionalmente uma distribuição Binomial Negativa, sendo que a variância tem, portanto, uma relação quadrática com a média. Apesar de o modelo Quasi-Poisson poder ser interpretado como um GLM, não existe nenhuma distribuição pertencente à família exponencial que tenha a média e a variância relacionadas linearmente (Turkman and Silva, 2000).
Sendo numerosas as aplicações de modelos lineares generalizados com as distribuições acima referidas em áreas como a economia, a tecnologia, a medicina, a farmácia, entre outras, destaque-se a área da ecologia, onde os modelos de regressão para dados de contagem têm sido cada vez mais usados: na estimação da abundância de animais (Hoef and Boveng, 2007; Pope et al., 2000; Seavy et al., 2005; White and Bennetts, 1996); no impacto da poluição atmosférica na saúde pública (McNeney and Petkau, 1994; Tadano et al., 2009) e no estudo do efeito das caraterísticas da paisagem e dos cervos no número de colisões entre veículos e cervos (Noro et al., 2010), por exemplo.
Uma outra caraterística de dados de contagem, principalmente dados ecológicos, é a tendência para conterem muitos valores exatamente iguais a zero. Quando este número de zeros é tão elevado que os dados não podem ser ajustados a distribuições standard (Poisson, Binomial Negativa,...), para a sua modelação terão de ser considerados modelos de regressão com um número excessivo de zeros; por exemplo, modelos de zeros inflacionados e modelos com barreira. Uma descrição teórica destes modelos pode ser encontrada em Cameron and Trivedi (1998), Hilbe (2011), Lambert (1992), Ridout et al. (1998) e Zuur et al. (2009). Os modelos em questão diferem na forma como lidam com os zeros. Nos modelos de zeros inflacionados, as contagens nulas podem ser originárias de um processo de contagem, sendo designadas por zeros verdadeiros, ou de uma massa pontual em zero, sendo designadas por zeros falsos. Sob o ponto de vista prático, os zeros verdadeiros correspondem aos indivíduos que, de facto, não se encontram presentes quando o processo de amostragem é efetuado, enquanto que os zeros falsos dizem respeito aos indivíduos que não foram observados por problemas no processo de amostragem. No caso dos modelos com barreira, o processo de contagem é truncado em zero e, portanto, não pode produzir contagens nulas. Este mesmo modelo apenas separa as contagens em zeros (ausências) *versus* não zeros (presenças). Nas últimas duas décadas, modelos com um número excessivo de zeros têm sido bastante requisitados em diversos estudos: número de defeitos num processo de fabricação (Lambert, 1992); número de crianças com cáries (Böhning et al., 1997); dados de violência doméstica (Famoye and Singh, 2006); análise do desenvolvimento motor de uma criança (Cheung, 2002), entre outros. Mais recentemente têm vindo a ser utilizados em estudos de ecologia, nomeadamente, quando se pretende apurar qual a abundância (presença/ausência) de um animal num determinado local e num período de tempo definido. Exemplos destes estudos foram publicados por Joseph et al. (2009), Kéry (2008), Martin et al. (2005), Potts and Elith (2006), Royle et al. (2005) e Wenger and Freeman (2008).

Em dados de contagem, é usual a mesma unidade experimental ser observada ao longo do tempo (dados longitudinais), pelo que se espera que haja uma correlação não nula entre as várias observações das unidades individuais, levando à violação do pressuposto de independência entre todas as observações. Uma forma de se ter em consideração os dados correlacionados é modelar explicitamente a estrutura de correlação, utilizando-se o modelo marginal (Cabral and Gonçalves, 2011; Diggle et al., 2002; Fitzmaurice et al., 2004). Este modelo foi proposto por Liang and Zeger (1986) e recorre ao método de equações de estimação generalizadas para estimar os parâmetros do modelo. Outra alternativa para modelar dados longitudinais passa por considerar modelos

lineares generalizados mistos (GLMM[3]). Estes modelos são uma extensão dos GLM dado que permitem a inclusão de efeitos aleatórios no preditor linear para que se tenha em conta a correlação existente entre as várias observações da mesma unidade experimental (Breslow and Clayton, 1993; Cabral and Gonçalves, 2011; Diggle et al., 2002; Fitzmaurice et al., 2004; Hedeker, 2005). Estes efeitos aleatórios seguem uma distribuição normal multivariada com valor esperado igual a zero e matriz de variância-covariância definida. A escolha entre os dois modelos para a análise de dados longitudinais recai sobre o objetivo da inferência que se pretende efetuar. Caso se pretenda inferir sobre a população, deve-se usar o modelo marginal. Caso se pretenda inferir sobre o indivíduo (a unidade experimental mais usual), o modelo escolhido deve ser o modelo linear generalizado misto. Exemplos de aplicação destes modelos são os estudos de: Kleinschmidt et al. (2001) sobre a taxa de incidência de malária na África do sul, no período de um ano; Vahtera et al. (2006) sobre a eventual discrepância entre sexos na vulnerabilidade após a morte ou doença de um familiar, com amostras recolhidas durante 6 anos e Edwards et al. (2010) e Tvardíková (2010) sobre a abundância de aves.

Existe, de facto, uma grande diversidade de modelos de regressão que se ajustam adequadamente a dados de contagem. A implementação de todos os modelos supra citados foi efetuada recorrendo a bibliotecas adequadas no software R versão 2.14.2 (R Development Core Team, 2012). Ao longo desta tese, sempre que uma biblioteca tenha sido usada, o seu nome será explicitamente mencionado.

1.1 Estrutura da tese

Este trabalho de investigação encontra-se segmentado em diversos capítulos, sendo que no início de cada capítulo é apresentada uma breve introdução ao tema a tratar e uma descrição do conteúdo do capítulo. Esta organização permite ao leitor ter uma ideia geral da investigação, remetendo os detalhes do estudo para o corpo integral do capítulo.

A introdução e os objetivos desta dissertação são apresentados no **capítulo 1**. É efetuada uma contextualização de caráter introdutório relativamente aos modelos de regressão para dados de contagem e abundância de espécies.

O **capítulo 2** inclui a base teórica relativa aos modelos lineares generalizados, que sustenta todas as metodologias abordadas nesta tese. Os modelos de regressão mais comuns na modelação de dados de contagem são analisados detalhadamente neste capítulo.

Modelos de regressão com um número excessivo de zeros são apresentados no **capítulo 3**. As principais diferenças na modelação entre modelos de regressão com zeros inflacionados e modelos com barreira encontram-se aqui descritas pormenorizadamente.

No **capítulo 4** é apresentada a parte do Projeto do CIBIO sobre a qual a aplicação desta dissertação incide. Inclui-se também uma revisão metodológica acerca dos métodos de censos e abundância de aves.

A análise crítica dos resultados obtidos para a primeira parte do projeto do CIBIO é apresentada no **capítulo 5**. Inicialmente, são descritos a base de dados e o modo como a preparação da análise foi efetuada. Posteriormente, é feita uma análise exploratória aos dados e uma implementação dos

[3] **G**eneralized **L**inear **M**ixed **M**odels.

modelos considerados no capítulo 2 e 3. Os resultados obtidos são interpretados sob o ponto de vista biológico e comparados com estudos de referência na literatura.

A metodologia referente à modelação de dados longitudinais assim como a aplicação aos dados do Projeto do CIBIO são explicitados no **capítulo 6**. Começamos por explanar teoricamente o modelo linear generalizado misto e o modelo marginal e seguimos depois para uma análise exploratória aos dados em estudo e ao ajustamento dos modelos longitudinais, acompanhados ainda de uma análise contextualizada dos resultados obtidos.

No **capítulo 7** são referidas as principais conclusões, as limitações inerentes ao estudo, e algumas propostas de trabalho futuro.

Capítulo 2

Modelos Lineares Generalizados

Os modelos são representações abstratas e simplificadas da realidade. Matematicamente, os modelos podem ser *determinísticos*, situação em que os resultados estão definidos com precisão, ou *probabilísticos*, onde os resultados são variáveis, com variabilidade devida a fatores aleatórios desconhecidos. Um modelo *estatístico* é um modelo que contém uma componente probabilística.
No quotidiano surgem diversos problemas cujo objetivo é analisar a influência que uma ou mais variáveis (*explicativas*), medidas em unidades experimentais como indivíduos, animais ou objetos, têm sobre uma variável de interesse a que damos o nome de *variável resposta*. De um modo geral, este problema é abordado através de modelos de regressão, que relacionam a variável de interesse com as variáveis explicativas (Turkman and Silva, 2000).
O modelo estatístico clássico para a análise de regressão, o *modelo linear*, surgiu no início do século XIX, através de Legendre e Gauss. Muito resumidamente, este modelo exprime a média da resposta como uma combinação linear das variáveis explicativas, e é aplicado na situação em que a resposta segue uma distribuição normal. Para dar resposta a situações que não se encaixavam no contexto dos modelos lineares, foram criados vários modelos para respostas não-normais - os *modelos lineares generalizados*.
Estes modelos foram introduzidos por Nelder and Wedderburn (1972). Estes modelos apresentam uma estrutura linear nas variáveis explicativas, assumem que a variável resposta segue uma distribuição que pertence à família exponencial, e relacionam, eventualmente de forma não linear, a média da resposta com a estrutura linear das variáveis explicativas. A estimação destes modelos é baseada no método da máxima verosimilhança, onde a maximização da função de verosimilhança é obtida através de métodos numéricos iterativos.
Neste capítulo será exposta toda a fundamentação teórica acerca destes modelos. Os modelos de regressão Logística, Poisson, Binomial Negativa e Quasi-Poisson são exemplos de modelos lineares generalizados que serão abordados com detalhe nas secções seguintes.

2.1 Família Exponencial

Seja Y uma variável aleatória. A distribuição de probabilidade de Y pertence à família exponencial com parâmetro de dispersão ϕ se a sua função de densidade de probabilidade é da forma

$$f(y,\theta,\phi) = exp\left\{\sum_{k=1}^{q} \frac{\theta_k T_k(y) - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

onde

- $\theta = (\theta_1, ..., \theta_q)$ são parâmetros de localização, designados por parâmetros canónicos
- ϕ é o parâmetro de escala ou parâmetro de dispersão
- as funções $T_1, ..., T_q$ dependem apenas da variável aleatória Y
- a função b depende apenas dos parâmetros θ
- a função a depende apenas do parâmetro de dispersão ϕ
- a função c depende apenas da variável aleatória Y e do parâmetro de dispersão ϕ.

Se considerarmos $a(\phi) = \dfrac{\phi}{\omega}$, onde ω é um peso conhecido para as observações de Y (este peso pode variar conforme as observações, usualmente toma o valor 1), $q = 1$ e $T(y) = y$, a equação (2.1) fica reduzida a

$$f(y, \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{2.2}$$

e neste caso diz-se que a distribuição de Y está na *forma canónica*.
Daqui em diante iremos assumir que todas as observações são provenientes de uma distribuição da família exponencial com função densidade de probabilidade que se encontra na forma canónica.
McCullagh and Nelder (1989) mostram que, se Y é uma variável aleatória com uma distribuição que pertence à família exponencial na forma canónica, então

$$E(Y) = \mu = b'(\theta) \tag{2.3}$$

$$Var(Y) = \sigma^2 = b''(\theta)a(\phi) = \phi b''(\theta)/\omega \tag{2.4}$$

onde $b'(\theta)$ e $b''(\theta)$ são a primeira e a segunda derivadas de $b(\theta)$, respetivamente. O termo $b''(\theta)$ pode também ser interpretado como a variância de μ, $Var(\mu)$, (Turkman and Silva, 2000).

Na secção que se segue apresentamos casos particulares de distribuições pertencentes à família exponencial que serão futuramente utilizadas nesta tese.

2.1.1 Exemplos

Em todos os exemplos aqui apresentados assume-se que o valor de ω é igual a 1.

2.1.1.1 Distribuição Normal

Se Y segue uma distribuição Normal com valor esperado μ e variância σ^2, $Y \sim N(\mu, \sigma^2)$, a função densidade de probabilidade de Y é dada por

$$\begin{aligned} f(y|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{\mu y - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]. \end{aligned} \tag{2.5}$$

para $y \in \mathbb{R}$. Esta função é do tipo (2.2) com

$$\begin{aligned}\theta &= \mu, \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2} \\ a(\phi) &= \sigma^2, \quad c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right) \\ E(Y) &= b'(\theta) = \theta = \mu, \quad Var(Y) = a(\phi)b''(\theta) = \sigma^2\end{aligned}$$

Repare-se que o parâmetro canónico é a média μ e que o parâmetro de dispersão é a variância, $\phi = \sigma^2$.

2.1.1.2 Distribuição de Poisson

Caso a variável aleatória Y siga uma distribuição de Poisson com média μ, $P(\mu)$, a sua função de probabilidade traduz-se por

$$\begin{aligned}f(y|\mu) &= \mu^y \frac{e^{-\mu}}{y!} \\ &= \exp\left(\log(\mu)y - \mu - \log(y!)\right)\end{aligned} \tag{2.6}$$

Aqui,

$$\begin{aligned}\theta &= \log(\mu), \quad b(\theta) = \mu = exp(\theta) \\ a(\phi) &= 1, \quad c(y,\phi) = -\log(y!) \\ E(Y) &= b'(\theta) = \exp(\theta) = \mu, \quad Var(Y) = a(\phi)b''(\theta) = \exp(\theta) = \mu\end{aligned}$$

A distribuição não apresenta parâmetro de dispersão ϕ e o parâmetro canónico é $\log(\mu)$.

2.1.1.3 Distribuição Binomial

Considere-se que $Y \sim B(n,\pi)$, onde n é o número de experiências de Bernoulli [1] de um determinado acontecimento e π é a probabilidade de sucesso desse mesmo acontecimento em cada experiência. Tem-se que Y é uma variável aleatória que conta o número de sucessos em n experiências. A função de probabilidade de Y é dada por

$$\begin{aligned}f(y\,|\,n,\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= \exp\left[y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}\right]\end{aligned} \tag{2.7}$$

Neste caso,

$$\begin{aligned}\theta &= \log\left(\frac{\pi}{1-\pi}\right), \quad b(\theta) = -n\log(1-\pi) = n\log(1+\exp(\theta)) \\ a(\phi) &= 1, \quad c(y,\phi) = \log\binom{n}{y} \\ E(Y) &= b'(\theta) = n\pi, \quad Var(Y) = a(\phi)b''(\theta) = n\pi(1-\pi)\end{aligned}$$

Não existe parâmetro de dispersão e o parâmetro canónico é $\log\left(\frac{\pi}{1-\pi}\right)$, também designado por $logit(\pi)$.

[1] As experiências têm resultados independentes e identicamente distribuídos.

2.1.1.4 Distribuição Binomial Negativa

A distribuição Binomial Negativa pode ser obtida de 12 maneiras diferentes (Hilbe, 2011). Neste trabalho iremos abordar duas dessas maneiras, reportando-nos à distribuição Binomial Negativa canónica (NB-C) e ao modelo tradicional de regressão binomial negativa (NB2). Mostraremos que ambas as formas podem ser obtidas da função de probabilidade da Binomial Negativa.
Seja Y uma variável aleatória que segue uma distribuição Binomial Negativa com parâmetros k e p,

$$Y \sim BN(k, p).$$

A variável aleatória Y representa o número de insucessos anteriores a k sucessos, num conjunto de acontecimentos independentes e com a mesma probabilidade de sucesso, p. A função de probabilidade de Y é dada por

$$\begin{aligned} f(y|p,k) &= \binom{y+k-1}{k-1} p^k (1-p)^y \qquad (2.8) \\ &= \exp\left[y\log(1-p) + k\log(p) + \log\left(\binom{y+k-1}{k-1} \right) \right] \end{aligned}$$

Nesta situação, a distribuição binomial negativa está representada na forma canónica (NB-C) onde

$$\begin{aligned} \theta &= \log(1-p), \quad b(\theta) = -k\log(p) \\ a(\phi) &= 1, \quad c(y,\phi) = \log\left(\binom{y+k-1}{k-1} \right) \end{aligned}$$

A média e a variância de Y são expressas, respetivamente, por

$$E(Y) = b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = \frac{k(1-p)}{p} \qquad (2.9)$$

$$\begin{aligned} Var(Y) &= b''(\theta) = \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} \qquad (2.10) \\ &= \frac{k}{p^2}(1-p)^2 + \frac{-k}{p}(-(1-p)) \\ &= \frac{k(1-p)}{p^2}. \end{aligned}$$

Portanto não existe parâmetro de dispersão e o parâmetro canónico é dado por $\log(1-p)$.

Resta-nos abordar a descrição da distribuição Binomial Negativa NB2. Para tal, descrevemos primeiro a distribuição de Poisson mista.
A função de probabilidade de uma distribuição de Poisson usual, com média μ, é dada por

$$f(y) = P\{Y = y\} = \frac{e^{-\mu}\mu^y}{y!}$$

para $\mu > 0$ e onde $y = 0, 1, 2, 3, ...$
Podemos também considerar a situação em que a média é ela própria uma variável aleatória, digamos $Y \sim P(\mu V)$ onde V é uma variável aleatória que representa a heterogeneidade não observada, com distribuição conhecida, U, e satisfazendo $E(V) = 1$ (Lawless, 1987). A distribuição

de U é designada por *distribuição de estrutura* [2] (Grandell, 1997).
Em Grandell (1997), mostra-se que:

$$\begin{aligned} E(Y) &= E(\mu V) = \mu E(V) = \mu \\ &\text{e} \\ Var(Y) &= E(\mu V) + Var(\mu V) = \mu + \mu^2 Var(V) = \mu + \alpha\mu^2, \end{aligned}$$

onde $\alpha = Var(V)$. Em particular, esta distribuição permite a ocorrência de sobredispersão.
A distribuição não condicionada de Y é conhecida por *distribuição de Poisson mista* e tem função de probabilidade dada por

$$P\{Y = y\} = \int_0^{+\infty} \frac{e^{-(\mu v)}(\mu v)^y}{y!} u(v)\partial v \tag{2.11}$$

onde u é a função densidade de probabilidade de V.
Para qualquer observação v de V, tem-se que

$$Y|v \sim P(\mu v)$$

e recuperamos o caso da distribuição de Poisson usual (Lawless, 1987).
Uma distribuição binomial negativa pode ser encarada como um caso particular de uma distribuição de Poisson Mista quando se considera que V segue a distribuição $\Gamma\left(\frac{1}{Var(V)}, \frac{1}{Var(V)}\right) = \Gamma\left(\frac{1}{\alpha}, \frac{1}{\alpha}\right)$. Nesta situação, a distribuição *não condicionada* de Y é $Y \sim BN(\mu, \alpha)$ (Agresti, 2007; Lawless, 1987; Rodríguez, 2007).
A distribuição de Poisson obtém-se como uma distribuição limite da distribuição Binomial Negativa, quando $\alpha = 0$.
Substituindo g em (2.11) pela expressão da função densidade de probabilidade de uma distribuição $\Gamma(\frac{1}{\alpha}, \frac{1}{\alpha})$ e fazendo manipulações algébricas (Hilbe, 2011) obtém-se a expressão da função de probabilidade correspondente à distribuição $BN(\mu, \alpha)$ (NB2),

$$\begin{aligned} f(y|\mu, \alpha) &= \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \\ &= \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i}. \end{aligned} \tag{2.12}$$

Fazendo

$$k = \frac{1}{\alpha} \qquad \text{e} \qquad p = \frac{1}{1 + \alpha\mu} \tag{2.13}$$

obtém-se

$$f(y|p, k) = \binom{y + k - 1}{k - 1} p^k (1 - p)^y,$$

ou seja, a função de probabilidade apresentada em (2.8). Deste modo, conclui-se que as expressões da função de probabilidade da NB-C e da NB2 são equivalentes. A função de probabilidade da Binomial Negativa pode assim ser expressa de diversas formas, dependendo da parametrização

[2] Structure distribution.

escolhida (Hilbe, 2011).
Como Hilbe (2011) refere, o modelo NB2 é criado convertendo a função de ligação canónica e a inversa desta nos modelos lineares generalizados obtidos pela NB-C para a forma log (considera o log como função de ligação).
Assim, em termos de μ e α temos

$$\begin{aligned} \theta &= \log\left(\frac{\alpha\mu}{1+\alpha\mu}\right), \quad b(\theta) = 1/\alpha \log(1+\alpha\mu) \\ E(Y) &= b'(\theta) = \mu \\ Var(Y) &= b''(\theta) = \mu(1+\alpha\mu). \end{aligned}$$

2.2 Componentes de um modelo linear generalizado

Os modelos lineares generalizados apresentam três componentes distintas: (i) componente aleatória - identifica a variável aleatória resposta Y e especifica uma distribuição para Y pertencente à família exponencial; (ii) componente sistemática - especifica as variáveis explicativas do modelo, também conhecidas como covariáveis ou variáveis independentes, e considera uma combinação linear dessas variáveis; (iii) função de ligação - estabelece a ligação entre as componentes aleatória e sistemática (Agresti, 2007; McCullagh and Nelder, 1989).

2.2.1 Componente aleatória

Designe-se por $\mathtt{X} = (X_1, ..., X_p)$ o vetor das covariáveis de interesse. Para uma amostra aleatória de tamanho n, designamos por $\mathtt{x}_i^T = (x_{i1}, ...x_{ip})$ a observação do indivíduo i. A componente aleatória de um modelo linear generalizado refere que a distribuição de Y condicionada por $\mathtt{X}$ pertence à família exponencial e que portanto satisfaz

$$E(Y_i|\mathtt{x_i}) = \mu_i = b'(\theta_i), \quad i = 1, ..., n.$$

2.2.2 Componente sistemática

As covariáveis $X_1, ..., X_p$ produzem uma estrutura linear η com caráter preditivo dada por

$$\eta = \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \tag{2.14}$$

onde $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ é o vetor que consiste dos coeficientes de regressão, normalmente desconhecidos.
A componente sistemática do modelo pode ser escrita, de modo equivalente, na forma:

$$\eta_i = \tilde{\mathtt{x}}_i^T \beta, \quad i = 1, ..., n,$$

onde β é o vetor dos coeficientes de regressão e $\tilde{\mathtt{x}}_i = (1, \mathtt{x}_i^T)^T$.
Em notação matricial,

$$\eta = X\beta,$$

onde X é matriz de especificação de dimensão $n \times (p+1)$, isto é, é a matriz cuja 1ª coluna consiste apenas de 1's e cujas restantes colunas consistem dos vetores coluna $\mathtt{x}_i^T$ e β é o vetor dos parâmetros de regressão de dimensão (p+1).

2.2.3 Função de ligação

A função de ligação de um modelo linear generalizado é uma função g, monótona e diferenciável, que relaciona a média da resposta, μ, com o preditor linear η

$$\eta = g(\mu).$$

A função de ligação $g(.)$ estabelece assim a ligação entre a componente aleatória e a componente sistemática do modelo. Quando a função de ligação torna o preditor linear η igual ao parâmetro canónico θ da família exponencial, diz-se que a função de ligação é a *função de ligação canónica*. O modelo de regressão linear é um caso particular dos modelos lineares generalizados, em que resposta segue uma distribuição normal e a função de ligação é a identidade.

2.3 Inferência

Partindo do pressuposto de que o modelo criado é adequado e está bem formulado, é necessário proceder à realização de inferências sobre esse mesmo modelo. Nos modelos lineares generalizados, os métodos de inferência estatística baseiam-se, essencialmente, na função de verosimilhança.
Em certas situações, o pressuposto de que o modelo é adequado e está bem formulado não é realista. Isto acontece, por exemplo, quando existe sobredispersão e portanto, é necessário alterar a variância da resposta através da introdução de um parâmetro de dispersão. Como refere Turkman and Silva (2000), neste contexto, não é possível especificar completamente o modelo dado que não existe nenhuma distribuição pertencente à família exponencial que apresente aqueles valores de média e variância. A solução para este problema passa por considerar a função de quasi-verosimilhança para realizar as inferências necessárias. A função de quasi−verosimilhança será também abordada no final desta secção.
Para a simplificação dos cálculos seguintes, assumamos, desde já, que a matriz de especificação X é de caraterística completa.

2.3.1 Log−verosimilhança, função score e matriz de informação de Fisher

A presente secção descreve a função log−verosimilhança, a *função score* e a matriz de informação de Fisher, baseada, essencialmente no capítulo 2 de Fahrmeir and Tutz (2001) e no capítulo 2 de Turkman and Silva (2000).
Os coeficientes de regressão de um modelo linear generalizado são estimados através do método da máxima verosimilhança.
Considerando a definição de modelo linear generalizado apresentado na secção 2.2, a função de verosimilhança do modelo, em função de β, é dada por (Turkman and Silva, 2000),

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f(y_i, \theta_i, \phi) \\ &= \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \\ &= \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)\right\}. \end{aligned} \tag{2.15}$$

O logaritmo da função de verosimilhança é assim traduzido por

$$\begin{aligned} \log(L(\beta)) = l(\beta) &= \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \\ &= \sum_{i=1}^{n} l_i(\beta), \end{aligned} \tag{2.16}$$

onde l_i é a contribuição de cada observação y_i para a verosimilhança (Fahrmeir and Tutz, 2001; McCullagh and Nelder, 1989; Turkman and Silva, 2000).
O estimador de máxima verosimilhança de β corresponde à solução do sistema de equações de verosimilhança:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta_j} = 0, \quad j = 0, 1, ..., p. \tag{2.17}$$

Para cada unidade i de observação, tem-se (McCullagh and Nelder, 1989; Turkman and Silva, 2000)

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\partial l_i(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i(\beta)}{\partial \beta_j}, \quad j = 0, ..., p. \tag{2.18}$$

Tendo em conta a função log $-$verosimilhança, equação (2.16), e sabendo que $b'(\theta_i) = \mu_i$ (ver secção 2.1), deduz-se que

$$\frac{\partial l_i(\theta_i)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}.$$

Na secção 2.1, vimos também que $Var(Y_i) = a(\phi)b''(\theta_i)$ e portanto

$$b''(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i} = \frac{Var(Y_i)}{a(\phi)}.$$

Por fim, sabemos que $\eta_i = \tilde{\mathbf{x}}_i^T \beta$ (secção 2.2), pelo que se conclui que

$$\frac{\partial \eta_i(\beta)}{\beta_j} = \tilde{x}_{ij}.$$

Reescrevendo a expressão (2.18) , tem-se que

$$\begin{aligned} \frac{\partial l_i(\beta)}{\partial \beta_j} &= \frac{(y_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \tilde{x}_{ij} \\ &= \frac{(y_i - \mu_i)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \tilde{x}_{ij} \end{aligned} \tag{2.19}$$

e portanto as equações de verosimilhança para β são

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \tilde{x}_{ij} = 0, \quad j = 0, 1, ..., p. \tag{2.20}$$

A (primeira) derivada da função log $-$verosimilhança em ordem a β é denominada por *função score* e é dada por

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta), \tag{2.21}$$

onde $s_i(\beta)$ é a contribuição individual da função score, isto é, é o vetor de componentes $\frac{\partial l_i(\beta)}{\partial \beta_j}$ acima descrito.
A matriz de covariância da *função score*,

$$cov(s(\beta)) = I(\beta) = E\left[-\frac{\partial s(\beta)}{\partial \beta}\right] \tag{2.22}$$

é designada por *matriz de informação de Fisher*.
A matriz de informação de Fisher corresponde ao simétrico do valor esperado da matriz Hessiana da função log −verosimilhança:

$$I(\beta) = -E\left(\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}\right) \quad = \quad E\left(\frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_k}\right). \tag{2.23}$$

O facto de a igualdade apresentada em (2.23) não ser trivial leva-nos a explicitar a sua prova de seguida.
Consideremos que a função verosimilhança é uma função tal que $L(\beta) = f(\mathtt{x};\beta)$ e que $E_\beta(c) = \int cf(\mathtt{x};\beta) \quad d\mu(x)$.

Como referido acima, por definição,

$$I(\beta) \quad = \quad E\left[\left(\frac{\partial}{\partial \beta}\log f(X;\beta)\right)^2\right] = E\left[\left(\frac{\partial f}{\partial \beta}\frac{1}{f}\right)^2\right]. \tag{2.24}$$

Proposição: Sob condições gerais de regularidade, tem-se (Lehmann and Casella, 1998)

a)

$$E\left(\frac{\partial}{\partial \beta}\log f_\beta(X)\right) = 0 \tag{2.25}$$
$$I(\beta) = Var\left(\frac{\partial^2}{\partial \beta^2}\log f_\beta(X)\right)$$

b)

$$I(\beta) = -E\left(\frac{\partial^2}{\partial \beta^2}\log f_\beta(X)\right). \tag{2.26}$$

Demonstração:

b)

$$\begin{aligned} \frac{\partial^2}{\partial \beta^2}\log f_\beta(x) &= \frac{\partial}{\partial \beta}\left(\frac{\frac{\partial f_\beta}{\partial \beta}}{f_\beta}\right) \\ &= \left(\frac{\partial^2}{\partial \beta^2} f_\beta(x)\right)\frac{1}{f_{\beta(x)}} - \left(\frac{\frac{\partial}{\partial \beta} f_\beta(x)}{f_\beta(x)}\right)^2 \end{aligned} \tag{2.27}$$

Tomando o valor esperado,

$$E\left(\frac{\partial^2}{\partial\beta^2}\log f_\beta(x)\right) = E\left(\left(\frac{\partial^2}{\partial\beta^2}f_\beta(x)\right)\frac{1}{f_\beta(x)}\right) - E\left(\left(\frac{\frac{\partial}{\partial\beta}f_\beta(x)}{f_\beta(x)}\right)^2\right). \tag{2.28}$$

Pela definição (2.24), sabe-se que $-E\left(\left(\frac{\frac{\partial}{\partial\beta}f_\beta(x)}{f_\beta(x)}\right)^2\right) = I(\beta)$. Resta-nos provar que $E\left(\left(\frac{\partial^2}{\partial\beta^2}f_\beta(x)\right)\frac{1}{f_\beta(x)}\right)$ é zero.

Assumindo que $\int f_\beta(x)d\mu(x) = 1$, tem-se que

$$\begin{aligned} E\left(\left(\frac{\partial^2}{\partial\beta^2}f_\beta(x)\right)\frac{1}{f_\beta(x)}\right) &= \int\left(\frac{\partial^2}{\partial\beta^2}f_\beta(x)\right)\frac{1}{f_\beta(x)}f_\beta(x)d\mu(x) \\ &= \frac{\partial^2}{\partial\beta^2}\int f_\beta(x)d\mu(x) \\ &= 0, \end{aligned} \tag{2.29}$$

como queríamos demonstrar.

Deste modo, a matriz de informação de Fisher é definida por

$$\begin{aligned} -E\left(\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) &= E\left(\frac{\partial l_i}{\partial\beta_j}\frac{\partial l_i}{\partial\beta_k}\right) \\ &= E\left[\left(\frac{(y_i-\mu_i)\tilde{x}_{ij}}{Var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right)\left(\frac{(y_i-\mu_i)\tilde{x}_{ik}}{Var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right)\right] \\ &= E\left[\frac{(y_i-\mu_i)^2\tilde{x}_{ij}\tilde{x}_{ik}}{(Var(Y_i))^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right] \\ &= \frac{\tilde{x}_{ij}\tilde{x}_{ik}}{Var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2, \end{aligned}$$

pelo que se conclui que o elemento genérico de ordem (j,k) da matriz de informação de Fisher é expresso por

$$I(\beta)_{j,k} = -\sum_{i=1}^{n} E\left(\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n}\frac{\tilde{x}_{ij}\tilde{x}_{ik}}{Var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{2.30}$$

Na forma matricial pode-se escrever

$$I(\beta) = X^T W X$$

onde W é a matriz diagonal de ordem n cujo i-ésimo elemento é

$$\begin{aligned} \varpi_i = \frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{Var(Y_i)} &= \frac{\omega_i\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{\phi b''(\theta)} \\ &= \frac{\omega_i}{\phi b''(\theta)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2}. \end{aligned} \tag{2.31}$$

2.3.2 Estimação dos parâmetros do modelo

O estimador de máxima verosimilhança de β é obtido através das equações de verosimilhança definidas em (2.20). Tal como referem Fahrmeir and Tutz (2001), este estimador não corresponde necessariamente a um máximo global da função log −verosimilhança $l(\beta)$. Contudo, em muitos modelos a função $l(\beta)$ é estritamente côncava e, por conseguinte, existe um único máximo local que coincide naturalmente com o máximo global. Nesta situação, o estimador de máxima verosimilhança é único. As equações de verosimilhança são, de um modo geral, não lineares o que torna necessária a utilização de métodos iterativos para a sua resolução. Os métodos iterativos mais usados na estimação numérica das equações de verosimilhança são o método iterativo dos mínimos quadrados reponderados[3] e o método de Newton-Raphson.

2.3.2.1 Método iterativo dos mínimos quadrados reponderados

Nesta secção pretendemos mostrar que o estimador de máxima verosimilhança de β pode ser obtido através do método iterativo dos mínimos quadrados reponderados e que esse estimador é equivalente ao obtido pelo algoritmo de Fisher.
Seja $s(\beta)$ o vetor de *scores* definido em (2.21). Dada uma estimativa arbitrária do parâmetro $\beta^{(0)}$, o método de *scores* de Fisher define iterações sucessivas através da equação:

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + F(\widehat{\beta}^{(k)})^{-1} s(\widehat{\beta}^{(k)}), \quad k = 0, 1, 2, ... \tag{2.32}$$

onde a inversa da matriz de informação de Fisher (que se supõe existir),F^{-1}, e o vetor de scores, $s(.)$, são calculados em $\widehat{\beta}^{(k)}$.

Posteriormente, é necessário calcular o preditor linear estimado $g(\widehat{\mu}_i) = \tilde{\mathbf{x}}_i^T \widehat{\beta} = \widehat{\eta}_i$, que irá ser usado para obter os valores ajustados $\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i)$. Utilizando os valores estimados anteriormente define-se uma variável dependente de trabalho T [4]; para o indivíduo i e na iteração k, a variável T vale (Rodríguez, 2007):

$$\begin{aligned} t_i^{(k)} &= \widehat{\eta}_i^{(k)} + (y_i - \widehat{\mu}_i^{(k)}) \frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}} \\ &= g(\widehat{\mu}_i^{(k)}) + (y_i - \widehat{\mu}_i^{(k)}) g'(\widehat{\mu}_i^{(k)}), \end{aligned} \tag{2.33}$$

onde o termo $\dfrac{\partial \eta_i}{\partial \mu_i}$ representa a derivada da função de ligação em ordem a μ_i; de facto,

$$\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i) \Longleftrightarrow \widehat{\eta}_i = g(\widehat{\mu}_i) \Longleftrightarrow \frac{\partial \widehat{\eta}_i}{\partial \mu_i} = g'(\widehat{\mu}_i).$$

De seguida, são calculados os pesos iterativos:

$$\varpi_i^{(k)} = \frac{1}{Var(y_i) \left(\dfrac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}} \right)^2}. \tag{2.34}$$

[3] Turkman and Silva (2000) referem-se a este método como método iterativo de mínimos quadrados ponderados.
[4] Working dependent variable.

No caso da família exponencial e considerando a transformação usual, $a_i(\phi) = \phi/\omega_i$, obtém-se:

$$\varpi_i^{(k)} = \frac{\omega_i^{(k)}}{\phi b''(\theta_i)\left(\frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}}\right)^2}. \tag{2.35}$$

Este peso é inversamente proporcional à variância da variável t_i. Mais precisamente, $\varpi_i = \frac{1}{Var(t_i)}$. De facto, a variância da variável t_i é

$$Var(t_i) = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 Var(Y_i - \widehat{\mu}_i) = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 Var(Y_i).$$

Partindo da equação (2.35) e fazendo pequenas manipulações algébricas, tem-se:

$$\varpi_i = \frac{\omega_i}{\phi b''(\theta_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2} = \frac{1}{a_i(\phi) b''(\theta_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2} = \frac{1}{Var(Y_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2} = \frac{1}{Var(t_i)},$$

Uma melhor estimativa de β é obtida tendo em conta a variável dependente de trabalho e os pesos obtidos. Deste modo, o estimador dos mínimos quadrados reponderados é traduzido por (Rodríguez, 2007)

$$\widehat{\beta}^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} T^{(k)}, \tag{2.36}$$

onde X representa a matriz do modelo em questão, W é a matriz diagonal dos pesos, com entradas ϖ_i dadas pela equação (2.35) e T é o vetor com entradas t_i dadas pela equação (2.33).
O procedimento repete-se até que as estimativas se alterem a menos de uma constante pré-especificada. Um critério de paragem é, por exemplo (Fahrmeir and Tutz, 2001),

$$\frac{\| \widehat{\beta}^{(k+1)} - \widehat{\beta}^{(k)} \|}{\| \widehat{\beta}^{(k)} \|} \leq \varepsilon, \tag{2.37}$$

para um valor de $\varepsilon > 0$ previamente definido.
De forma resumida, o cálculo do estimador de máxima verosimilhança processa-se, iterativamente, da seguinte forma:

i. Dado $\widehat{\beta}^{(k)}$ (k=0,1,2,..), calcula-se $t_i^{(k)}$ através da expressão (2.33) e $W^{(k)}$ através de (2.35).

ii. A nova iteração $\widehat{\beta}^{(k+1)}$ é calculada através da expressão (2.36).

As iterações param quando é atingido um critério adequado, como por exemplo o critério (2.37).
O algoritmo do método iterativo dos mínimos quadrados reponderados difere do algoritmo de Newton-Raphson na medida em que o primeiro usa a matriz de informação de Fisher (matriz Hessiana esperada) em vez da matriz Hessiana observada no segundo (Fahrmeir and Tutz, 2001; Hilbe, 2011). Este algoritmo torna-se vantajoso dado que, geralmente, é mais fácil determinar a matriz de informação de Fisher e esta é sempre semi-definida positiva. Quando estamos na presença de um modelo de regressão com função de ligação canónica, as matrizes Hessiana observada e esperada são iguais.
McCullagh and Nelder (1989) provaram que ambos os métodos conduzem ao mesmo algoritmo.

No software R, o ajustamento a um modelo linear generalizado pode ser efetuado através da função *glm()* da biblioteca *stats*. O algoritmo acima descrito para a estimação dos coeficientes de regressão é o utilizado nesta função. Para uma melhor elucidação deste algoritmo, apresentamos de seguida a sua aplicação em diversos modelos de regressão.

Aplicação do algoritmo na Regressão Linear

Para dados em que a variável resposta segue uma distribuição normal, a função de ligação canónica é dada por $\eta_i = \mu_i$ e portanto $\dfrac{\partial \eta_i}{\partial \mu_i} = 1$. Substituindo estas igualdades na equação (2.33) tem-se:

$$t_i = \widehat{\eta_i} + (y_i - \widehat{\mu_i}) \times 1 = y_i.$$

Para além disso, $b''(\theta_i) = \frac{Var(y_i)\omega_i}{\phi} = \frac{\sigma^2 * 1}{\sigma^2} = 1$ pelo que $\varpi_i = \frac{1}{\sigma^2}$. Nesta situação os pesos são constantes o que implica que não é necessária nenhuma iteração.

Aplicação do algoritmo na Regressão Logística

A regressão logística é um modelo linear generalizado com resposta binomial. A função de ligação mais usada é

$$\eta_i = logit(\pi_i) \tag{2.38}$$

onde π_i representa a probabilidade de sucesso ou insucesso de um determinado acontecimento. Para a aplicação do algoritmo é conveniente escrever a função de ligação em termos de μ_i:

$$\eta_i = log\left(\frac{\pi_i}{1-\pi_i}\right) = log\left(\frac{\mu_i}{n_i - \mu_i}\right) = log(\mu_i) - log(n_i - \mu_i), \tag{2.39}$$

onde μ_i é a média da resposta e n_i o número de observações na condição i. Derivando a equação anterior em ordem a μ_i obtemos

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i} + \frac{1}{n_i - \mu_i} = \frac{n_i}{\mu_i(n_i - \mu_i)} = \frac{1}{n_i \pi_i (1-\pi_i)}. \tag{2.40}$$

Nesta caso, a variável dependente de trabalho é então da forma

$$t_i = \eta_i + \frac{y_i - n_i \pi_i}{n_i \pi_i (1-\pi_i)} \tag{2.41}$$

e os pesos são dados por

$$\varpi_i = \frac{1}{n_i \pi_i (1-\pi_i)} \left[n_i \pi_i (1-\pi_i)\right]^2 = n_i \pi_i (1-\pi_i), \tag{2.42}$$

considerando $\omega_i = 1$.

Aplicação do algoritmo na Regressão de Poisson

Na regressão de Poisson a variável resposta segue uma distribuição de Poisson com função de ligação

$$\eta_i = log(\mu_i). \tag{2.43}$$

A derivada da função de ligação em ordem a μ é, portanto,

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i}. \tag{2.44}$$

Assim, a variável dependente de trabalho é da forma

$$t_i = \eta_i + \frac{y_i - \mu_i}{\mu_i} \tag{2.45}$$

e o processo iterativo do peso é dado por

$$\varpi_i = \frac{1}{\mu_i \frac{1}{\mu_i^2}} = \mu_i, \tag{2.46}$$

tendo em conta que $\omega_i = 1$.

Aplicação do algoritmo na Regressão Quasi-Poisson

Na regressão Quasi-Poisson a variável resposta segue uma distribuição Quasi-Poisson com função de ligação

$$\eta_i = log(\mu_i). \tag{2.47}$$

Assim, a derivada da função de ligação é

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i}. \tag{2.48}$$

De acordo com a fórmula apresentada nesta secção, a variável dependente de trabalho é expressa por

$$t_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}. \tag{2.49}$$

Sabe-se que, numa regressão Quasi-Poisson a variável aleatória Y segue uma distribuição Quasi-Poisson com função de probabilidade igual à da distribuição de Poisson e com média da resposta μ e variância $\phi\mu$. O processo iterativo do peso é, deste modo, dado por

$$\varpi_i = \frac{1}{\phi\mu_i \frac{1}{\mu_i^2}} = \frac{\mu_i}{\phi}, \tag{2.50}$$

ou seja, na regressão Quasi-Poisson os pesos são diretamente proporcionais à média (Hoef and Boveng, 2007).

Aplicação do algoritmo na Regressão Binomial Negativa

Na regressão binomial negativa, a variável resposta segue uma distribuição binomial negativa com função de ligação

$$\eta_i = \log(\mu_i), \tag{2.51}$$

e portanto, a derivada a função de ligação é expressa por

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i}. \tag{2.52}$$

Neste caso, a variável dependente de trabalho é definida como

$$t_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}. \tag{2.53}$$

Tem-se $E(Y) = \mu$ e $Var(Y) = \mu + \alpha\mu^2$, pelo que o processo iterativo do peso é traduzido por

$$\varpi_i = \frac{1}{(\mu + \alpha\mu)\frac{1}{\mu_i^2}} = \frac{\mu}{1 + \alpha\mu}. \tag{2.54}$$

Nesta situação, os pesos apresentam uma relação côncava com a média. Isto significa que pequenos valores para a média traduzem pesos bastantes pequenos, mas à medida que a média aumenta, o peso é nivelado para $1/\alpha$ (Hoef and Boveng, 2007).

2.3.2.2 Estimação do parâmetro de dispersão

O parâmetro de dispersão do modelo, ϕ, quando desconhecido, pode ser estimado por (McCullagh and Nelder, 1989)

$$\widehat{\phi} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{\omega_i (y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \tag{2.55}$$

$$= \frac{\chi^2}{n-(p+1)}, \tag{2.56}$$

sendo $\widehat{\phi}$ um estimador consistente de ϕ e χ^2 a estatística de Pearson generalizada.

2.3.3 Testes de Hipóteses

Os testes de hipóteses são uma ferramenta de modelação estatística que nos permite tomar uma decisão para a população acerca de determinadas hipóteses, com base numa amostra. No contexto dos modelos lineares generalizados, pretende-se, por exemplo, averiguar se um determinado conjunto de covariáveis é estatisticamente significativo para o modelo, ou, analisar a significância estatística de cada um dos parâmetros, individualmente, ou comparar a qualidade do ajustamento de dois modelos, entre outras questões. Os testes de hipóteses que vamos apresentar de seguida dizem respeito à significância estatística dos coeficientes de regressão.

2.3.3.1 Teste de Wald

Suponhamos que se pretende testar a hipótese nula

$$H_0 : \beta_j = 0, \quad j = 0, ..., p$$

que indica que o coeficiente independente $\beta_0 = 0$ é irrelevante para o modelo ($j = 0$) ou que a variável explicativa X_j não deve constar do modelo de regressão ($j \neq 0$).
A estatística de teste é, para amostras grandes,

$$W_j = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \sim N(0,1)$$

onde *se* diz respeito ao erro padrão.

2.3.3.2 Teste da razão de verosimilhanças

O teste da razão de verosimilhanças pretende comparar o ajustamento de dois modelos encaixados. Comecemos por comparar um modelo ajustado ω, qualquer, com o modelo completo ou saturado Ω. O modelo saturado é o modelo que estima um parâmetro para cada observação, logo, é o que apresenta uma maior função de verosimilhança. De forma análoga, o modelo nulo (modelo de regressão com apenas um parâmetro, constante) é o que apresenta menor função de verosimilhança. De acordo com Gaio (2012) sejam

- $\widehat{\mu}_i$ o valor ajustado da observação i fornecido pelo modelo ω
- y_i o valor ajustado da observação i fornecido pelo modelo Ω
- $\widehat{\theta}_i$ os parâmetros canónicos estimados pelo modelo ω
- $\overline{\theta}_i$ os parâmetros canónicos estimados pelo modelo Ω
- L_ω a função de verosimilhança do modelo ω
- L_Ω a função de verosimilhança do modelo Ω.

De acordo com a definição da função $log-$verosimilhança (eq.(2.16)) e assumindo que não existem pesos associados às observações, isto é, $a(\phi) = \phi$, o critério da razão de verosimilhanças entre o modelo ajustado ω e o modelo completo Ω é

$$\begin{aligned} -2\log\left(\frac{L_\omega}{L_\Omega}\right) &= -2\left(l_\omega - l_\Omega\right) \\ &= -2\left[\sum_{i=1}^{n}\frac{y_i\widehat{\theta}_i - b(\widehat{\theta}_i)}{\phi} + c(y_i,\phi) - \left(\sum_{i=1}^{n}\frac{y_i\overline{\theta}_i - b(\overline{\theta}_i)}{\phi} + c(y_i,\phi)\right)\right] \\ &= 2\sum_{i=1}^{n}\frac{y_i(\overline{\theta}_i - \widehat{\theta}_i) - b(\overline{\theta}_i) + b(\widehat{\theta}_i)}{\phi} \\ &= \frac{D(y,\widehat{\mu})}{\phi} \end{aligned} \tag{2.57}$$

sendo que o numerador não depende de parâmetros desconhecidos.
A quantidade

$$\begin{aligned} D(y,\widehat{\mu}) &= -2\phi(l_\omega - l_\Omega) \\ &= 2\sum_{i=1}^{n} y_i(\overline{\theta_i} - \widehat{\theta}_i) - b(\overline{\theta}_i) + b(\widehat{\theta}_i) \end{aligned} \tag{2.58}$$

é designada por *desviância* (na secção 2.3.4.1 esta quantidade será descrita com detalhe).
Pretende-se agora saber se o modelo ω faz um bom ajustamento aos dados. Para isso, considere-se o seguinte teste de hipóteses, sob a hipótese nula:

H_0: A qualidade de ajustamento do modelo ω é igual à qualidade de ajustamento do modelo saturado.

Cada um destes grupos representa um padrão de covariáveis.
Sob H_0 e quando os dados se encontram agrupados em grupos relativamente grandes, a estatística de teste é dada por

$$D_\omega \sim \chi^2_{1-\alpha}(J - (p+1)), \tag{2.59}$$

onde D_ω corresponde à desviância do modelo ω, J é o número de padrões de covariáveis diferentes existentes nos dados, p é o número de parâmetros do modelo ω a menos da constante e χ^2 é a distribuição Qui-Quadrado. Rejeita-se H_0 com um nível de significância α caso $D_\omega > \chi^2_{1-\alpha}(J - (p+1))$.
Quando os dados se encontram agrupados em grupos pequenos, um elevado valor destas estatísticas

pode não ser indicação de um mau ajustamento do modelo aos dados. Nesta situação, são necessários mais testes e análises dos resíduos do modelo.
Por fim, considere-se dois modelos encaixados, o modelo ω_1 e o modelo ω_2 ($\omega_1 \in \omega_2$), com número de parâmetros p_1 e p_2, respetivamente, satisfazendo $p_1 < p_2$. Os modelos dizem-se encaixados dado que o conjunto dos parâmetros do modelo ω_1 é um subconjunto dos parâmetros do modelo ω_2.
Nesta caso, o logaritmo da razão de verosimilhanças para os modelos em questão é

$$\begin{aligned} -2\log\left(\frac{L_{\omega_1}}{L_{\omega_2}}\right) &= -2\log\left(\frac{\frac{L_{\omega_1}}{L_\Omega}}{\frac{L_{\omega_2}}{L_\Omega}}\right) \\ &= -2\log\left(\frac{L_{\omega_1}}{L_\Omega}\right) + 2\log\left(\frac{L_{\omega_2}}{L_\Omega}\right) \\ &= \frac{D_{\omega_1} - D_{\omega_2}}{\phi}. \end{aligned}$$

Caso não se conheça o parâmetro de dispersão ϕ, este será estimado a partir do modelo ω_2.
Para que possamos avaliar a qualidade do ajustamento entre os dois modelos encaixados, aplica-se o teste da razão de verosimilhanças para modelos encaixados, sob a hipótese nula:

H_0: os dois modelos têm a mesma qualidade de ajustamento

onde a estatística de teste é dada por

$$G = \frac{D_{\omega_1} - D_{\omega_2}}{\phi} \sim \chi^2_{1-\alpha}(p_2 - p_1).$$

A decisão passa por rejeitar a hipótese nula com nível de significância α se $G > \chi^2_{1-\alpha}(p_2 - p_1)$.

2.3.4 Qualidade do ajustamento

2.3.4.1 Desviância

Conforme referido na secção anterior, o modelo saturado (modelo Ω) é utilizado para avaliar a qualidade do ajustamento de um certo modelo (modelo ω), através da introdução de uma medida de distância baseada no critério da razão de verosimilhanças.
Na secção 2.3.3.2 vimos que o critério da razão de verosimilhança é dado por

$$-2\log\left(\frac{L_\omega}{L_\Omega}\right) = \frac{D(y,\widehat{\mu})}{\phi},$$

onde

$$\begin{aligned} D(y,\widehat{\mu}) &= -2\phi(l_\omega - l_\Omega) \\ &= 2\sum_{i=1}^{n} y_i(\overline{\theta_i} - \widehat{\theta}_i) - b(\overline{\theta}_i) + b(\widehat{\theta}_i) \end{aligned} \tag{2.60}$$

é conhecida como desviância.
A razão de log $-$verosimilhanças

$$-2\log\left(\frac{L_\omega}{L_\Omega}\right) = D^*(y,\widehat{\mu})$$

é designada por desviância re-escalonada por corresponder ao quociente entre a desviância e o parâmetro de dispersão ϕ. Quando ϕ=1, por exemplo na distribuição binomial e poisson (caso não haja sobre-dispersão), a desviância coincide com a desviância re-escalonada.
A desviância de um modelo avalia a discrepância entre os valores observados (valores ajustados pelo modelo saturado) e os valores ajustados pelo modelo ou seja, avalia a qualidade do ajustamento. Deste modo, o valor de D é sempre maior ou igual a zero e será tanto maior quanto maior for a discrepância entre o modelo estimado e os dados. Um modelo com um ajustamento perfeito terá desviância igual a zero, como é o caso do modelo saturado. A adição de covariáveis no modelo nulo leva a uma redução do valor da desviância (McCullagh and Nelder, 1989). Para o modelo linear a desviância é dada pela soma dos quadrados dos resíduos.

2.3.4.2 Estatística χ^2 de Pearson generalizada

Outra importante medida de discrepância é a estatística χ^2 de Pearson generalizada, que toma a forma

$$X_P^2 = \sum_{i=1}^{n} \frac{\omega_i(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \tag{2.61}$$

onde $V(\widehat{\mu}_i)$ é a função de variância estimada para a distribuição do modelo em causa. Para a distribuição normal, X_P^2 é mais uma vez a soma dos quadrados dos resíduos, enquanto que para a distribuição de Poisson ou Binomial esta é a estatística χ^2 de Pearson original (McCullagh and Nelder, 1989).
Apesar de a estatística de Pearson ser uma medida alternativa da qualidade do ajustamento, a diferença entre estatísticas de Pearson não pode ser usada para comparação de modelos encaixados, contrariamente ao que acontece com a desviância. A diferença entre as estatísticas de Pearson não pode ser utilizada dado que não é conhecida a distribuição para a diferença ente estas estatísticas.

2.3.5 Resíduos

Os resíduos são utilizados para explorar a adequação do ajustamento do modelo no que diz respeito à escolha da função de variância, da função de ligação e dos termos no preditor linear. Para além disso, os resíduos indicam também a presença de valores anómalos no modelo, isto é, observações mal ajustadas que não são bem explicadas pelo modelo (McCullagh and Nelder, 1989).
No modelo de regressão linear, a variável resposta pode ser escrita na forma

$$\begin{aligned} Y(X) &= \mu(X) + u \\ &= \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + u, \end{aligned}$$

onde $u \sim N(0, \sigma^2(x))$ são os resíduos.
Os resíduos medem a discrepância entre os valores observados y_i e os valores ajustados $\widehat{\mu}_i$. Para os modelos lineares generalizados é necessário estender a definição de resíduos para que estes possam ser aplicados a todos as distribuições e não só à distribuição normal.
Na secção seguinte apresentamos dois tipos de resíduos, os resíduos de Pearson e os resíduos de desviância.

2.3.5.1 Resíduos de Pearson

O resíduo de Pearson para uma dada observação i é dado por

$$R_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{var(Y_i)}}} = \frac{(y_i - \widehat{\mu}_i)\omega_i}{\sqrt{\widehat{\phi} V(\widehat{\mu}_i)}}.$$

Este resíduo corresponde à contribuição de cada observação para o cálculo da estatística de Pearson. Pretende-se que os resíduos estimados sejam padronizados para que seja possível fazer uma análise adequada. O facto de estes estarem padronizados implica que têm variância constante.
Dado que se tem $var(Y_i - \widehat{\mu}_i) \approx var(Y_i)(1 - h_{ii})$ (Turkman and Silva, 2000), o resíduo de Pearson estandardizado é

$$R_P^* = \frac{(y_i - \widehat{\mu}_i)\omega_i}{\sqrt{\widehat{\phi} V(\widehat{\mu}_i)(1 - h_{ii})}},$$

onde h_{ii} são as leverages do modelo ajustados. Nos casos em que as leverages não são excessivamente elevadas, não existe grande diferença entre os gráficos dos resíduos de Pearson brutos e dos resíduos de Pearson estandardizados, excluindo a escala que apresentam.

2.3.5.2 Resíduos de desviância

Como referido anteriormente, a desviância é usada como uma medida de discrepância nos modelos lineares generalizados. Cada observação i contribui com uma quantidade d_i para esta medida de discrepância e portanto,

$$D(y, \widehat{\mu}) = \sum_{i=1}^{n} d_i.$$

O resíduos de desviância da i−ésima observação é assim definido por

$$R_D = sinal(y_i - \widehat{\mu}_i)\sqrt{d_i}.$$

Note-se que o resíduo de desviância aumenta com $y_i - \widehat{\mu}_i$ onde $\sum R_D^2 = D$. O resíduo de desviância estandardizado expressa-se por

$$R_D^* = \frac{R_D}{\sqrt{\widehat{\phi}(1 - h_{ii})}}.$$

2.3.6 Seleção de Modelos

Quando se seleciona um modelo, é necessário ter em mente que o modelo selecionado não é o único "correto". Como refere Agresti (2007), qualquer modelo é uma simplificação da realidade logo, não existem modelos verdadeiros. Há apenas modelos aproximados da realidade que causam perda de informação. Deste modo, é necessário proceder à seleção do "melhor"modelo, dentro daqueles que foram ajustados, para explicar o fenómemo em estudo.
Quando dois modelos não são encaixados, não é possível utilizar o teste de razão de verosimilhanças, pelo que se torna aconselhável utilizar um critério que meça a quantidade de informação que o modelo recolhe dos dados, por oposição ao ruído que o modelo não consegue explicar.

Agresti (2007) afirma que o melhor critério de seleção de modelos é o *critério de informação de Akaike* (AIC) [5], proposto por Akaike (1974). Este critério é definido por:

$$AIC = -2l + 2p, \tag{2.62}$$

onde l representa o logaritmo da verosimilhança do modelo e p o número total de parâmetros do modelo.
Dado um conjunto de diversos modelos ajustados aos dados, o "melhor"modelo é aquele que possui o menor valor de AIC. Note-se que o AIC é uma medida da qualidade do ajustamento que penaliza modelos com muitos parâmetros. É importante salientar, também, que o AIC não traduz, no sentido absoluto, se um modelo faz um bom ajustamento aos dados ou não. Apenas revela se um determinado modelo é preferível relativamente a outro.
Um outro critério de informação é o *critério de informação Bayesiano* (BIC) [6]. Diversos autores preferem a utilização deste critério em vez do AIC uma vez que este, para além de penalizar o número de parâmetros do modelo, penaliza também o número de observações. Contudo, nesta dissertação utilizaremos o critério AIC para a seleção de modelos.

2.3.7 Quasi−verosimilhança

Muitas vezes os modelos que formulamos não são apresentados de acordo com a formulação descrita na secção 2.2. Isto porque, em muitas situações, existe sobredispersão no modelo e portanto existe a necessidade de alterar a variância através da introdução de um parâmetro de dispersão (ver secção 2.4.1). Assim, o modelo deixa de estar especificado uma vez que não existe nenhuma distribuição dentro da família exponencial que apresente esses valores de média e variância. Este modelo fica então especificado apenas pela média e variância, pelo que não é possível fazer inferências acerca dele através da função de verosimilhança. Contudo, é possível fazer inferências sobre este modelo, considerando modelos quasi-verosimilhança. Os modelos de quasi-verosimilhança não têm em conta as suposições da família exponencial, apenas consideram a estrutura da média e da variância. Deste modo, não é necessário assumir qualquer distribuição para o modelo, sendo apenas necessário especificar o primeiro e segundo momentos (Fahrmeir and Tutz, 2001).

Considere-se o vetor resposta Y e a *função score* definida por

$$U = U(\mu, y) = \frac{Y - \mu}{\phi V(\mu)}, \tag{2.63}$$

que apresenta as seguinte propriedades em comum com a derivada da função $\log -$ verosimilhança:

$$\begin{aligned} E(U) &= 0, \\ Var(U) &= \frac{1}{\phi V(\mu)}, \\ -E\left(\frac{\partial U}{\partial \mu}\right) &= \frac{1}{\phi V(\mu)}. \end{aligned} \tag{2.64}$$

Dado que a função score é a derivada da função $\log -$verosimilhança é de esperar que o integral de U,

$$Q(\mu, y) = \int_y^\mu \frac{y - t}{\phi V(t)} dt \tag{2.65}$$

[5] Akaike information criterion (AIC).
[6] Bayesian information criterion (BIC).

caso exista, se comporte como uma função de log −verosimilhança. $Q(\mu, y)$ é definida por função de quasi−verosimilhança ou, mais corretamente, por função de quasi− log −verosimilhança (McCullagh and Nelder, 1989).
Se considerarmos que temos n observações de variáveis aleatórias independentes, a quasi−verosimilhança para os dados completos é o somatório de cada contribuição individual

$$Q(\mu, y) = \sum_{i=1}^{n} Q_i(\mu_i, y_i). \tag{2.66}$$

Turkman and Silva (2000) referem que $Q(\mu, y)$ pode mesmo ser uma função de log −verosimilhança. Para isso basta provar que se existe uma função de log −verosimilhança l tal que

$$\frac{\partial l}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)},$$

com $E(Y) = \mu$ e $var(Y) = \phi V(\mu)$, então l tem a estrutura correspondente a uma função de log −verosimilhança da família exponencial.
Nesta caso, a função de quasi-desviância é dada por

$$D(y, \mu) = -2\phi Q(\mu, y) = 2\int_{\mu}^{y} \frac{y - t}{V(t)} dt \tag{2.67}$$

e é sempre estritamente positiva exceto quando $y = \mu$. Note-se que a desviância apenas depende de y e μ, não depende do parâmetro ϕ.

Na secção 2.2 vimos que $g(\mu_i) = \eta_i$ e portanto, temos que $\mu_i = g^{-1}(\eta_i) = g^{-1}(\tilde{\mathbf{x}}_i^T \beta)$. Substituindo μ_i na equação de quasi−verosimilhança (eq. (2.66)),

$$Q(\mu, y) = \sum_{i=1}^{n} \frac{y_i - g^{-1}(\tilde{\mathbf{x}}_i^T \beta)}{var(Y_i)}.$$

Para obtermos o sistema de equações de estimação da função quasi−verosimilhança para os parâmetros de regressão β basta igualarmos a zero as derivadas de $Q(\mu_i, y_i)$ em ordem a β_j, $\quad j = 0, ..., p$,

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\tilde{x}_{ij}}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} \quad = \quad 0. \tag{2.68}$$

Repare-se que o sistema de equações não depende do parâmetro de dispersão ϕ e que este coincide com as equações de verosimilhança da equação (2.20).
A função $s^*(\beta) = \frac{\partial Q}{\partial \beta}$ é conhecida como função quasi−score e as equações em (2.68) são chamadas de equações de quasi−verosimilhança, sendo as estimativas resultantes estimativas de quasi−máxima verosimilhança.
Os modelos quasi−verosimilhança também utilizam o método iterativo de mínimos quadrados reponderados para estimar os coeficiente do modelo (Hoef and Boveng, 2007).

2.4 Exemplos

2.4.1 Regressão de Poisson

Um modelo linear generalizado com resposta de Poisson é designado por regressão de Poisson. A regressão de Poisson é utilizada para modelar o número de ocorrências de um determinado acontecimento ou a taxa de ocorrência de um acontecimento em função de um grupo de variáveis explicativas. Assim, este tipo de regressão é utilizada quando se lida com dados na forma de contagem. Exemplos destes dados incluem o número de ciclones tropicais que cruzam a costa norte da Austrália, o número de chamadas recebidas por uma operadora ou o número de acidentes numa determinada estrada.
Como definido na secção 2.1.1.2, se Y é uma variável aleatória que segue uma distribuição de Poisson $P(\mu)$, a sua função de probabilidade é

$$f(y|\mu) = \mu^y \frac{e^{-\mu}}{y!}, \tag{2.69}$$

onde μ representa o número médio de ocorrências de um dado acontecimento, $\mu > 0$. Uma caraterística da distribuição de Poisson reside no facto de a média da variável aleatória ser igual à variância desta, isto é,

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

o que implica que qualquer fator que afete a média afetará também a variância.
Uma propriedade da distribuição de Poisson é que o somatório de variáveis aleatórias independentes que seguem uma distribuição de Poisson é também uma distribuição de Poisson. Matematicamente, se Y_1 e Y_2 são duas variáveis aleatórias independentes tais que $Y_i \sim P(\mu_i)$, com $i = 1, 2$, então

$$Y_1 + Y_2 \sim P(\mu_1 + \mu_2). \tag{2.70}$$

Este resultado generaliza-se, naturalmente, à soma de um número qualquer de parcelas, e garante que os resultados produzidos por um modelo de Poisson a partir de dados individuais ou agrupados são equivalentes. Em termos de estimação obtém-se exatamente a mesma função de verosimilhança.

Seja $Y_1, ..., Y_n$ uma amostra aleatória de uma variável aleatória Y que representa o número de ocorrências de um acontecimento raro num determinado período de tempo ou espaço. Dado um vetor de variáveis explicativas $X = (X_1, ..., X_p)$ e uma observação $\mathtt{x_i}^T = (x_{i1}, ..., x_{ip})$ do indivíduo i, assume-se que

$$Y|X = \mathtt{x_i} \sim P(\mu(\mathtt{x_i}))$$

onde

$$\mu_i = \mu(\mathtt{x_i})$$

representa o número médio de ocorrências de um determinado acontecimento dada a observação $\mathtt{x_i}$. Naturalmente tem-se que

$$\mu_i = E(Y|X = \mathtt{x_i}) = Var(Y|X = \mathtt{x_i}).$$

Pretende-se modelar a média de $Y|X = \mathtt{x_i}$ como combinação linear das variáveis explicativas. Poderia escrever-se um modelo linear simples da forma

$$\mu(\mathtt{x_i}) = \tilde{\mathtt{x}}_i^T \beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}.$$

Contudo, este modelo é desajustado uma vez que o preditor linear, do lado direito, pode assumir qualquer valor real enquanto que o termo do lado esquerdo (a média de Poisson) só pode tomar valores não negativos. Uma solução para este problema passa por considerar a transformação logarítmica como função de ligação do modelo linear generalizado

$$\log(\mu(\mathtt{x_i})) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}.$$

Assim, o modelo de regressão de Poisson é definido por

$$Y|X = \mathtt{x_i} \sim P(\mu(\mathtt{x_i})) \tag{2.71}$$

e

$$\log(\mu(\mathtt{x_i})) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \tag{2.72}$$

Note-se que no modelo de regressão de Poisson o logaritmo é a função de ligação canónica para a distribuição de Poisson.
Neste modelo, os coeficientes de regressão β_j, $j = 0, ..., p$, representam a variação esperada no logaritmo da média por unidade de variação na variável explicativa $\mathtt{x}_j$.
Da equação (2.72) obtém-se um modelo multiplicativo para a média da resposta

$$\begin{aligned} \mu(\mathtt{x_i}) &= exp\{\beta_0 + \beta_1 x_{i1} + ... + \beta_K p x_{ip}\} \\ &= e^{\beta_0} e^{\beta_1 x_{i1}} ... e^{\beta_p x_{ip}}. \end{aligned}$$

Esta equação traduz assim o efeito multiplicativo, $exp(\beta_j)$, de cada cada uma das variáveis explicativas X_j sobre a média da resposta. Uma vantagem de usar a função logaritmo como função de ligação é o facto de os efeitos das variáveis explicativas serem sempre multiplicativos em vez de aditivos.

Dados agrupados

Através da propriedade apresentada em (2.70) é possível analisar dados individuais ou agrupados com resultados equivalentes. Suponha-se que existem N_i unidades de exposição do padrão de covariáveis i, onde Y_i é o número de ocorrências observadas nas N_i unidades de exposição do padrão de covariáveis i. Para $l = 1, ..., N_i$, considere-se que Y_{il} representa o número de ocorrências observadas na $l-$ésima unidade individual de exposição do padrão de covariáveis i. Assumindo a condição de independência, se $Y_{il} \sim P(\mu_i)$, para $l = 1, 2, ..., N_i$, então $Y_i \sim P(N_i \mu_i)$, pela propriedade da soma de variáveis independentes.
Em termos de regressão tem-se que

$$\begin{aligned} \log(E(Y_{il})) &= \log(\mu_i) \\ &= \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \end{aligned}$$

Assim, o modelo para as médias por grupo é

$$\begin{aligned} \log(E(Y_i)) = \log(N_i \mu_i) & \\ &= \log(N_i) + \log(\mu_i) \\ &= \log(N_i) + \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \end{aligned}$$

Os totais por grupo seguem assim um modelo log linear exatamente com os mesmos coeficientes β que as médias individuais, à exceção do termo independente que é alterado de $\log(N_i)$.
O termo $\log(N_i)$ é denominado por *offset* do modelo dado que não terá uma estimativa correspondente no modelo.
Em termos de estimação obtém-se exatamente a mesma função de verosimilhança caso se trabalhe com dados de contagem individuais, Y_{ij}, ou em grupo, Y_i (Rodríguez, 2007).

Interpretação dos parâmetros do modelo - Risco Relativo

O risco relativo (RR) corresponde ao quociente entre o número esperado de eventos num grupo exposto e num grupo não exposto. Pode ser encarado como uma medida de associação entre o número de eventos e uma exposição (Gaio, 2012). O risco relativo é utilizado na interpretação dos parâmetros β do modelo, supondo que não existem interações entre as variáveis explicativas $X_1, ..., X_p$.
Comecemos por considerar a situação em que X_j toma valores **contínuos**. Aumentando a componente j de $\mathtt{x} = (\mathtt{x_1}, ..., \mathtt{x_p})$ de uma unidade,

$$\mathtt{x}^{(+1)j} = (\mathtt{x_1}, ..., \mathtt{x_{j-1}}, \mathtt{x_j} + 1, \mathtt{x_{j+1}}, ..., \mathtt{x_p}),$$

tem-se

$$\begin{aligned} \mu(\mathtt{x}) &= exp(\beta_0 + ... + \beta_j \mathtt{x}_j + ... + \beta_p \mathtt{x}_p) \\ \mu(\mathtt{x}^{(+1)j}) &= exp(\beta_0 + ... + \beta_j \mathtt{x}_j + \beta_j + ... + \beta_p \mathtt{x}_p), \end{aligned}$$

pelo que

$$\frac{\mu\left(\mathtt{x}^{(+1)j}\right)}{\mu(\mathtt{x})} = exp(\beta_j).$$

Deste modo, o risco relativo para a resposta é dado por

$$RR\left(\mathtt{x}^{(+1)j}, \mathtt{x}\right) = \frac{E\left(Y|X = \mathtt{x}^{(+1)j}\right)}{E\left(Y|X = \mathtt{x}\right)} = exp(\beta_j),$$

o que significa que um aumento de uma unidade em X_j provoca uma alteração de e^{β_j} no risco relativo para a resposta, mantendo as restantes variáveis explicativas constantes.
Este raciocínio pode ser generalizado caso consideremos uma aumento de c unidades na coordenada j de $\mathtt{x} = (\mathtt{x_1}, ..., \mathtt{x_p})$. Neste caso tem-se,

$$\mathtt{x}^{(+c)j} = (\mathtt{x_1}, ..., \mathtt{x_j} + \mathtt{c}, ..., \mathtt{x_p})$$

e portanto

$$RR\left(\mathtt{x}^{(+c)j}, \mathtt{x}\right) = \frac{E\left(Y|X = \mathtt{x}^{(+c)j}\right)}{E\left(Y|X = \mathtt{x}\right)} = exp(c\beta_j).$$

Analogamente à interpretação anterior, o risco relativo traduz que um aumento de c unidades em X_j provoca uma alteração de $c\beta_j$ no logaritmo da média da resposta.
Considerando agora que temos duas observações, $\mathtt{x_1} = (\mathtt{x_{11}}, ..., \mathtt{x_{1p}})$ e $\mathtt{x_2} = (\mathtt{x_{21}}, ..., \mathtt{x_{2p}})$, tem-se

$$\log\left(RR(\mathtt{x_1}, \mathtt{x_2})\right) = \log(\mu(\mathtt{x_1})) - \log(\mu(\mathtt{x_2})) = (\tilde{\mathtt{x}}_1^T - \tilde{\mathtt{x}}_2^T)\beta$$

e portanto

$$\begin{aligned} RR(\mathtt{x_1}, \mathtt{x_2}) &= e^{(\tilde{\mathtt{x}}_1^T - \tilde{\mathtt{x}}_2^T)\beta} \\ &= e^{\sum_{j=0}^{p}(\mathtt{x}_{1j} - \mathtt{x}_{2j})\beta_j} \end{aligned}$$

Consideremos agora a situação em que X_j toma valores **dicotómicos**. Assuma-se, sem perda de generalidade, que as categorias de X_j estão codificadas em 0 e 1. Tem-se que

$$\begin{aligned} \log(RR) &= \log(RR(X_j = 1, X_j = 0)) \\ &= \log(E(Y|X_j = 1)) - \log(E(Y|X_j = 0)) \\ &= \beta_0 + ... + \beta_{j-1}\mathtt{x}_{j-1} + \beta_j + \beta_{j+1}\mathtt{x}_{j+1} + ... + \beta_p\mathtt{x}_p \\ &\quad -\beta_0 - ... - \beta_{j-1}\mathtt{x}_{j-1} - \beta_{j+1}\mathtt{x}_{j+1} - ... + \beta_p\mathtt{x}_p \\ &= \beta_j \end{aligned}$$

e por isso,

$$RR = e^{\beta_j}$$

Isto significa que, na população em estudo, o número esperado de ocorrências entre os indivíduos com $X_j = 1$ seja e^{β_j} vezes o número esperado de ocorrências entre os indivíduos com $X_j = 0$, mantendo as restantes variáveis explicativas constantes.
Considere-se agora que a variável X_j toma valores **politómicos**.
Este situação é uma generalização do caso em que X_j é dicotómica. A variável politómica com C categorias é representada por C-1 variáveis dummy, sendo uma das categorias designada por 0. O modelo de Poisson estima depois um RR para cada uma das classes 1,2,...,C-1, usando a classe 0 como referência.

A constante β_0 do modelo de regressão corresponde ao logaritmo do número médio de ocorrências com características nulas em todas as variáveis explicativas em estudo.
De um modo geral, o risco relativo pode ser interpretado da seguinte forma:

- RR>1 indica existência de associação positiva entre a exposição e o risco de um determinado evento acontecer
- RR próximo de 1 indica a inexistência de associação entra a exposição e o risco
- 0<RR<1 indica a existência de associação negativa entre a exposição e o risco de um determinado evento acontecer.

Intervalos de confiança para o risco relativo

Sob a hipótese nula

$$H_0 : RR = 0$$

de inexistência de associação entre a resposta e o fator em estudo, é possível mostrar-se que uma estimativa $\widehat{RR}$ do risco relativo segue aproximadamente uma distribuição $\log - normal$ de valor

esperado igual a 1. Deste modo, intervalos de confiança a $(1-\alpha)\%$ para RR são dados pela fórmula (Gaio, 2012)

$$\left(\widehat{RR} \quad e^{\pm z_{\alpha/2}\sqrt{\frac{1}{O_1}+\frac{1}{O_2}}}\right)$$

onde O_1 e O_2 representam o número de eventos observados nos expostos e nos não expostos, respetivamente, e $z_{\frac{\alpha}{2}} = 1.96$ é o correspondente quantil da distribuição normal.
O risco relativo estimado pelo modelo será estatisticamente significativo sempre que o correspondente intervalo de confiança não contiver o valor 1.

Estimação pelo método da máxima verosimilhança

Como visto anteriormente, a estimação dos modelos lineares generalizados é feita através do método da máxima verosimilhança.
A função de verosimilhança para n observações independentes de Poisson é o produto das probabilidades dado pela equação (2.69). O logaritmo da função de verosimilhança do modelo de poisson é, a menos de uma constante $\log(y_i!)$,

$$\log(L(\beta)) = \sum \{y_i \log(\mu(\mathtt{x_i})) - \mu(\mathtt{x_i})\} \tag{2.73}$$

onde $\mu(\mathtt{x_i})$ depende do vetor de covariáveis $\mathtt{x_i}$ e $log(\mu(\mathtt{x_i}))$ é dado pela equação (2.72). Derivando o lado direito da igualdade 2.73 em ordem a β e igualando a zero, mostra-se que as estimativas de máxima verosimilhança $\widehat{\beta}$ de β, e portanto $\widehat{\mu}$, satisfazem

$$Xy = X\widehat{\mu}, \tag{2.74}$$

onde X é a matriz de especificação, y é o vetor resposta e $\widehat{\mu}$ é o vetor dos valores previsto pelo modelo.
Na verdade, a equação acima descrita verifica-se não só no modelo de regressão de Poisson assim como em todos os modelos lineares generalizados com função de ligação canónica. Contudo, esta equação não é satisfeita no modelo probit para dados binários (Rodríguez, 2007). Em particular, se o modelo linear generalizado estiver de acordo com o definido acima e possuir termo constante assim como função de ligação canónica, a soma dos valores observados é igual à soma dos valores previstos pelo modelo. Para que possamos saber qual a estimativa de cada parâmetro de regressão é necessário, como já foi referido anteriormente, recorrer ao método iterativos de mínimos quadrados reponderados (ver secção 2.3.2.1).

Qualidade do ajustamento

Considerando que y_i são os valores observados e $\widehat{\mu}_i$ os valores previstos pelo modelo, a desviância (ver equação (2.60)) para respostas de Poisson é dada por

$$\begin{aligned} D(y,\widehat{\mu}) &= 2\sum (y_i \log(y_i) - y_i - \log(y_i!) - y_i \log(\widehat{\mu}_i) + \widehat{\mu}_i + \log(y_i!)) \\ &= 2\sum \left(y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right). \end{aligned} \tag{2.75}$$

O primeiro termo representa duas vezes a soma dos valores observados multiplicado pelo logaritmo do quociente entre os valores observados e previsto pelo modelo. Já o segundo termo diz respeito à soma das diferenças entre os valores observados e previsto, que é usualmente zero (Rodríguez, 2007). Isto porque, se derivarmos a função de verosimilhança em ordem a β e igualarmos a zero, obtemos a equação (2.74) e, caso exista uma constante β_0 no modelo, a primeira coluna da matriz X consiste unicamente de 1's e portanto

$$\sum y_i = \sum \widehat{\mu}_i,$$

ou equivalentemente,

$$\sum (y_i - \widehat{\mu}_i) = 0.$$

Assim, para modelos com termo constante, a desviância reduz-se a

$$D(y, \widehat{\mu}) = 2 \sum \left(y_i \log \left(\frac{y_i}{\widehat{\mu}_i} \right) \right).$$

A qualidade do ajustamento de uma modelo de regressão de Poisson pode também ser avaliada pela estatística de Pearson generalizada,

$$X_P^2 = \sum \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

O numerador corresponde ao quadrado da diferença entre os valores observados e os valores previstos enquanto que o denominador corresponde à variância do valor observado.

Sobredispersão

A imposição pelo modelo Poisson da variância ser igual ao valor médio, produz, com alguma frequência, problemas de sobredispersão. Isto é, na maior parte dos padrões de covariáveis a variância é superior à média. Assuma-se agora que a variância é proporcional à média,

$$Var(Y) = \phi E(Y) = \phi \mu,$$

onde ϕ representa o parâmetro de dispersão.
Em termos de estimação, as estimativas pontuais são exatamente iguais às da situação em que não existe sobredispersão mas a variância dos estimadores $\widehat{\beta}$ é inflacionada pelo fator ϕ.
Muitos autores sugerem a estimação de ϕ dada pelo quociente entre a estatística χ^2 de Pearson e o número de graus de liberdade correspondente (Rodríguez, 2007),

$$\widehat{\phi} = \frac{\chi^2}{n - (p+1)},$$

sendo que $\phi > 1$ indica sobredispersão. Hoef and Boveng (2007) referem que o facto da sobredispersão ser tão comum levou a que se desenvolvessem modelos para este tipo de dados, incluindo modelos como a Binomial Negativa, Quasi-Poisson (Wedderburn, 1974), Poisson generalizado (Consul, 1989) e zeros inflacionados (Lambert, 1992).
Soluções possíveis para resolver um problema de sobredispersão num modelo de Poisson consistem então no ajustamento a um modelo de regressão Binomial Negativa ou no ajustamento a um modelo de regressão Quasi-Poisson, considerando também o logaritmo como função de ligação.

2.4.2 Regressão Quasi-Poisson

O modelo Quasi−Poisson pode ser interpretado como um modelo linear generalizado. Considere-se que Y é uma variável aleatória que segue uma distribuição de Quasi-Poisson com função de probabilidade igual à da distribuição de Poisson satisfazendo

$$\begin{aligned} E(Y) &= \mu \\ Var(Y) &= \phi\mu \end{aligned} \tag{2.76}$$

onde E(Y) é o valor esperado da variável aleatória Y, var(Y) é a variância de Y e ϕ é o parâmetro de dispersão. Note-se que a variância estabelece uma relação linear com a média. Como referem Hoef and Boveng (2007) a relação próxima entre a equação (2.76) e a média e variância de uma distribuição de Poisson, juntamente com o logaritmo como função de ligação justificam o facto deste modelo ser conhecido como modelo Quasi−Poisson, denotado por $Y \sim P(\mu, \phi)$. Este modelo é caraterizado pelos dois primeiros momentos, média e variância (Wedderburn, 1974).
Para o modelo de regressão Quasi−Poisson assume-se que

$$\begin{aligned} Y|X &= \mathtt{x_i} \sim P(\mu, \phi) \\ &\text{e} \\ \log(\mu(\mathtt{x_i})) &= \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \end{aligned} \tag{2.77}$$

Os parâmetros do modelo Quasi-poisson são estimados através da quasi-verosimilhança que recorre posteriormente ao método iterativo dos mínimos quadrados reponderados (Hoef and Boveng, 2007). A interpretação destes parâmetros é efetuada através do risco relativo, tal como na regressão de Poisson.

2.4.3 Regressão Binomial Negativa

Como referido anteriormente, quando se fala em modelo de regressão Binomial Negativa estamos a referir-nos à parametrização NB2. A regressão Binomial Negativa é utilizada, por exemplo, quando existe sobredispersão no modelo de Regressão de Poisson, isto é, quando a variância excede a média e apresenta uma relação quadrática com a média.
Considere-se então que $Y \sim BN(\mu, \alpha)$ com função probabilidade

$$\begin{aligned} f(y;\mu,\alpha) &= \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \\ &= \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}, \end{aligned} \tag{2.78}$$

onde α é denominado por parâmetro de heterogeneidade ou parâmetro ancilar. Uma particularidade da distribuição Binomial Negativa é o facto de a variância exceder a média, isto é,

$$E(Y) = \mu \qquad \text{e} \qquad Var(Y) = \mu + \alpha\mu^2.$$

Repare-se que, quando o valor do parâmetro de heterogeneidade (α) tende para zero, a distribuição Binomial Negativa tende para a distribuição de Poisson. Considere-se a variável aleatória Y, com n observações que representa o número de ocorrência de um determinado acontecimento num certo período de tempo ou espaço. Dado um vetor de variáveis explicativas $X = (X_1, ..., X_p)$ e uma observação $\mathtt{x_i}^T = (x_{i1}, ..., x_{ip})$ do indivíduo i, assume-se que

$$Y|X = \mathtt{x_i} \sim BN(\mu(\mathtt{x_i}), \alpha) \tag{2.79}$$

onde

$$\mu_i = \mu(\mathtt{x_i})$$

representa o número médio de ocorrências de um determinado acontecimento dada a observação $\mathtt{x_i}$. Naturalmente, tem-se que

$$\mu_i = E(Y|X = \mathtt{x_i}) \qquad \text{e} \qquad Var(Y|X = \mathtt{x_i}) = \mu_i + \alpha\mu_i^2.$$

O modelo de regressão Binomial Negativa é expresso por (2.79) e por

$$\log(\mu(\mathtt{x_i})) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \tag{2.80}$$

onde o log representa a função de ligação do modelo em questão.
Tomando a exponencial da equação (2.80), obtemos um modelo multiplicativo para a média da resposta.
A interpretação dos parâmetros no modelo de regressão Binomial Negativa é executada de acordo com o risco relativo e, portanto, é análoga à interpretação dos parâmetros no modelo de regressão de Poisson.

Estimação pelo método da máxima verosimilhança

Dado que o modelo de regressão Binomial Negativa é um modelo linear generalizado, os parâmetros deste modelo são obtidos através do método da máxima verosimilhança.
A função log −verosimilhança para n observações da distribuição Binomial Negativa é dada por (Hilbe, 2011),

$$\begin{aligned} \log(L(\beta)) = l(\beta) \quad = \quad & \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \left(\frac{1}{\alpha}\right)\log(1+\alpha\mu_i) \right. \\ & \left. + \log\Gamma\left(y_i + \frac{1}{\alpha}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\alpha}\right) \right\}, \end{aligned} \tag{2.81}$$

onde $\mu(\mathtt{x_i}) = exp\left\{\beta_0 + \beta_1 \mathtt{x}_{i1} + ... + \beta_p \mathtt{x}_{ip}\right\}$.

Qualidade do ajustamento

A desviância da regressão Binomial Negativa, conhecida como uma medida de discrepância entre os valores ajustados e os valores observados é traduzida pela seguinte expressão

$$D(y, \widehat{\mu}) = 2\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - \left(\frac{1}{\alpha} + y_i\right)\log\left(\frac{1+\alpha y_i}{1+\alpha\widehat{\mu}_i}\right) \right\}. \tag{2.82}$$

Nesta situação, a estatística de Pearson generalizada é

$$\chi^2_P = \frac{(y_i - \widehat{\mu}_i)}{\widehat{\mu}_i + \alpha\widehat{\mu}_i^2}. \tag{2.83}$$

2.4.4 Regressão Logística

A regressão logística é um modelo linear generalizado com resposta que segue uma distribuição binomial.
Como visto anteriormente, a função densidade de probabilidade da distribuição Binomial é

$$f(y|n,\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0,1,...,n \tag{2.84}$$

onde n é o número de tentativas de um determinado acontecimento e π a probabilidade de sucesso desse mesmo acontecimento. Nesta situação, $\pi^y(1-\pi)^{n-y}$, é a probabilidade de se obter y sucessos e $n-y$ insucessos para um valor de y específico e o coeficiente de combinatória é o número de possibilidades de se obter y sucessos em n tentativas.
A média e a variância da distribuição binomial são

$$\begin{aligned} E(Y) &= n\pi \\ &e \\ Var(Y) &= n\pi(1-\pi). \end{aligned} \tag{2.85}$$

Note-se que a média e a variância dependem ambas da probabilidade π o que implica que qualquer fator que afete esta probabilidade irá afetar também a média e a variância (Rodríguez, 2007).
Em termos de regressão, considere-se $Y_1,...,Y_n$ uma amostra aleatória binária Y codificada por 0 e 1, onde Y representa o número de sucessos numa sequência de n tentativas. Dado um vetor de variáveis explicativas $X = (X_1,...,X_p)$ e uma observação $\mathtt{x_i}^T = (x_{i1},...,x_{ip})$ do indivíduo i, assume-se que

$$Y|X = \mathtt{x_i} \sim B(1, \pi(\mathtt{x_i})) \tag{2.86}$$

onde

$$\pi(\mathtt{x_i}) = P(Y = 1|X = \mathtt{x_i})$$

é a probabilidade de sucesso para Y dada a observação $\mathtt{x_i}$. De acordo com a definição de média e variância apresentadas acima para a distribuição binomial tem-se que para $X = \mathtt{x_i}$

$$E(Y|X = \mathtt{x_i}) = \mu = \pi(\mathtt{x_i})$$

e

$$Var(Y|X = \mathtt{x_i}) = \pi(\mathtt{x_i})(1 - \pi(\mathtt{x_i})).$$

Pretende-se modelar a média da resposta, isto é, as probabilidades $\pi(\mathtt{x_i})$ em função do vetor de variáveis explicativas $\mathtt{x_i}$. Uma solução passa por transformar a probabilidade $\pi_i = \pi(\mathtt{x_i})$ nos odds

$$odds_i = \frac{\pi_i}{1-\pi_i}, \tag{2.87}$$

que representam a razão entre a probabilidade de sucesso e a probabilidade de insucesso para a resposta. O facto de as probabilidades π_i variarem entre 0 e 1 implica que os odds variem entre 0 e $+\infty$. Por fim, aplica-se a transformação logarítmica à equação (2.87) e tem-se

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right), \tag{2.88}$$

que já varia entre $-\infty$ e $+\infty$. Note-se que quando a probabilidade π_i é próxima de zero, o valor do *logit* tende para $-\infty$. Já quando a probabilidade π_i é próxima de 1, o valor do odds aproxima-se de $+\infty$. Logits negativos correspondem a probabilidades inferiores a 0.5 e logits positivos a probabilidades superiores a 0.5.
O modelo de regressão logística é assim definido por

$$Y|X \quad = \quad \mathtt{x_i} \sim B(1, \pi(\mathtt{x_i})) \tag{2.89}$$

e

$$logit(\pi_i) \quad = \quad \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \tag{2.90}$$

A função *logit* é a função de ligação para o modelo linear generalizado com resposta binomial. Nesta caso, os coeficientes $\beta_j, j = 0, .., p$, representam a alteração no *logit* da probabilidade associada à mudança de uma unidade na $j-$ésima covariável, mantendo as restantes variáveis explicativas constantes.
Invertendo a função *logit* (2.90) obtém-se a média da resposta de Y, condicionada pelas variáveis explicativas X,

$$\begin{aligned} \pi(\mathtt{x}) &= P(Y = 1|X_1 = \mathtt{x_1}, ..., X_p = \mathtt{x_p}) \\ &= \frac{e^{\beta_0+\beta_1\mathtt{x_1}+...+\beta_p\mathtt{x_p}}}{1 + e^{\beta_0+\beta_1\mathtt{x_1}+...+\beta_p\mathtt{x_p}}} \\ &= \frac{1}{1 + e^{-(\beta_0+\beta_1\mathtt{x_1}+...+\beta_p\mathtt{x_p})}}. \end{aligned}$$

Para uma dada observação i, o valor $\widehat{\pi_i}$ é uma estimativa da média (proporção) de observações da população em causa tendo em conta as variáveis explicativas que apresentam evidência para o sucesso.

Interpretação dos parâmetros do modelo - Odds Ratio

Como descrito anteriormente, o odds corresponde ao quociente entre a probabilidade de sucesso π e a probabilidade de insucesso $1 - \pi$ de um determinado acontecimento,

$$odds = \frac{\pi}{1 - \pi}.$$

A título de exemplo, um odds de $\frac{1}{3}$ para um acontecimento significa que a probabilidade do acontecimento não ocorrer (insucesso) é três vezes maior do que a probabilidade do acontecimento ocorrer (sucesso).
Já um odds ratio, OR, corresponde a um quociente entre dois odds, calculados para dois indivíduos ou para dois grupos de indivíduos. Considere-se que Y é uma variável dicotómica que indica doença e E uma variável também dicotómica indicadora de exposição de um fator de risco. O OR da exposição para a doença é

$$OR \quad = \quad \frac{\frac{P(Y=1|E)}{1-P(Y=1|E)}}{\frac{P(Y=1|\overline{E})}{1-P(Y=1|\overline{E})}} = \frac{\frac{P(Y=1|E)}{P(Y=0|E)}}{\frac{P(Y=1|\overline{E})}{P(Y=0|\overline{E})}}, \tag{2.91}$$

que traduz o odds para a doença nos indivíduos expostos e o odds para a doença nos indivíduos não expostos. Por exemplo, se Y corresponde à presença de cancro do pulmão e E diz respeito aos indivíduos fumadores, um $\widehat{OR} = 2$ significa que, na população em estudo, o cancro do pulmão

é duas vezes superior nos indivíduos fumadores face aos não fumadores (Hosmer and Lemeshow, 2000).
Esta interpretação para o odds ratio é baseada no facto de este, em muitas situações, poder ser interpretado como uma quantidade que aproxima o risco relativo. Tendo em conta a definição de risco relativo e odds ratio tem-se que (Gaio, 2012)

$$\frac{OR}{RR} = \frac{1 - P(Y = 1|E = 0)}{P(Y = 1|E = 1)}$$

e portanto, o odds ratio aproxima o risco relativo se ambas as probabilidades P(Y=1|E=1) e P(Y=1|E=0) são baixas. Os odds ratio são uma importante medida de associação utilizada na regressão logística para a interpretação dos parâmetros de regressão β. Consideremos então que as variáveis explicativas $X_1, ..., X_p$ são fatores ou variáveis contínuas sem interações.
Comecemos por analisar o caso em que X_j toma valores **contínuos**. Aumentando a coordenada de j de $\mathtt{x} = (\mathtt{x_1}, ..., \mathtt{x_p})$ de uma unidade,

$$\mathtt{x}^{(+1)j} = (\mathtt{x_1}, ..., \mathtt{x_j} + \mathtt{1}, ..., \mathtt{x_p}),$$

tem-se, de acordo com a definição de OR na equação (2.91),

$$\begin{aligned}
\log(OR(\mathtt{x}^{(+1)_j}, \mathtt{x})) &= \log\left(\frac{odds\left(\mathtt{x}^{(+1)_j}\right)}{odds(\mathtt{x})}\right) \\
&= \log\left(\frac{\frac{\pi\left(\mathtt{x}^{(+1)j}\right)}{1-\pi\left(\mathtt{x}^{(+1)j}\right)}}{\frac{\pi(\mathtt{x})}{1-\pi(\mathtt{x})}}\right) \\
&= logit\left(\mathtt{x}^{(+1)_j}\right) - logit(\mathtt{x}) \\
&= (\mathtt{x_j} + \mathtt{1})\beta_j - \mathtt{x}_j\beta_j \\
&= \beta_j.
\end{aligned}$$

Logo,

$$OR\left(\mathtt{x}^{(+1)_j}, \mathtt{x}\right) = e^{\beta_j},$$

onde e^{β_j} corresponde ao odds ratio para o sucesso por aumento de uma unidade em X_j, mantendo as restantes variáveis explicativas constantes. Este raciocínio pode ser generalizado caso consideremos uma aumento de c unidades na coordenada j de $\mathtt{x} = (\mathtt{x_1}, ..., \mathtt{x_p})$. Neste caso tem-se, de forma análoga,

$$OR\left(\mathtt{x}^{(+c)_j}, \mathtt{x}\right) = e^{c\beta_j}, \tag{2.92}$$

o que significa que, para cada aumento de c unidades em X_j, o risco de sucesso (Y=1) varia de $e^{c\beta_j}$ vezes, mantendo as restantes variáveis explicativas constantes.
Considerando agora duas observações, $\mathtt{x_1} = (\mathtt{x_{11}}, ..., \mathtt{x_{1p}})$ e $\mathtt{x_2} = (\mathtt{x_{21}}, ..., \mathtt{x_{2p}})$, pode escrever-se que

$$\log(OR(\mathtt{x_1}, \mathtt{x_2})) = logit(\mathtt{x_1}) - logit(\mathtt{x_2}) = (\tilde{\mathtt{x}}_1^T - \tilde{\mathtt{x}}_2^T)\beta$$

e portanto

$$\begin{aligned} OR(\mathtt{x_1}, \mathtt{x_2}) &= e^{(\tilde{\mathtt{x}}_1^T - \tilde{\mathtt{x}}_2^T)\beta} \\ &= e^{\sum_{j=0}^{p}(\mathtt{x_{1j}} - \mathtt{x_{2j}})\beta_j}. \end{aligned}$$

Na situação em que X_j toma valores **dicotómicos** e supondo que as categorias desta variável estão codificadas em 0 e 1

$$\begin{aligned} \log(OR) &= \log(OR(X_j = 1, X_j = 0)) \\ &= \log\left(\frac{\frac{\pi(X_j=1)}{1-\pi(X_j=1)}}{\frac{\pi(X_j=0)}{1-\pi(X_j=0)}}\right) \\ &= logit(X_j = 1) - logit(X_j = 0) \\ &= \beta_0 + ... + \beta_{j-1}\mathtt{x}_{j-1} + \beta_j + \beta_{j+1}\mathtt{x}_{j+1} + ... + \beta_p\mathtt{x}_p \\ &\quad -\beta_0 - ... - \beta_{j-1}\mathtt{x}_{j-1} - \beta_{j+1}\mathtt{x}_{j+1} - ... - \beta_p\mathtt{x}_p \\ &= \beta_j \end{aligned}$$

e portanto

$$OR = e^{\beta_j}.$$

Deste resultado retira-se que, na população em estudo, é mais (ou menos) provável e^{β_j} vezes que a ocorrência do sucesso (Y=1) ocorra nos indivíduos com X=1 do que nos indivíduos com X=0, considerando as restantes covariáveis constantes.
Por fim, quando X_j toma valores **politómicos**, estamos perante a generalização da situação em que X_j é dicotómica. A variável politómica com C categorias é representado por $C-1$ variáveis dummy, sendo uma das categorias a categoria de referência, designada por 0. O modelo de regressão logística estima odds ratio para cada uma das classes 1,2,...,C-1, usando a classe 0 como referência.

A constante β_0 do modelo de regressão corresponde ao log-odds de um indivíduo onde as variáveis explicativas são todas nulas. De forma equivalente, e^{β_0} é interpretada como o odds para o sucesso na ausência de variáveis explicativas.

Na situação em que temos uma regressão simples, isto é, que apenas possui uma variável explicativa, o odds ratio obtido diz-se *odds ratio bruto*. Se estivermos na presença de uma regressão multivariada (existe mais do que uma variável explicativa) então o odds ratio diz-se *odds ratio ajustado*.

Resumidamente, os odds ratio podem ser interpretados da seguinte forma:

- OR>1 indica que existe associação positiva entre a probabilidade de sucesso e a exposição
- OR próximo de 1 indica que a probabilidade de sucesso e insucesso para uma determinada exposição são iguais
- 0<OR<1 indica que existe associação negativa entre a probabilidade de sucesso e a exposição.

Intervalos de confiança para o odds ratio

O intervalo com $100(1-\alpha)\%$ de confiança para o odds ratio $OR(x^{(+1)_j}, x)$ é

$$exp\left(\beta_j \pm N_{1-\alpha/2} se(\widehat{\beta}_j)\right),$$

onde $N_{1-\alpha/2}$ representa o $(1-\alpha/2)-$quantil da distribuição normal reduzida, $N(0,1)$.
O odds ratio estimado pelo modelo será estatisticamente significativo sempre que o correspondente intervalo de confiança não contiver o valor 1.

Estimação pelo método da máxima verosimilhança

Para n observações independentes da distribuição Binomial, resulta que a função de verosimilhança é

$$L(\beta) = \prod_{i=1}^{n} \pi(\mathtt{x}_i)^{y_i}(1-\pi(\mathtt{x}_i))^{1-y_i}$$

e portanto, o logaritmo da função de verosimilhança para um modelo de regressão logística, excluindo o termo combinatório, traduz-se em

$$\log(L(\beta)) = \sum_{i=1}^{n} \left[y_i \log(\pi(\mathtt{x}_i)) + (1-y_i)\log(1-\pi(\mathtt{x}_i)) \right].$$

Conforme já referido, em todos os modelos lineares generalizados com termo constante e função de ligação canónica, à exceção do modelo com função de ligação *probit*, a soma dos valores observados é igual à soma dos valores previsto pelo modelo. Assim, resulta que

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\pi}(\mathtt{x}_i).$$

Qualidade do ajustamento

Uma medida de discrepância entre os valores observados e ajustados é a desviância (secção 2.3.4.1). Tendo em conta que no modelo saturado se tem $\pi_i = y_i$ e no modelo em causa $\pi_i = \widehat{\mu}_i$ ($\widehat{\mu}_i$ é o valor ajustado da observação i fornecido pelo modelo em causa), a desviância no modelo de regressão logística é expressa da seguinte forma

$$D(y, \widehat{\mu}) = 2\sum_{i=1}^{n} \left(y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (1-y_i)\log\left(\frac{1-y_i}{1-\widehat{\mu}_i}\right)\right). \tag{2.93}$$

Uma medida alternativa da avaliação da qualidade do ajustamento é a estatística χ^2 de Pearson, que para dados binomiais é (Rodríguez, 2007)

$$X_P^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i(1-\widehat{\mu}_i)}.$$

Capítulo 3

Modelos de contagem com um número excessivo de zeros

Em muitas situações, dados de contagem existentes na vida real são caraterizados por um número excessivo de zeros e pela presença de sobredispersão. Quando este número de zeros é tão elevado que os dados não podem ser ajustados a uma distribuição de Poisson ou Binomial Negativa (distribuições usualmente consideradas em dados de contagem) para a sua modelação, terão de ser considerados modelos de contagem que lidam com um número excessivo de zeros.

Neste capítulo serão abordados modelos com zeros inflacionados e modelos com barreira. Estes modelos surgem, assim, com o intuito de modelar o excesso de zeros e, consequentemente, sobredispersão existente numa grande parte de dados de contagens reais. As suas aplicações estendem-se a diversas áreas tais como: ecologia, modelando as presenças de determinada espécie num determinado habitat (Martin et al., 2005); produção, análise do número de defeitos num processo de fabricação (Lambert, 1992); medicina (saúde pública), análise do número de crianças com cáries (Böhning et al., 1997); entre outras.

3.1 Modelos com zeros inflacionados

Os modelos com zeros inflacionados aqui descritos terão em conta a distribuição de Poisson e a distribuição Binomial Negativa, uma vez que estas são as distribuições mais utilizadas em dados de contagem (Cameron and Trivedi, 1998).
Os modelos que contêm o termo "zeros inflacionados"na sua designação modelam as contagens como uma mistura de duas distribuições com dois processos subjacentes: (i) um processo que trata do excesso de zeros, modelado por uma massa pontual em zero; e (ii) um processo que trata das contagens (incluindo algumas contagens nulas), modelado por uma distribuição de Poisson ou por uma distribuição Binomial Negativa (Lambert, 1992). O modelo assume que o primeiro processo ocorre com uma probabilidade π_i, enquanto que o segundo processo ocorre com uma probabilidade $1-\pi_i$ e gera uma contagem a partir de um modelo de Poisson com média μ_i ou a partir de um modelo Binomial Negativa (μ_i, α), (Lam et al., 2006). Em particular, considere-se que as variáveis resposta, $Y = (Y_1, ..., Y_n)^T$, onde n é o número de observações, são independentes. Para cada variável Y_i, existem dois processos possíveis para cada modelo considerado. Em suma:

$$Y_i \sim \begin{cases} 0 & \text{com probabilidade } \pi_i \\ Poisson(\mu_i) \quad \text{ou} \quad BinomialNegativa(\mu_i, \alpha) & \text{com probabilidade } 1-\pi_i, \end{cases}$$

onde π_i corresponde à probabilidade de existir um zero falso.

Entende-se por zeros falsos os zeros que não derivam de uma contagem de Poisson ou Binomial Negativa e por zeros verdadeiros as contagens nulas que estão sujeitas a uma distribuição de Poisson ou Binomial Negativa (Zuur et al., 2009). A figura seguinte ilustra como se comportam os modelos com zeros inflacionados.

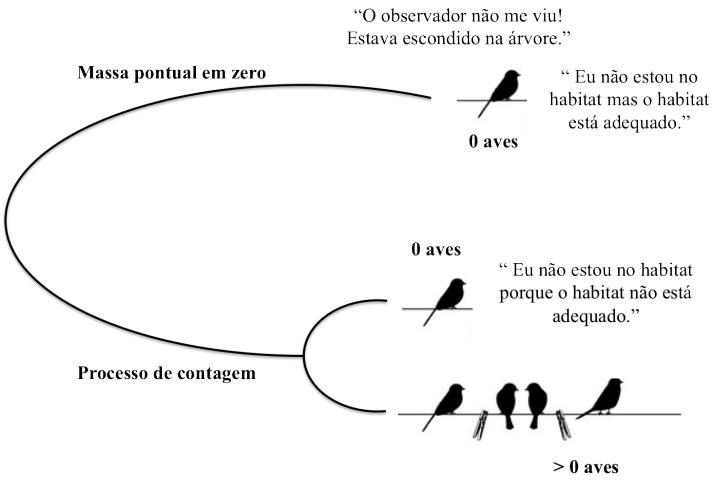


Figura 3.1: Zeros falsos *vs* zeros verdadeiros nos modelos com zeros inflacionados (Figura inspirada em Zuur et al. (2009)).

Suponha-se que se pretende analisar a presença de determinada espécie de aves num habitat. De acordo com a figura 3.1, considerar-se-ão como zeros falsos as aves que estão presentes no habitat mas que não se conseguem observar (as aves podem estar camufladas ou, no momento da amostragem terem saído à procura de alimento, por exemplo) - zeros que derivam de uma massa pontual em zero. Já os zeros verdadeiros correspondem às aves que não se encontram no habitat porque migraram, por exemplo - zeros que derivam de um processo de contagem.
Quando consideramos como exemplo dados acerca do número de artigos publicados pelos estudantes nos últimos 3 anos de doutoramento (Rodríguez, 2007), as contagens nulas correspondem à não publicação de artigos. Assim, os zeros falsos equivalem àqueles estudantes para quem publicar não é relevante porque, por exemplo, pretendem entrar no mundo do trabalho e não estão interessados em publicar. As não publicações dos estudantes que consideram as mesmas importantes mas não conseguem publicar por determinado motivo referem-se aos zeros verdadeiros.

3.1.1 Modelo de regressão de Poisson com zeros inflacionados

O modelo de regressão de Poisson com zeros inflacionados foi referenciado originalmente na literatura da econometria por Mullahy (1986). Como referido anteriormente, este modelo pretende modelar o excesso de zeros presente nos dados. Para uma amostra com n observações, a variável resposta é traduzida por Y_i, $i = 1, 2, ..., n$ e o vetor das covariáveis em estudo é dado por $\mathtt{X} = (X_1, ..., X_p)$. Considerando $y_i, ..., y_n$ realizações da variável resposta Y_i, o modelo de zeros inflacionados de Poisson assume que (Cheung, 2002):

$$P(Y_i = y_i|\mathtt{x_i}) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i} & \text{para } y_i = 0 \\ (1-\pi_i)\dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & \text{para } y_i \geq 1, \end{cases} \tag{3.1}$$

onde $\mu_i = \mu(\mathtt{x_i})$.
Observações com $y_i = 0$ provêm de dois grupos (Cheung, 2002; Lewsey and Thomson, 2004): o primeiro não está sujeito a um processo de Poisson e o segundo faz parte de uma distribuição de Poisson com média μ_i mas que apenas toma o valor de zero, dado que $\frac{e^{-\mu_i}\mu_i^0}{0!} = e^{-\mu_i}$. É de realçar que quando se observa uma contagem nula não se sabe de que grupo esta contagem é originária.
Tendo em conta a definição de valor esperado de uma variável aleatória e sabendo que $Var(Y) = E(Y^2) - E(Y)^2$, a média e a variância de uma variável aleatória que segue uma distribuição de Poisson com zeros inflacionados são dadas (Zuur et al., 2009), respetivamente, por :

$$E(Y_i) = \mu_i \times (1 - \pi_i) \tag{3.2}$$

$$Var(Y_i) = (1 - \pi_i) \times (\mu_i + \pi_i \times \mu_i^2). \tag{3.3}$$

Um modelo de regressão de Poisson com zeros inflacionados (modelos ZIP− *Zero Inflated Poisson models*) modela a média μ de uma variável de Poisson através de uma regressão de Poisson e a probabilidade π de existir um zero falso através de uma regressão logística com função de ligação *logit*. Mais precisamente, os parâmetros $\mu_i = \mu(\mathtt{x_i})$ e $\pi_i = \pi(\mathtt{x_i}), i = 1, ..., n$ satisfazem (Lambert, 1992; Ridout et al., 1998):

$$\log(\mu_i) = X_i\beta \tag{3.4}$$

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = Z_i\gamma, \tag{3.5}$$

onde X e Z são as matrizes das observações das covariáveis de interesse e β e γ (vetores) são os parâmetros de regressão. Nestas duas equações de regressão, as duas matrizes de covariáveis podem ou não coincidir (Lambert, 1992; Ridout et al., 1998). Quando estas são coincidentes e os parâmetros μ e π não estão relacionados, o modelo de regressão de Poisson com zeros inflacionados requer o dobro dos parâmetros de uma regressão de Poisson. Quando a probabilidade π não depende de covariáveis, a matriz Z reduz-se a uma coluna de 1's e o modelo de de Poisson com zeros inflacionados requer apenas mais um parâmetro do que a regressão de Poisson. Porém, existem poucas aplicações com informação prévia acerca de como π e μ estão relacionados. Numa tentativa de resolver este problema, Lambert (1992) sugeriu a seguinte parametrização:

$$\log(\mu_i) = X_i\beta \tag{3.6}$$

$$logit(\pi_i) = -\tau X_i\beta, \tag{3.7}$$

onde τ é um parâmetro de escala desconhecido. Este modelo é denominado por modelo de regressão de Poisson com zeros inflacionados (modelo ZIP(τ)). Tem-se que $\pi_i = (1 + \mu_i^\tau)^{-1}$. De facto,

$$\begin{aligned} &\log\left(\frac{\pi_i}{1-\pi_i}\right) = -\tau X_i\beta \Leftrightarrow \pi_i = \frac{e^{-\tau X_i\beta}}{1+e^{-\tau X_i\beta}} \Leftrightarrow \\ &\pi_i = \frac{e^{-\tau\log(\mu_i)}}{1+e^{-\tau\log(\mu_i)}} \Leftrightarrow \pi_i = \frac{\mu_i^{-\tau}}{1+\mu_i^{-\tau}} = \frac{1}{\mu_i^\tau + 1} \end{aligned} \tag{3.8}$$

Estes autores também referem a possibilidade de utilização de outras funções de ligação para a regressão binomial, para além do *logit*, como por exemplo a função $\log - \log$ (definida por $\log(\log(\pi))$ e a função $\log - \log$ *do complementar* (definida por $\log(-\log(1-\pi))$.

3.1.2 Modelo de regressão Binomial Negativa com zeros inflacionados

O modelo de regressão Binomial Negativa com zeros inflacionados é utilizado quando os dados apresentam, para além de um excesso de zeros, presença de sobredispersão. Este modelo é análogo ao modelo de regressão de Poisson com zeros inflacionados, à exceção de que neste modelo, a variável Y_i segue uma distribuição Binomial Negativa com parâmetros (μ_i,α) com zeros inflacionados que tem como função de probabilidade (Ridout et al., 2001):

$$P(Y_i = y_i|\mathtt{x}_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\dfrac{1}{1+\alpha\mu}\right)^{\alpha^{-1}} & \text{para } y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\dfrac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\dfrac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} & \text{para } y_i \geq 1. \end{cases}$$

onde a constante não negativa α é o parâmetro de heterogeneidade, que se supõe não depender das covariáveis.
Pode-se mostrar (Ridout et al., 2001; Zuur et al., 2009) que a média e a variância de uma distribuição Binomial Negativa com zeros inflacionados é dada por:

$$E(Y_i) = \mu_i \times (1-\pi_i) \tag{3.9}$$

$$Var(Y_i) = (1-\pi_i) \times \mu_i \times (1+\pi_i\mu_i+\alpha\mu_i). \tag{3.10}$$

Como era previsível, quando α tende para zero a distribuição em causa reduz-se para uma distribuição de Poisson com zeros inflacionados.

3.2 Modelos com barreira

Os modelos com barreira são outro tipo de modelos utilizados para analisar dados de contagem com zeros inflacionados. O primeiro modelo com barreira foi proposto por Mullahy (1986) e posteriormente popularizado por Cameron and Trivedi (1986, 1998). Estes modelos tratam os dados como presença (contagens positivas) *vs* ausência (contagens nulas), onde os dados referentes às presenças são modelados por um processo de contagem. Sejam $Y_1, ..., Y_n$ as variáveis resposta de uma amostra com n observações, considere-se o seguinte processo binomial que determina quando a variável Y_i, $i = 1, 2, .., n$, toma o valor zero ou um valor positivo (McDowell, 2003):

$$P(Y_i = y_i) = \begin{cases} 1-\pi_i, & y_i = 0 \\ \pi_i, & y_i = 1, 2, 3, ..., \end{cases}$$

onde π_i é a probabilidade de se observar uma contagem não nula para a observação i.
O modelo com barreira é descrito como um modelo que consiste de duas partes (Zuur et al., 2009):

1. Os dados são considerados apenas como presenças (contagens não nulas) *vs* ausências (contagens nulas), não se fazendo distinção entre zeros falsos e zeros verdadeiros como nos modelos apresentados nas subsecções 3.1.1 e 3.1.2). A probabilidade de se observar uma contagem não nula é modelada através de uma regressão logística - modelo de resposta binária.

2. As contagens positivas (presenças) são modeladas, usualmente, pela distribuição de Poisson/Binomial Negativa truncada - modelo de contagem truncado.

A figura 3.2 ilustra como se pode interpretar um modelo com barreira.

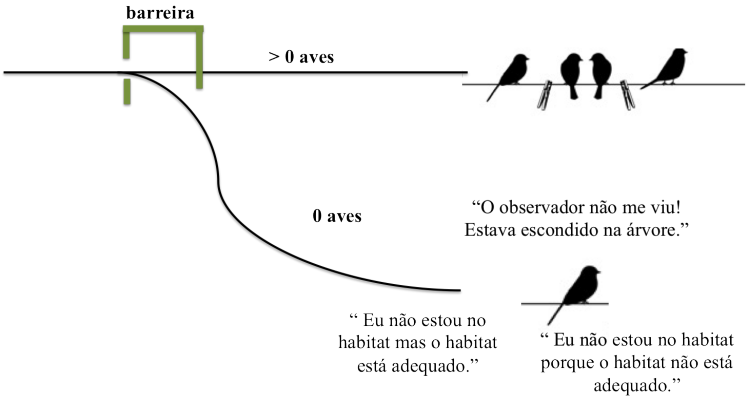


Figura 3.2: Modelo com barreira (Figura inspirada em Zuur et al. (2009)).

Formalmente, o modelo com barreira combina um modelo de contagem truncado com um modelo de barreira de zeros (Zeileis et al., 2008). O nome *barreira* surge do facto de contagens positivas terem de ultrapassar a barreira dos zeros, independentemente do mecanismo em causa. Eventualmente, a barreira pode estar num outro qualquer valor não nulo, dependendo do contexto do problema (Cameron and Trivedi, 1998).
Matematicamente, o modelo com barreira pode ser escrito da seguinte forma (Zeileis et al., 2008)

$$P_{barreira}(y;\mathtt{x},\mathtt{z},\beta,\gamma) = \begin{cases} P_{zero}(0;\mathtt{z},\gamma) & y=0 \\ (1-P_{zero}(0;\mathtt{z},\gamma)) * P_{contagem}(y;\mathtt{x},\beta)/(1-P_{contagem}(0;\mathtt{x},\beta)) & y>0, \end{cases}$$

onde $\mathtt{x}$ e $\mathtt{z}$ são os vetores das covariáveis de interesse e β e γ os parâmetros de regressão do modelo de contagem truncado e do modelo de zeros, respetivamente.

3.2.1 Modelo de regressão de Poisson com barreira

Conforme explicitado na secção 3.2, nos modelos com barreira, o processo de contagem não produz zeros. Em particular, considere-se que as contagens positivas são modeladas por uma distribuição de Poisson truncada. A função de probabilidade do modelo de regressão de Poisson com barreira é dada por (Ridout et al., 1998; Zuur et al., 2009):

$$P(Y_i = y_i|\mathtt{x}_i) = \begin{cases} 1-\pi_i & \text{para } y_i = 0 \\ \pi_i \frac{e^{-\mu_i}\mu_i^{y_i}}{(1-e^{-\mu_i})y_i!} & \text{para } y_i \geq 1. \end{cases} \tag{3.11}$$

onde π_i é a probabilidade de se observar uma contagem não nula. Como referido, para se conseguir obter uma contagem não nula é necessário atravessar a barreira do zero e, portanto, o processo de contagem de Poisson tem de excluir a probabilidade de existirem zeros. Nesta situação estamos na presença de uma distribuição de Poisson truncada. Assim, a segunda equação do sistema (3.11), traduz que a probabilidade de existir uma contagem não nula é igual à probabilidade de não existir um zero multiplicada pela função de probabilidade de uma distribuição de Poisson truncada. A distribuição de Poisson truncada é obtida dividindo a função de probabilidade de Poisson por 1 menos a função de probabilidade de Poisson quando $y_i = 0$.
A média e a variância de um modelo com barreira com distribuição de Poisson truncada é dada pelas expressões seguintes (Zuur et al., 2009)

$$E(Y_i) = \mu_i \times \frac{\pi_i}{1-e^{-\mu_i}} \tag{3.12}$$

$$Var(Y_i) = (\mu_i + \mu_i^2) \times \frac{\pi_i}{1 - e^{-\mu_i}} - \left(\frac{\pi_i}{1 - e^{-\mu_i}} \times \mu_i \right)^2. \tag{3.13}$$

O modelo de regressão de Poisson com barreira modela a média μ_i de uma variável de Poisson truncada através de uma regressão de Poisson e modela também a probabilidade de existir uma contagem (estritamente) positiva, π_i, através de uma regressão binomial com uma função de ligação adequada. Por exemplo Ridout et al. (1998) consideraram

$$\log(\mu_i) = X_i\beta \tag{3.14}$$

$$log(-log(1 - \pi_i)) = Z_i\gamma, \tag{3.15}$$

onde X e Z são as matrizes das observações das covariáveis de interesse, e os vetores β e γ são os parâmetros de regressão. Uma outra hipótese seria obviamente considerar um modelo de regressão logística para π_i.
Na situação em que as covariáveis são as mesmas, quando considerada uma regressão binomial com função de ligação $\log - \log$ complementar para π, um teste de hipóteses com hipótese nula $H_0 : \beta = \gamma$ testa se um modelo com barreira é necessário ou não (Zeileis et al., 2008); caso $\beta = \gamma$, o modelo em causa é reduzido ao modelo de Regressão de Poisson e portanto não é necessário considerar um modelo com barreira (Ridout et al., 1998).

Prova:

- Numa regressão de Poisson tem-se que:

$$\log(\mu) = X\beta \Rightarrow \mu = e^{X\beta}. \tag{3.16}$$

A probabilidade de uma observação ser nula na distribuição de Poisson é obtida através da equação (2.6), considerando $y = 0$,

$$P(Y = 0) = e^{-\mu}. \tag{3.17}$$

- Numa regressão binomial com função de ligação $\log - \log$ complementar para π tem-se

$$\log(-\log(1 - \pi)) = X\beta \Rightarrow 1 - \pi \quad = \quad e^{\left(-e^{X\beta}\right)} \tag{3.18}$$

Considerando y=0, a função densidade de probabilidade de uma binomial (eq. (2.8)) é

$$P(Y = 0) = 1 - \pi = e^{\left(-e^{X\beta}\right)}. \tag{3.19}$$

Como queríamos demonstrar, a probabilidade de uma observação ser nula na regressão de Poisson é igual à probabilidade de uma observação ser nula numa regressão binomial com função de ligação $\log - \log$ do complementar para π, considerando as mesmas covariáveis e caso $\beta = \gamma$. Isto significa que, neste caso particular, não é necessário existir uma barreira para que haja contagens positivas.

3.2.2 Modelo de regressão Binomial Negativa com barreira

Seja Y uma variável aleatória que segue uma distribuição Binominal Negativa de parâmetros μ e α. A função de probabilidade do modelo de regressão binomial negativa com barreira é expressa por (Hilbe, 2011; Zuur et al., 2009),

$$P(Y_i = y_i|\mathtt{x}_i) = \begin{cases} 1-\pi_i & \text{para } y_i = 0 \\ \pi_i \left[\frac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\right] / \left(1-\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\right) & \text{para } y_i \geq 1, \end{cases} \tag{3.20}$$

onde π_i é a probabilidade de se observar uma contagem não nula. Neste modelo, o processo de contagem é independente das contagens nulas e, consequentemente, a probabilidade de existir uma contagem não nula é modelada de acordo com a probabilidade de uma distribuição Binomial Negativa truncada (segunda equação do sistema (3.20)).
A média e a variância de um modelo de regressão Binomial Negativa com zeros inflacionados é dada, respetivamente, por (Zuur et al., 2009)

$$E(Y_i) = \mu_i \times \frac{\pi_i}{1-\left(\frac{1}{\alpha\mu_i+1}\right)^{\alpha^{-1}}} \tag{3.21}$$

$$Var(Y_i) = \frac{\pi_i}{1-\left(\frac{1}{\alpha\mu_i+1}\right)^{\alpha^{-1}}} \times (\mu_i^2 + \mu_i + \alpha\mu_i^2) - \left(\frac{\pi_i}{1-\left(\frac{1}{\alpha\mu_i+1}\right)^{\alpha^{-1}}}\right)^2. \tag{3.22}$$

Tal como no modelo de regressão de Poisson com barreira, o modelo de regressão Binomial Negativa com barreira, modela a média μ_i de uma variável de Poisson truncada através de uma regressão de Poisson e modela a probabilidade de se observar uma contagem não nula, π_i, através de uma regressão binomial com uma função de ligação adequada (*logit* ou $\log-\log$ complementar para π_i).

3.3 Inferência

3.3.1 Estimação dos parâmetros do modelo

A estimação dos parâmetros em ambos os modelos é baseada no método da máxima verosimilhança, tal como acontece nos modelos lineares generalizados. Consequentemente, algoritmos como o Newton-Raphson e o método iterativo dos mínimos quadrados reponderados podem ser considerados.

Modelo com zeros inflacionados

No modelo de zeros inflacionados, as contagens nulas podem ser originárias do processo de contagem ou de uma massa pontual em zero. Consequentemente, a estimação dos parâmetros do modelo tem de ser conjunta, isto é, a função de $\log-$verosimilhança envolve o ajustamento conjunto de duas regressões: regressão logística e regressão de Poisson ou Binomial Negativa.

- **Modelo de regressão de Poisson com zeros inflacionados - ZIP**

A função de log −verosimilhança para o modelo ZIP é dada por (Dalrymple et al., 2003; Lambert, 1992)

$$\begin{aligned} l_{ZIP} &= \sum_{y_i=0} \log\left[e^{\mathbf{z}_i^T\gamma} + e^{-e^{(\mathbf{x}_i^T\beta)}}\right] \\ &\quad + \sum_{y_i>0}(y_i\mathbf{x}_i^T\beta - e^{\mathbf{x_i}^T\beta}) - \sum_{i=1}^{n}\log(1+e^{\mathbf{z}_i^T\gamma}) \\ &\quad - \sum_{y_i>0}\log(y_i!). \end{aligned} \tag{3.23}$$

Lambert (1992) desenvolveu um algoritmo para a estimação dos parâmetros de regressão nos modelos ZIP, contudo, algoritmos como o Newton-Raphson e o método iterativo dos mínimos quadrados reponderados podem ser utilizados.

- **Modelo de regressão Binomial Negativa com zeros inflacionados - ZINB**

Nesta situação, a função log −verosimilhança é expressa por (Hilbe, 2011)

$$\begin{aligned} l_{ZINB} &= \sum_{y_i=0}\log\left[e^{\mathbf{z}_i^T\gamma} + \left(\frac{1}{1+\alpha e^{\mathbf{x}_i^T\beta}}\right)^{\alpha^{-1}}\right] \\ &\quad + \sum_{y_i>0}\mathbf{z}_i^T\gamma + \log(\Gamma(y_i+\alpha^{-1})) \\ &\quad - \log(yi!) - \log(\Gamma(\alpha^{-1})) + y_i\left[\log(\alpha e^{\mathbf{x}_i^T\beta}) - \log(1+\alpha e^{\mathbf{x}_i^T\beta})\right] \\ &\quad + \frac{1}{\alpha}\log\left(\frac{1}{1+\alpha e^{\mathbf{x}_i^T\beta}}\right) + \sum_{i=1}^{n} - \log(1+e^{\mathbf{z}_i^T\beta}) \end{aligned} \tag{3.24}$$

Modelo com barreira

Como mencionado anteriormente, nos modelos com barreira, o processo de contagem não pode produzir contagens nulas. Desta forma, existe um único processo que gera as contagens nulas. O facto de os zeros serem independentes das contagens não nulas, possibilita que a função de log −verosimilhança do modelo em causa possa ser estimada separadamente como o somatório de duas funções de log −verosimilhança distintas.

- **Modelo de regressão de Poisson com barreira**

Neste modelo, a função log −verosimilhança pode ser escrita da seguinte forma (Dalrymple et al., 2003)

$$\begin{aligned} l_{Poisson_{barreira}} &= \sum_{y_i>0}\mathbf{z}_i^T\gamma - \sum_{i=1}^{n}\log(1+e^{\mathbf{z}_i^T\gamma}) \\ &\quad + \sum_{y_i>0}\left[y_i\mathbf{x}_i^T\beta - e^{\mathbf{x}_i^T\beta} - \log(1-e^{-e^{\mathbf{x}_i^T\beta}}) - \log(y_i!)\right] \\ &= l(\mathbf{z}_i^T\gamma) + l(\mathbf{x}_i^T\beta), \end{aligned} \tag{3.25}$$

onde $l(\mathbf{z}_i^T\gamma)$ é a função log −verosimilhança baseada num modelo de regressão logística e $l(\mathbf{x}_i^T\beta)$ é a função log −verosimilhança baseada num modelo de regressão de Poisson truncado.

- **Modelo de regressão Binomial Negativa com barreira**

No modelo em questão, a função log −verosimilhança é expressa por (Hilbe, 2011)

$$\begin{aligned} l_{BinomialNegativa_{barreira}} &= \sum_{y_i>0} \mathtt{z}_i^T\gamma - \sum_{i=1}^{n}\log(1+e^{\mathtt{z}_i^T\gamma}) \\ &\quad + \sum_{y_i>0}\Bigg[\log(\Gamma(y_i+\alpha^{-1}) - \log(\Gamma(y_i+1)) - \log(\Gamma(\alpha^{-1})) - \frac{1}{\alpha}\log(1+\alpha\mathtt{x}_i^T\beta) \\ &\quad + y_i\log\left(\frac{\alpha e^{\mathtt{x}_i^T\beta}}{1+\alpha e^{\mathtt{x}_i^T\beta}}\right) - \log\left(1-\left(\frac{1}{1+\alpha e^{\mathtt{x}_i^T\beta}}\right)^{\alpha^{-1}}\right)\Bigg] \\ &= l(\mathtt{z}_i^T\gamma) + l(\mathtt{x}_i^T\beta), \end{aligned} \tag{3.26}$$

onde $l(\mathtt{z}_i^T\gamma)$ é a função log −verosimilhança baseada num modelo de regressão logística e $l(\mathtt{x}_i^T\beta)$ é a função log −verosimilhança baseada num modelo de regressão Binomial Negativa truncado.

3.3.2 Testes de hipóteses, análise de resíduos e seleção do modelo

Nos modelos com um número excessivo de zeros, a estimação dos parâmetros de regressão é baseada no método da máxima verosimilhança o que significa que, quando os modelos são encaixados, pode-se utilizar o teste de razão de verosimilhanças para comparar os dois modelos. Para testar a significância dos coeficientes de regressão recorre-se ao teste de Wald, à semelhança do que acontece nos modelos lineares generalizados.
A análise dos resíduos é feita, essencialmente, através da análise gráfica.
Para que se possa comparar dois modelos não encaixados, recorre-se ao *critério de Akaike (AIC)* ou ao *critério de informação Bayesiano (BIC)*. O modelo que apresentar um menor valor de AIC ou BIC é o modelo eleito.

3.4 Modelos com zeros inflacionados *versus* Modelos com barreira

Uma das principais diferenças entre os modelos com zeros inflacionados e os modelos com barreira apresentados reside na forma como estes modelos lidam com os tipos de zeros. Como visto anteriormente, os modelos com zeros inflacionados fazem distinção entre zeros falsos e zeros verdadeiros, enquanto que os modelos com barreira, apenas separam as contagens em presenças (valores não nulos) *versus* zeros (valores nulos). Para além disso, estes modelos diferem também ao nível do processo de contagem: o processo de contagem associado a modelos com zeros inflacionados pode produzir zeros, enquanto que o processo de contagem associado a modelos com barreira não pode produzir zeros uma vez que a distribuição presente neste modelo é truncada. Já a média da resposta no modelo com zeros inflacionados é modelada através de uma regressão de Poisson e no modelo com barreira é modelada através de uma regressão de Poisson truncada. Nos modelos com zeros inflacionados obtém-se a probabilidade de existir um zero falso e posteriormente é possível calcular a probabilidade de existir um zero. Já nos modelos com barreira obtém-se diretamente a probabilidade de existir uma contagem nula.
Ambos os modelos são modelos genéricos e adaptam-se facilmente a outro valor que não o nulo, isto é, adaptam-se também a observações em excesso de valores não nulos.
Como escreve Miller (2007), a escolha entre os dois modelos, modelo com zeros inflacionados e modelo com barreira, depende se o investigador acredita que os zeros são todos "reais", ou se

alguns dos zeros são erros aleatórios.
Na literatura acerca da inflação de zeros, os resultados de dados, tantos simulados como reais são, muitas vezes, contraditórios. Lambert (1992) concluiu que o modelo de regressão de Poisson com zeros inflacionados faz um melhor ajustamento aos dados do que o modelo de regressão Binomial Negativa, o qual se revelou superior ao modelo de regressão de Poisson. Por outro lado, Greene (1994) concluiu que o modelo de regressão binomial negativa é superior ao modelo de regressão de Poisson com zeros inflacionados e que este último, faz um melhor ajustamento do que o modelo de regressão de Poisson usual. No trabalho desenvolvido por Miller (2007), este autor refere que Slymen et al. (2006) descobriram que o modelo de regressão de Poisson com zeros inflacionados e o modelo de regressão Binomial Negativa com zeros inflacionados são equivalentes. Já Welsh et al. (1996) concluíram que o modelo de regressão de Poisson com barreira e o modelo de regressão de Poisson com zeros inflacionados são iguais, enquanto que Pardoe and Durham (2003) consideram que o modelo de regressão Binomial Negativa com zeros inflacionados é superior, tanto ao modelo de regressão de Poisson como aos modelos com barreira.
Os artigos acima descritos apresentam diferenças em termos da proporção de zeros existente nos dados, assim como na distribuição para as contagens não nulas e, portanto, é natural que modelos diferentes originem resultados distintos, dependendo das caraterísticas dos dados.

Capítulo 4

Projeto do CIBIO

O estudo em análise nesta tese surge na sequência do projeto "Estrutura verde urbana: Estudo da Relação entre a Morfologia do Espaço Público e a Diversidade de Flora e Fauna na cidade do Porto"(http://bio-diver-city.fc.up.pt) desenvolvido pelo CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, da Universidade do Porto. Pretende-se analisar o efeito de diferentes variáveis ambientais sobre a abundância de aves das ordens *Columbiformes* e *Passeriformes* nos espaços verdes de acesso público da cidade do Porto. Este projeto enquadra apenas áreas verdes de acesso público uma vez que a influência do homem sobre a biodiversidade será mais evidente, não só pela presença constante de pessoas nestes locais, decorrente de movimentos pendulares, de recreio ou manutenção dos espaços, mas também da organização espacial que o homem impõe à estrutura verde (Farinha-Marques et al., 2011a).
A cidade do Porto possui 95 áreas verdes de acesso público das quais, 74 estão incluídas na categoria de Parques e Jardins e as restantes 21 na categoria de Praças ajardinadas, como se pode visualizar na figura 4.1.

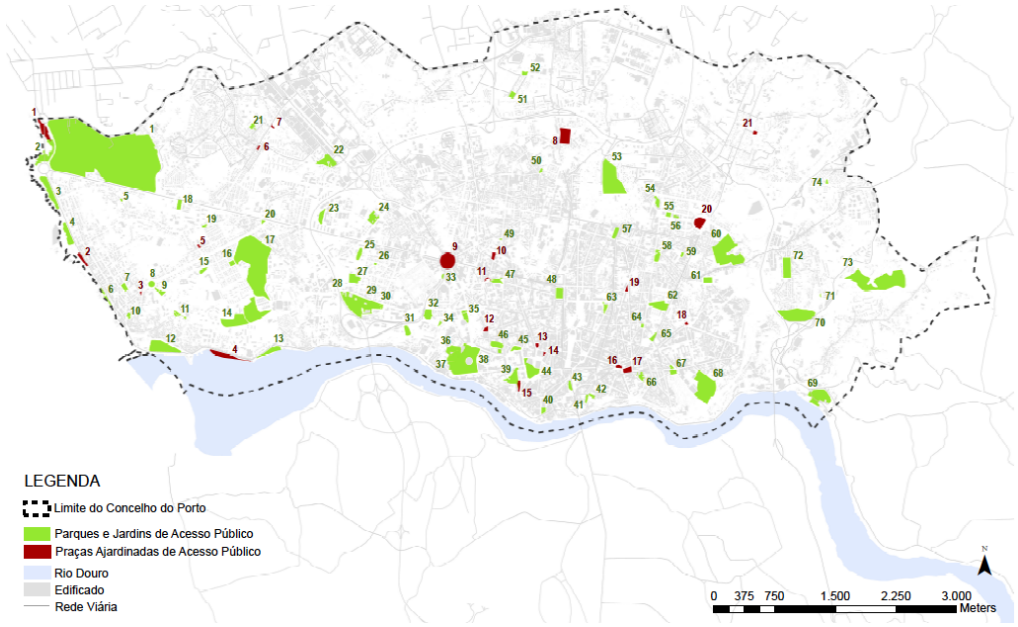


Figura 4.1: Carta de Parques, Jardins e Praças ajardinadas de acesso público da cidade do Porto (Fonte: Farinha-Marques et al., 2011).

Conforme referido por Farinha-Marques et al. (2011a), devido a constrangimentos de caráter logístico e de tempo, revelou-se fundamental a escolha de um grupo restrito de áreas a amostrar. Neste contexto, identificaram-se espaços representativos das áreas verdes do Porto, de acordo com as semelhanças existentes entre os vários espaços quanto a um conjunto selecionado de variáveis ambientais (Rita Gaio e Cláudia Fernandes- comunicação pessoal). Esse conjunto consistiu das variáveis: área total, área impermeável, área de coberto vegetal, presença/ausência de água, função dominante, idade e caráter do espaço. Um modelo de misturas finitas identificou 4 grupos de espaços verdes na totalidade dos espaços existentes, e as caraterísticas ambientais de cada grupo foram também analisadas. Detetaram-se ainda duas observações extremas (outliers) para as variáveis ambientais diretamente relacionadas com a área do jardim, Parque da Cidade e Parque de Serralves (Figura 4.2). A qualidade da classificação foi garantida por elevadas probabilidades de cada espaço verde pertencer ao seu grupo. Dentro de cada grupo foram selecionados alguns dos espaços verdes de acordo com as condições necessárias à realização das metodologias de levantamento de dados sobre flora e fauna e de acordo com a posição geográfica dos mesmos.

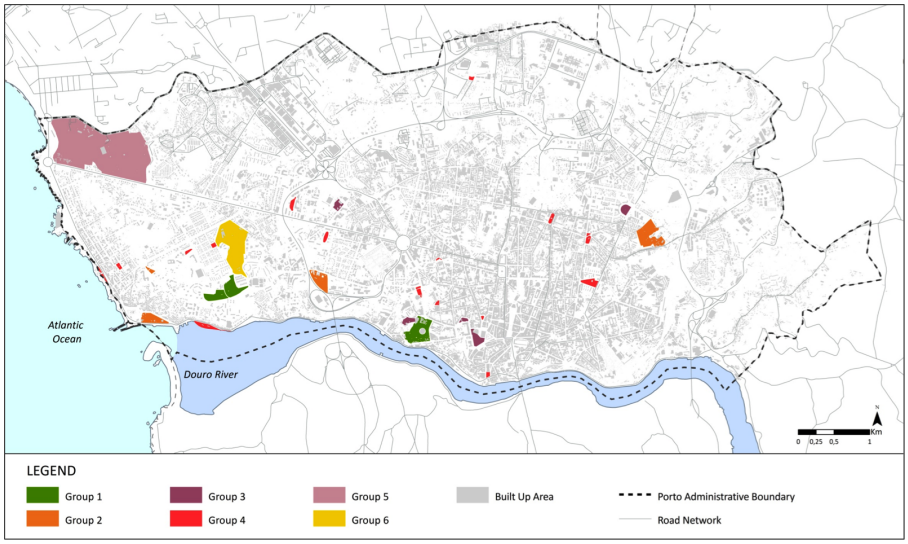


Figura 4.2: Grupos resultantes da análise de clusters (Fonte: CIBIO - comunicação pessoal).

Para o efeito, foram selecionados, numa primeira fase, 15 espaços verdes de acesso público, a saber:

- Jardim da Avenida Brasil
- Jardim de Bartolomeu Velho
- Jardim do Passeio Alegre
- Parque da Pasteleira
- Jardim de Alfredo Keil
- Jardim Machado de Assis
- Jardim Botânico
- Jardim de Sophia (Praça da Galiza)
- Jardim da Casa Tait

- Jardim do Palácio de Cristal
- Jardim da Cordoaria
- Jardim do Carregal
- Jardim da Praça Marquês de Pombal
- Jardim da Praça Rainha D. Amélia
- Jardim da Cantareira.

Os dados longitudinais estudados nesta dissertação foram recolhidos em duas épocas distintas. Na primeira época, a equipa do CIBIO conseguiu amostrar, apenas, 15 das 95 áreas verdes da cidade do Porto, devido a constrangimentos de caráter logístico e de tempo, como já referido. Estas 15 áreas verdes foram amostradas durante seis meses - abril, maio, junho, julho, dezembro de 2011 e janeiro de 2012 e correspondem aos jardins acima referidos. Na segunda época, a equipa do CIBIO amostrou outros jardins, num total de 14, que foram selecionados de acordo com o critério apresentado nesta secção. Estes jardins foram também amostrados durante seis meses - dezembro de 2011, janeiro, abril, maio, junho e julho de 2012 e correspondem às seguintes áreas verdes:

- Parque da Cidade
- Jardim da Praça de Liége
- Jardim do Largo de D. João III
- Jardim da Fundação de Serralves
- Fundação Engenheiro António de Almeida
- Jardim na Rua de Manuel Bandeira
- Jardim na Praça do Infante D. Henrique
- Jardim do Largo Palmira Milheiro
- Parque de S. Roque
- Jardim Paulo Vallada
- Jardim da Praça Pedro Nunes
- Largo da Maternidade Júlio Dinis
- Praça Carlos Alberto
- Praça Velasquez.

A figura 4.3 ilustra os instantes em que os jardins referentes às duas épocas em estudo foram amostrados.

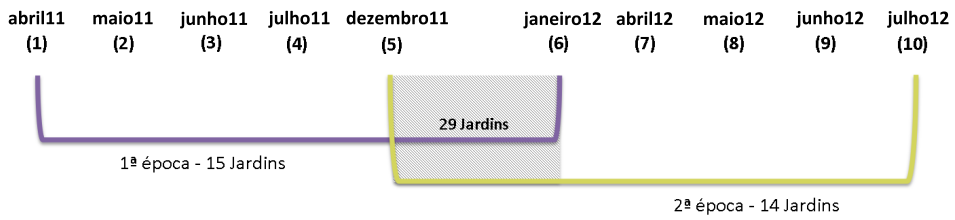


Figura 4.3: Esquema dos dados longitudinais.

Como é possível visualizar, os dados longitudinais em estudo possuem 10 tempos distintos. A equipa do CIBIO decidiu amostrar os jardins nos meses de Dezembro, Janeiro, Abril, Maio, Junho e Julho porque considera que estes meses são *"momentos chave"* para tentar captar o máximo de diversidade de espécies; as invernantes (no inverno), as residentes que se começam a reproduzir cedo (com picos de atividade no início da primavera) e as estivais reprodutoras que começam a reprodução mais tarde (picos no final da primavera e início do verão).
O esquema da figura 4.3 mostra a existência de dois meses comuns à 1ª e 2ª épocas, dezembro de 2011 e janeiro de 2012. Isto significa que, nestes meses, os jardins amostrados passam a ser 29. É importante realçar o facto de cada jardim ter sido amostrado seis vezes e, portanto, para cada jardim vão existir 4 meses em que estes não são amostrados, pelo que as suas observações, neste período, vão ser omissas. Apesar de a seleção dos jardins a serem amostrados nas duas épocas não ser aleatória, uma vez que foi efetuada uma análise de *clusters* e se teve em conta caraterísticas dos jardins para levantamento de dados sobre flora e fauna, a variável resposta não teve qualquer influência sobre a existência de valores omissos. Isto é, o facto de dispormos de jardins distintos nas duas épocas (apenas existem dois meses comuns) foi opção do CIBIO e nada teve a ver com o número de *Passeriformes* que foram observados.

Como referido, o presente estudo pretende inferir acerca do efeito de determinadas variáveis ambientais sobre a abundância de aves das ordens *Columbiformes* e *Passeriformes* nos jardins públicos da cidade do Porto. Para tal, a equipa do CIBIO considerou 7 variáveis ambientais. A tabela 4.1 carateriza cada uma destas variáveis, procede à sua descrição e identifica-as no software R.

Tabela 4.1: Tabela das variáveis ambientais.

Variável	**Descrição**	**Variável no R**
Área total (m^2)	área total do jardim ou praça ajardinada	at
Área de coberto vegetal (m^2)	área do jardim ou praça ajardinada que contém vegetação	avm
Área de coberto fanerófito (m^2)	área do jardim ou praça ajardinada que contém apenas árvores e arbustos	afm
Área permeável (m^2)	área do jardim ou praça ajardinada cuja superfície do solo é revestida por coberto vegetal ou terra	apm
Área de plano de água (m^2)	área de todos os elementos de água do jardim ou praça ajardinada: lagos, tanques, chafarizes, fontes e linhas de água a céu aberto	apa
Distância ao mar (m)	distância mínima entre o jardim ou praça ajardinada e a orla costeira (linha de fronteira do jardim e linha da costa)	dm
Distância ao rio (m)	distância mínima entre o jardim ou praça ajardinada e a linha que representa a margem direita do rio Douro	dr

4.1 Metodologias para censos e abundância de aves

Globalmente, o projeto CIBIO centra-se no estudo da biodiversidade nos jardins públicos da cidade do Porto. O termo *biodiversidade* (etimologicamente, do grego *biós*, vida e *diversidade*, variedade, multiplicidade) pode dar margem a inúmeras interpretações. A abordagem mais simples e direta do termo refere-se à universalidade dos seres vivos no planeta Terra e de todas as suas variações genéticas. A definição mais abrangente e preferencialmente adotada tem por base as United Nations (1992) que define a *biodiversidade* como "a variabilidade dos organismos vivos de toda a origem, compreendendo, entre outros, os ecossistemas terrestres, marinhos e outros ecossistemas aquáticos e os complexos ecológicos de que fazem parte; compreende ainda a diversidade dentro de espécies, entre espécies e de ecossistemas".

Os espaços verdes existentes nas cidades e aglomerados populacionais permitem o acolhimento e a movimentação de espécies comuns e eventualmente, espécies raras. Conforme citado por Farinha-Marques et al. (2011a), as estruturas verdes são elementos paisagísticos fundamentais para a promoção e conservação da biodiversidade nas áreas urbanas. Dado o valor elevado dos serviços que prestam aos habitantes (desde atividades de lazer até ao fornecimento de alimentos), os jardins, praças e parques que circundam as cidades são geralmente muito apreciados. De acordo com Farinha-Marques et al. (2011b) em estudo original de Melles (2005) o comportamento e a atividade humana, tais como o cultivo e a urbanização influenciam a biodiversidade urbana. Consequentemente, pode afirmar-se que a biodiversidade urbana reflete a cultura humana.

Um importante conceito em ecologia diz respeito à noção de *abundância*, encarada como a quantidade relativa de espécies num determinado ecossistema. Já, as medidas de abundância, que são aproximadas pela contagem do número de indivíduos numa área de amostra, são utilizadas

para indicar a "saúde"da população. Deste modo, este conceito tem um papel central em muitas teorias ecológicas e as técnicas de recolha utilizadas em organismos biológicos são diversas, desde dados sobre presença/ausência, até à estimação de abundância relativa, densidade ou tamanho da população. A dificuldade em obter valores de abundância que correspondam efetivamente à dimensão real das populações das aves é um facto incontestável (Rabaça, 1995). Na publicação de Rosenstock et al. (2002), é-nos referido que existem três tipo de fatores que, individualmente ou combinados com outros, podem influenciar a contagem/deteção aviária.
O primeiro, corresponde a variáveis que afetam a capacidade de o observador detetar e identificar corretamente a ave. Já, a performance do observador varia entre e dentro de indivíduos e é fortemente influenciada pelo treino, idade, experiência, motivação, capacidade auditiva e visual, saúde mental e nível de fadiga (Cyr, 1981; Kepler and Scott, 1981; Ramsey and Scott, 1981; Sauer et al., 1994).
O segundo fator decorre de variáveis ambientais que afetam o comportamento da ave e a eficiência do observador tais como, a velocidade do vento, temperatura, intensidade da luz (Verner, 1985) bem como a topografia e caraterísticas da vegetação (Dawson, 1981).
O terceiro fator que afeta a capacidade de deteção da ave diz respeito a atributos físicos e comportamentais das aves que podem tornar-se confusos para o observador. Exemplos destes atributos são o tamanho do corpo, a coloração das plumagens, caraterísticas vocais, comportamento de voo, entre outros (Cohen et al., 1960; Sayre et al., 1978; Wilson and Bart, 1985).
Existe uma grande variedade de métodos de censo utilizados na contagem de aves. Os métodos de censo de aves correspondem a estratégias de aplicação de técnicas de recolha de dados devidamente padronizadas, visando a obtenção, tanto quanto possível, de estimativas precisas e exatas. Dependendo do objetivo e das questões a que o estudo pretende responder, os métodos de censo de aves são habitualmente classificados em dois grupos (Rabaça, 1995; Rosenstock et al., 2002):

1. **métodos relativos** - usam contagens de aves como um índice para abundância relativa, sendo os resultados normalmente expressos em número de aves observadas por unidade de tempo ou distância;

2. **métodos absolutos** - estimam diretamente a densidade das aves, sendo os resultados usualmente expressos em número de indivíduos numa determinada superfície.

Centrados apenas no objetivo do estudo que passa por inferir acerca da abundância das aves, os exemplos de métodos de censos relativos são (Rabaça, 1995; Rosenstock et al., 2002; Verner, 1985):

- os *transetos lineares*, nos quais o observador percorre, com uma velocidade sensivelmente constante, um trajeto de comprimentos conhecido, registando todas as aves detetadas durante o percurso.

- os *métodos pontuais*, no qual a estratégia de recolha de dados se baseia no registo das aves detetadas pelo observador em determinados pontos da área de amostragem, durante um período de tempo previamente estabelecido.

- os *métodos de captura-recaptura* que consistem em obter uma amostra inicial de indivíduos que serão marcados ou identificados, dependendo da necessidade e do habitat da população em estudo, sendo, de seguida, devolvidos à população. Posteriormente é retirada uma segunda amostra, independente da primeira, na qual se contabilizam os indivíduos marcados da primeira amostra. Este método assume que, no caso da segunda amostra ser representativa da população total, a proporção de indivíduos marcados será aproximadamente igual à proporção dos indivíduos marcados na população total. A estimativa do número total de indivíduos na população é retirada da relação existente entre estas proporções (Oliveira, 2007).

É importante referir que nos transetos lineares e nos métodos pontuais, a deteção das aves é feita pela observação direta e pelo canto, enquanto que no método de captura-recaptura, como o próprio nome indica, a deteção não é feita através da observação direta das aves nem pelo canto, mas sim pela captura dos indivíduos e marcação dos mesmos. Dentro dos métodos apresentados, o mais comum na amostragem de aves é o método pontual, tal como referido por Rosenstock et al. (2002). Depois de serem recolhidas as estimativas de abundância, é importante analisá-las. Metodologias para a análise de dados de abundância são, muitas vezes, meramente descritivas, como é o caso do ranking de espécies. Esta metodologia pretende estudar a performance dos dados amostrados permitindo descrever qual o risco de extinção de determinada ave. As publicações de Linsdale (1924) e White (1942) utilizaram o ranking de espécies nos seus estudos.
A análise Bayesiana é também uma abordagem comum nos estudos ecológicos e ambientais (Bermudez and Bispo, 2012). Forcey et al. (2011) estudaram a influência do terreno e do clima sobre as aves aquáticas em seis estados dos Estados Unidos. Estes autores avaliaram variáveis de condição do terreno em três escalas logarítmicas diferentes e construíram um modelo de contagem espacial hierárquico *a priori*, usando a informação já publicada. O modelo foi ajustado através da estatística Bayesiana, recorrendo ao método Monte Carlo via Cadeias de Markov. Do estudo resultou que a vegetação, a área de pântano nas três escalas consideradas e as variáveis de precipitação influenciam positivamente a abundância das aves aquáticas. Um exemplo geral acerca da análise Bayesiana utilizando o método Monte Carlo via Cadeias de Markov na abundância animal é reportado por Brooks (1999). Esta publicação apresenta uma comparação entre a análise Bayesiana e a análise por máxima verosimilhança.
Uma outra metodologia utilizada na análise de abundância são os modelos lineares generalizados (Bermudez and Bispo, 2012). De um modo geral, em virtude de os dados de abundância, na sua grande maioria, serem apresentados como forma de contagem, isto é, como o número de aves observadas num determinado período de tempo, as distribuições mais comuns para modelar estes dados são a distribuição de Poisson, a distribuição Quasi-Poisson, e a distribuição Binomial Negativa (Hoef and Boveng, 2007; Nur et al., 1999; White and Bennetts, 1996). Seavy et al. (2005) recorreram à utilização destes modelos para estudarem a variação da abundância em cinco aves diferentes tendo em conta o habitat em que foram amostradas. Ambas as aves foram amostradas em 78 locais distintos e para cada um destes locais está associada a variável habitat (variável dicotómica) que indica se o local possui pouca (assume o valor 0) ou muita vegetação (assume o valor 1), de acordo com pressupostos definidos. Os autores propuseram um modelo de regressão de Poisson para interpretar o efeito do habitat sobre a abundância nas aves em estudo. Em ambos os modelos para as diversas espécies, a variável habitat revelou ser significativa. Através da estimação do parâmetro de dispersão, os autores concluiram que existia sobredispersão em 4 das 5 espécies em análise. Para estas 4 espécies, foi ajustado um modelo de regressão Binomial Negativa. Após o ajustamento, a variável habitat revelou continuar a ser significativa apenas para 1 das 4 espécies. Os autores deste estudo alertam ainda para o facto de o valor dos parâmetros estimados pelos dois modelos de regressão serem iguais, apenas varia o erro padrão. O modelo de regressão binomial negativa apresenta um erro padrão superior dado que considera a variância adicional existente. Os investigadores deste artigo reconhecem também que os modelos lineares generalizados são um método poderoso na análise de abundância e que a escolha do modelo de regressão depende dos dados em análise.
Contudo, uma caraterística comum aos dados ecológicos de contagens é a tendência para conterem muitos valores exatamente iguais a zero. Quando este número de zeros é tão elevado que os dados não podem ser ajustados a nenhuma distribuição standard, para a sua modelação terão de ser considerados modelos de contagem com um número excessivo de zeros. Martin et al. (2005) apresentam um estudo onde pretendem identificar qual o efeito do nível de alimento (baixo, médio, alto) sobre a abundância de 4 espécies diferentes. Estes autores ajustaram os dados em análise a

modelos de regressão de Poisson e Binomial Negativa e também a modelos de regressão de Poisson e Binomial Negativa com zeros inflacionados, devido ao número de zeros presentes nos dados. Dos resultados obtidos, concluiram que o modelo de regressão de Poisson foi o que evidenciou pior ajustamento. O modelo de regressão de Poisson com zeros inflacionados foi o modelo que melhor se ajustou aos dados da ave que continha um maior número de zeros. O modelo de regressão binomial negativa foi o modelo eleito para uma das aves em estudo, enquanto que o modelo de regressão binomial negativa com zeros inflacionados foi o que assumiu um melhor ajustamento aos dados das restantes duas espécies.

Joseph et al. (2009) aplicaram os quatro modelos referidos num estudo sobre seis aves do sul da Austrália. Diversas variáveis ambientais foram utilizadas como covariáveis tanto no modelo de contagem como no modelo de zeros. Estes autores demonstraram que os modelos que consideram a distribuição binomial negativa produzem sempre estimativas ecologicamente irrealistas para os parâmetros. Os modelos de zeros inflacionados são preferíveis, pois modelam um mecanismo ecológico em vez de um fenómeno estatístico gerando estimativas de parâmetros razoáveis. Os autores realçam ainda que é importante perceber-se quais as fontes que provocam variação nos dados e utilizar esta informação para escolher a distribuição apropriada. Outros exemplos de aplicação de modelos com um número excessivo de zeros em dados ornitológicos constam na publicação de Kéry (2008) e Royle et al. (2005).

Quando a abundância de aves é amostrada ao longo de diversos tempos distintos, os dados em análise são considerados longitudinais e, recorre-se, usualmente, à utilização de modelos lineares generalizados mistos e modelos marginais para a sua modelação.

Um estudo longitudinal sobre a abundância de aves onde o modelo escolhido para fazer o ajustamento aos dados foi o modelo linear generalizado misto é apresentado por Tvardíková (2010). A amostragem foi efetuada durante 10 dias do mês de julho de 2008 e a autora propôs-se a estudar a eventual diferença na abundância de espécies de aves entre dois tipos de florestas, primária e secundária. Os tipos de floresta, o período do dia em que a amostragem foi feita, o grupo de observadores e o método de amostragem implementado foram considerados como efeitos fixos do modelo. Os transetos, exemplo de um método de censo de aves, foi tratado como efeito aleatório e o logit foi a função de ligação escolhida. Da análise do estudo, a autora não encontrou diferenças significativas na abundância de aves entre os tipos de floresta. Já Edwards et al. (2010) investigaram o impacto dos tipos de habitat (plantações de óleo de palma, fragmento florestal, ou floresta contígua) e da área fragmentada na abundância de aves ao longo de 5 meses e em 22 locais distintos. A abundância foi modelada como uma variável que segue uma distribuição de Poisson, com função de ligação log. As diferenças entre os locais, dias e pontos de amostragem foram considerados efeitos aleatórios do modelo. Os fragmentos florestais revelaram não aumentar a abundância de aves, ao contrário das plantações de óleo de palma. Um exemplo de um modelo marginal para contagens (com função de ligação log) aplicado a dados ornitológicos foi estudado por Luther et al. (2008). Estes autores consideraram 18 locais para a amostragem e a duração do levantamento de dados foi de 4 meses, nos anos de 2001 e 2002. O estudo incide sobre qual o efeito de diversas variáveis de habitat e de paisagem sobre a abundância de aves no norte da Califórnia. A abundância de aves mostrou ser altamente correlacionada com todas as variáveis de paisagem consideradas, assim como com as variáveis ambientais locais.

Do levantamento metodológico é possível retirar que os dados de abundância não são fáceis de modelar e que existem, de facto, estudos contraditórios acerca da mesma metodologia. Não se pode assumir, portanto, que existe um modelo de eleição antes de se ajustarem diversos modelos aos dados em investigação.

Uma fator importante na análise da abundância das aves é perceber quais as variáveis ambientais que afetam a sua variação. Na descrição metodológica algumas variáveis foram já referidas, contudo, existem outras que são também relevantes. No artigo de revisão de Farinha-Marques

et al. (2011b), a fragmentação do habitat e a heterogeneidade do terreno apresentam um elevado impacto na biodiversidade. O isolamento do fragmento tem uma maior influência negativa na diversidade de espécies de aves ao longo do tempo do que o tamanho do fragmento (Crooks et al., 2001). Segundo Mörtberg (2001), a distância ao habitat natural ou a outra área verde irá afetar a taxa de sucesso de movimentos dispersos e a proximidade a estas áreas irá promover a abundância de aves. O uso de espécies nativas em áreas verdes promoverá também comunidades mais complexas de aves tal como referem White et al. (2005). De um modo geral, áreas muito urbanizadas possuem baixa diversidade de espécies de aves (Mörtberg, 2001), enquanto que, níveis intermediários de urbanização mostram beneficiar as aves (Palomino and Carrascal, 2007; Posa and Sodhi, 2006). Garden et al. (2007) sustentam que a presença de árvores decrépitas ou caídas influenciam positivamente a abundância e diversidade de certos grupos de fauna, tais como as aves. Para além disso, vegetação alta (árvores) e arbustos aumentam a presença das espécies de aves (Daniels and Kirkpatrick, 2006; Green and Baker, 2003; Posa and Sodhi, 2006), à exceção de aves adaptadas ao meio urbano que beneficiam de áreas com cobertura arbórea baixa (Palomino and Carrascal, 2007; Sandström et al., 2006). Como cita Brown and Freitas (2002) a presença de elementos de água traduz-se num efeito positivo na biodiversidade urbana.
Farinha-Marques et al. (2011b) acrescentam ainda que as aves são consideradas um bom grupo para avaliar a biodiversidade urbana e a qualidade ambiental, dado que estas são bastante sensíveis a mudanças de habitat.

Capítulo 5

Resultados – Parte 1

Neste capítulo pretende-se aplicar os modelos lineares generalizados e os modelos com um número excessivo de zeros aos dados disponibilizados pelo CIBIO-UP, com o objetivo de estudar o efeito de diferentes variáveis ambientais sobre a abundância de aves das ordens *Columbiformes* e *Passeriformes* nos espaços verdes de acesso público da cidade do Porto.
Assim, neste capítulo, começar-se-á por apresentar a base de dados cedida pelo CIBIO e a preparação da análise. Posteriormente, aplicar-se-ão os modelos referidos aos dados em estudo.
As análises estatísticas destes modelos foram efetuadas no software livre R versão 2.14.2 (R Development Core Team, 2012) e todas as bibliotecas implementadas serão indicadas no decorrer do estudo.
O estudo reportado neste capítulo foi parcialmente apresentado nas XIX Jornadas de Classificação e Análise de Dados - JOCLAD 2012 (Januário et al., 2012).

5.1 Base de dados e preparação da análise

Da base de dados fornecida pelo CIBIO constam 15 jardins (jardins selecionados na primeira época - ver capítulo 4), 29 espécies de aves da ordem *Passeriformes*, 3 espécies de aves da ordem *Columbiformes* e 7 variáveis ambientais medidas para cada um dos jardins (tabela 4.1). Para cada jardim, a equipa do CIBIO registou o número de aves observadas de cada uma das ordens anteriores num período de 8 minutos e a partir de uma localização privilegiada (Rabaça, 1995). A amostragem foi efetuada por observação direta das aves ou através do canto das mesmas, pelas primeiras horas da manhã (desde o nascer do sol até, aproximadamente, duas horas e meia depois) no mês de junho de 2011. O método de censo de aves aplicado neste estudo é denominado por método dos transetos pontuais sem limite de distância. Para alguns dos jardins em estudo foi considerado mais do que um local de observação (no máximo 3) devido às caraterísticas do jardim, nomeadamente a área. Para efeito de modelação considerou-se, para estes jardins, que o número de aves observadas, para cada ordem, é o número máximo de aves registadas entre os diferentes locais selecionados do jardim em questão. Sublinhe-se que a escolha dos locais de observação teve em conta caraterísticas que a equipa do CIBIO considera favoráveis para a observação das aves. Por fim, somou-se o número de aves observadas das diferentes espécies, para cada ordem, dando origem ao número de aves observadas em cada um dos 15 jardins de acesso público da cidade do Porto.
É importante realçar que a aplicação dos modelos de contagem (GLM e modelos com um número excessivo de zeros) aos dados obtidos pelo CIBIO constitui um exercício meramente académico. Na realidade, os dados são longitudinais (ver capítulo 4) e a sua modelação num único instante de tempo e com um número de observações manifestamente reduzido, é incorreta. De qualquer

forma fazêmo-lo aqui para as observações recolhidas no mês de junho de 2011 porque foram os dados inicialmente disponibilizados. Os dados longitudinais apenas foram obtidos no final do mês de julho de 2012.
Um outro ponto a ter em consideração é que estes modelos de contagem num único instante de tempo têm como resposta o máximo do número de aves observadas entre os vários pontos de localização de um mesmo jardim. Pelo contrário, na análise aos dados longitudinais (capítulo 6) a resposta consiste do número médio uma vez que a equipa do CIBIO considerou, posteriormente, ser o melhor método. Neste capítulo, esta resposta foi também considerada mas não se obtiveram quaisquer modelos significativos (resultados não apresentados neste trabalho).

5.1.1 *Passeriformes*

A ordem dos *Passeriformes* é a ordem que engloba os pássaros e passarinhos. Esta ordem assume-se como a maior das ordens, tanto em número de espécies como em número de indivíduos. Todas as espécies têm pés com três dedos para a frente e um para trás, próprios para se empoleirarem. Dado que neste grupo de aves existem espécies similares torna-se necessário aprender as caraterísticas das famílias. Bruun et al. (2002) referem que uma boa orientação passa pelo formato do bico, cores de plumagem e hábitos. Costa et al. (2011) descreve os *Passeriformes* como aves granívoras de bico cónico e grosso, gregárias e sedentárias.

Figura 5.1: Aves da ordem *Passeriformes*.

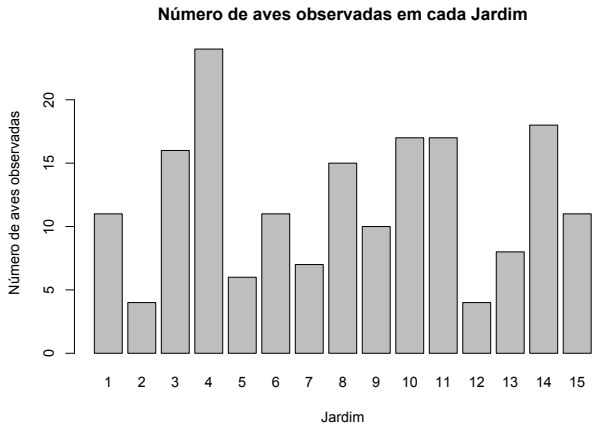


Figura 5.2: Número de *Passeriformes* observados em cada jardim.

Através da figura 5.2 é possível evidenciar que foram sempre observadas aves em todos os jardins

amostrados, sendo 4 o número mínimo e 24 o número máximo de *Passeriformes* observados.
Com o intuito de conhecermos como se comportam os dados em questão, começámos, numa primeira abordagem, por fazer uma análise exploratória dos mesmos (Figura 5.3).

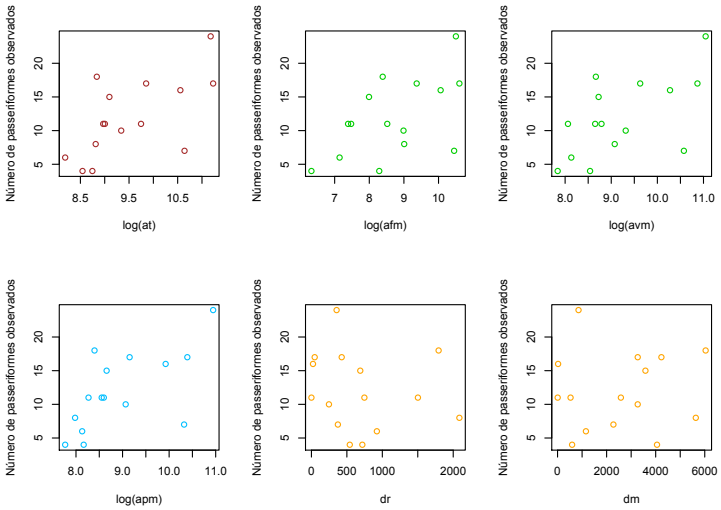


Figura 5.3: Análise exploratória das variáveis ambientais para os *Passeriformes*.

Repare-se que a escala das variáveis ambientais foi logaritmizada uma vez que esta apresentava valores elevados e, portanto, é mais fácil interpretar os dados utilizando a escala logarítmica. A variável *área de plano de água* não está representada na figura 5.3 dado que vai ser posteriormente dicotomizada (histograma na figura 5.4). Decidiu-se dicotomizar esta variável uma vez que os dados em análise possuem poucas observações e valores para a *área de plano de água* com uma amplitude muito grande e onde 1/3 dos jardins não possuem água.
Da figura 5.3 pode retirar-se que existe uma tendência ligeiramente crescente nos quatro primeiros gráficos. Contudo, a análise exploratória não é de todo conclusiva, facto que se deve, em parte, ao reduzido tamanho da amostra (apenas 15 jardins).
Recorde-se que o objetivo deste estudo passa por determinar qual o efeito das variáveis ambientais sobre a abundância de aves da ordem *Passeriformes* nos espaços verdes de acesso público da cidade do Porto. Daí que o termo "abundância"se reporte ao número de aves observadas num período de 8 minutos nos jardins em estudo e não à abundância relativa usual (%).
Face ao objetivo do estudo e dado que estamos a lidar com dados de contagem, recorremos ao modelo de regressão de Poisson para fazer um ajustamento aos dados. Comecemos por considerar um modelo de regressão de Poisson univariado, ou seja, um modelo que contém apenas uma variável explicativa. Nesta investigação as variáveis explicativas dizem respeito às variáveis ambientais. Para a implementação dos modelos lineares generalizados recorreu-se à função *glm()* da biblioteca *stats* do software R.

- **Modelo:** $log(\mu(x)) = \beta_0 + \beta_1 * v.ambiental$

Tabela 5.1: Output do Modelo de Regressão de Poisson Simples.

	Variável	Coef.	Erro Padrão	valor-p
Mod 1	constante	-0.342	0.741	0.644
	$\log(at)$	0.292	0.075	**1.03e-04**
Mod 2	constante	0.016	0.678	0.981
	$\log(avm)$	0.263	0.071	**2.04e-04**
Mod 3	constante	0.640	0.545	0.240
	$\log(afm)$	0.207	0.059	**5.18e-04**
Mod 4	constante	-0.105	0.673	0.876
	$\log(apm)$	0.284	0.072	**8.64e-05**
Mod 5	constante	2.434	0.132	**< 2e-16**
	I_{agua}	0.068	0.160	0.672
Mod 6	constante	2.441	1.259e-01	**< 2e-16**
	dm	1.480e-05	3.907e-05	0.705
Mod 7	constante	2.564	0.111	**< 2e-16**
	dr	−1.267e-04	1.258e-04	0.314

Conforme mencionado, algumas variáveis do estudo foram logaritmizadas devido ao tamanho da escala de cada uma destas. A variável explicativa I_{agua} diz respeito à variável *área de plano de água* dicotomizada à qual passamos a denominar por variável *presença de água*. Esta variável assume o valor 0 em 5 jardins (significa que existem 5 jardins que não possuem água) e regista o valor 1 nos restantes 10 jardins.
Da tabela 5.1 constata-se que todas as variáveis ambientais são estatisticamente significativas quando modeladas individualmente, à exceção da variável *presença de água*, *distância ao mar* e *distância ao rio*.
Tendo em conta que pretendemos determinar qual o melhor modelo que se ajusta aos dados e estando cientes do reduzido tamanho da amostra considerámos diversos modelos. Poderíamos ter optado por utilizar métodos de seleção de variáveis como o *stepwise*, por exemplo, mas devido ao tamanho amostral e ao facto de diversos autores referirem que nem sempre esse deve ser o método preferencial (Kleinbaum and Klein, 2002), optámos por determinar os modelos de forma heurística. Dos vários modelos criados, o modelo que melhor se ajusta aos dados tendo em conta o critério de *Akaike* é

- **Modelo 1:** $log(\mu) = \beta_0 + \beta_1 * log(avm)_c + \beta_2 * I_{agua}$

Para além da variável *área de coberto vegetal* ter sido logaritmizada, esta variável foi também centrada para que se possa inferir acerca da constante e^{β_0}. Nesta situação, e^{β_0} traduz o número médio de *Passeriformes* num jardim onde a *área de coberto vegetal* é igual à média da amostra e onde não existe água.

Os histogramas apresentados na figura 5.4 evidenciam a escala da variável *avm*, da variável *avm logaritmizada* e da *apa* (variável agora dicotomizada e designada por *presença de água*).
O output do modelo 1 revelou que ambas as variáveis são estatisticamente significativas e que a classe 0 da variável *presença de água* é a classe de referência, como se pode confirmar na tabela 5.2.
O facto de a variável I_{agua} apresentar um efeito negativo na abundância de *Passeriformes* levou a que se introduzisse a variável *número de Columbiformes* (a outra ordem de aves em estudo nesta dissertação) no modelo, com o intuito de percebermos se existiam efeitos de confundimento ou interação. Contudo, esta variável não mostrou ser significativa logo, o modelo eleito é o apresentado na Tabela 5.2.

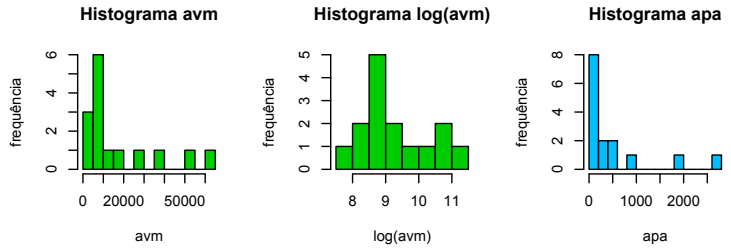


Figura 5.4: Histogramas da variável avm, $log(avm)$ e apa.

Tabela 5.2: Output do Modelo de Regressão de Poisson (Modelo 1).

Variável	**Coef.**	**Erro Padrão**	**valor-p**
constante	2.708	0.148	**<2e-16**
$\log(avm)_c$	0.373	0.092	**5.03e-05**
$I_{agua}(ref.0)$	-0.417	0.209	**0.047**
AIC: 91.78		$\frac{\chi^2}{df} = 1.9$ (c.p)	

Este modelo apresenta um AIC de 91.78 e uma constante de proporcionalidade igual a 1.9. Esta constante de proporcionalidade diz respeito ao parâmetro de dispersão ϕ definido na secção 2.3.2.2 e na secção 2.4.1. Como o valor de ϕ é superior a (1) um significa que existe sobredispersão no modelo.
A figura 5.5 ilustra alguns gráficos de diagnóstico com o intuito de se obter uma melhor compreensão acerca do modelo em estudo.

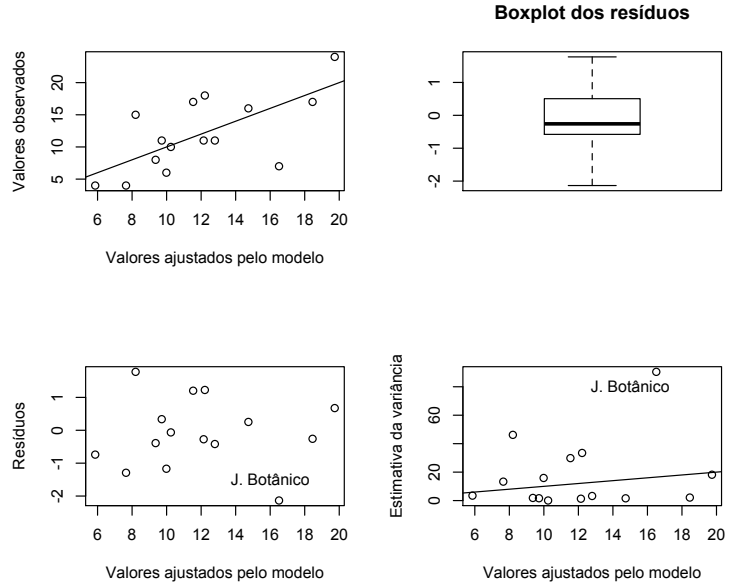


Figura 5.5: Gráficos de diagnóstico do modelo de Regressão de Poisson (Modelo 1).

Da combinação dos gráficos é possível verificar que o modelo em causa não traduz um bom ajustamento aos dados. Apesar disso, este modelo não apresenta resíduos muito altos, apenas um resíduo é inferior a -2. A mesma observação que possui o maior resíduo em valor absoluto é a que possui também uma estimativa da variância maior (entende-se por estimativa da variância o quadrado da diferença entre os valores observados e os valores ajustados, segundo Faraway (2006)). É também esta observação que corresponde ao Jardim Botânico.

Dado que o modelo em causa apresenta sobredispersão, decidimos ajustar aos dados um modelo de regressão Binomial Negativa com o objetivo de modelar a sobredispersão presente. O ajustamento dos dados em análise a este modelo de regressão foi efetuado no R, recorrendo à função *glm.nb()* da biblioteca *MASS*.

O modelo de regressão Binomial Negativa escolhido é expresso da seguinte forma

- **Modelo 2:** $log(\mu) = \beta_0 + \beta_1 * log(avm)_c + \beta_2 * I_{agua}$

Tabela 5.3: Output do Modelo de Regressão Binomial Negativa (Modelo 2).

Variável	**Coef.**	**Erro Padrão**	**valor-p**
constante	2.709	0.181	**2e-16**
$\log(avm)_c$	0.378	0.112	**7.5e-04**
$I_{agua}(ref.0)$	-0.421	0.249	0.092
α	0.042		
AIC: 92.47			

Tanto no modelo de regressão de Poisson (Tabela 5.2) como no modelo de regressão Binomial Negativa (Tabela 5.3), a variável $\log(avm_c)$ mostra ser estatisticamente significativa, ao contrário da variável I_{agua} que revela não ser estatisticamente significativa para o modelo em análise. Para além disso, note-se que o valor dos parâmetros de regressão estimados são praticamente iguais em ambos os modelos, mas o erro padrão é ligeiramente superior no modelo de regressão Binomial Negativa. Este facto decorre do modelo de regressão Binomial Negativa ter em conta a heterogeneidade presente e, portanto, multiplica os erros padrão pela raiz quadrada de $(1 + \alpha\mu)$ (Ismail and Jemain, 2007). O parâmetro ancilar α apresenta um valor muito próximo de zero, 0.04 e este último modelo apresenta um AIC de 92.47.
O mesmo tipo de gráficos de diagnóstico foram elaborados para o modelo de regressão em causa conforme evidenciado na figura 5.6.

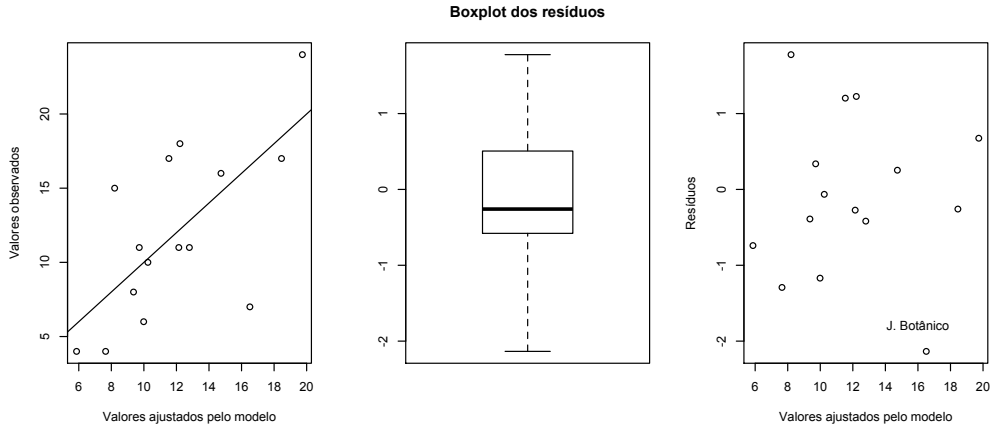


Figura 5.6: Gráficos de diagnóstico do modelo de Regressão Binomial Negativa (Modelo 2).

Apesar de não ser possível identificar facilmente através da análise gráfica (Figura 5.6), os resíduos neste modelo são inferiores aos resíduos do modelo de regressão de Poisson (Modelo 1).
Para comparar a qualidade do ajustamento dos dois modelos (Modelo 1 e Modelo 2) recorre-se ao teste da razão de verosimilhanças, uma vez que o modelo de regressão de Poisson e Binomial Negativa são considerados modelos encaixados (Zuur et al., 2009). Apesar de o resultado não ser estatisticamente significativo (*valor-p*=0.25), o facto de o parâmetro ancilar α estimado apresentar um valor muito próximo de zero, remete-nos para a escolha do modelo de regressão de Poisson. Realce-se o facto de que, para os dados em análise foram criados vários modelos de regressão Binomial Negativa com diversas combinações das variáveis ambientais. Contudo, o modelo aqui presente foi o modelo que melhor se ajustou aos dados, de entre os demais.
Conforme assinalado na análise de regressão de Poisson (Modelo 1), existe uma observação que se destaca relativamente às outras. Essa observação diz respeito, como citado anteriormente, ao Jardim Botânico. Numa tentativa de se perceber o porquê deste jardim se destacar dos restantes, consultou-se, com mais detalhe, as variáveis ambientais inerentes a este jardim.

Figura 5.7: Vista aérea do Jardim Botânico do Porto (Fonte: Farinha-Marques et al., 2011).

Figura 5.8: Plano geral do Jardim Botânico do Porto (Fonte: Farinha-Marques et al., 2011).

Tabela 5.4: Extrato dos dados do software R.

jardim	npass	at	afm	avm	apm	dr	dm	I_agua
Parque da Pasteleira	24	71564.76	36366.93	62724.39	56675.45	355	855	1
Jardim na Rua de Alf. Keil	6	3593.94	1264.31	3408.08	3404.06	923	1156	0
Jardim Machado de Assis	11	7854.29	5008.53	6561.83	5418.13	1502	2581	0
Jardim Botânico	7	41812.31	34561.77	39138.72	30378.10	373	2277	1

A tabela 5.4 descreve-nos as variáveis ambientais em análise de alguns dos jardins selecionados nesta pesquisa. A variável *npass* diz respeito à variável resposta do modelo, ou seja, corresponde ao número de *Passeriformes* observados em cada jardim, sendo que as restantes variáveis ambientais já foram definidas anteriormente.
Analisando cuidadosamente os dados referentes ao Jardim Botânico consta-te que a área de coberto vegetal (avm) representa praticamente a área total do jardim (93 %) e que, da *área de coberto vegetal*, 88% desta são árvores e arbustos (afm). As figura 5.7 e 5.8 comprovam estes resultados. Pelo facto deste jardim possuir uma vasta área de vegetação (na maior parte árvores e arbustos) pensa-se que, eventualmente, durante a amostragem, poderão ter ocorrido erros de observação que evidenciaram este jardim face aos demais. Apesar de estarmos cientes de que a amostra em estudo é de dimensão reduzida, decidimos retirar a observação referente a este para que possamos aferir qual o comportamento do modelo escolhido face a esta alteração.
O melhor modelo de regressão de Poisson obtido de entre os diversos modelos considerados foi o seguinte:

- **Modelo 3:** $log(\mu) = \beta_0 + \beta_1 * log(avm)_c + \beta_2 * I_{agua} + \beta_3 * ncolumb.$

Tabela 5.5: Output do Novo Modelo de Regressão de Poisson (Modelo 3).

Novo Modelo de Regressão de Poisson			
Variável	**Coef.**	**Erro Padrão**	**valor-p**
constante	2.863	0.158	**2e-16**
$\log(avm)_c$	0.481	0.095	**4.28e-07**
$I_{agua}(ref.0)$	-0.360	0.209	0.086
$ncolumb$	-0.013	0.007	**0.047**
AIC:77.839		$\frac{\chi^2}{df} = 1.06$ (c.p)	

O output representado na tabela 5.5 refere-se ao output do modelo de regressão de Poisson eleito sem a observação referente ao Jardim Botânico, o qual denominamos por, novo modelo de regressão de Poisson (Modelo 3). Este mesmo modelo incorpora a variável *ncolumb* que diz respeito ao número de *Columbiformes* observados durante oito minutos em cada um dos jardins em análise. Neste novo modelo apenas as variáveis $\log(avm)_c$ e *ncolumb* são estatisticamente significativas e o valor de AIC obtido é de 77.84. O parâmetro de dispersão obtido é aproximadamente igual a 1 e, por conseguinte, podemos afirmar que não existe praticamente sobredispersão neste novo modelo. Nesta situação, o número médio de *Passeriformes* num jardim onde a *área de coberto vegetal* é igual à média da amostra, onde não existe água e onde não existem *Columbiformes* é de $e^{2.863} \simeq 18$. Os gráficos de diagnóstico deste novo modelo são apresentados na figura 5.9.

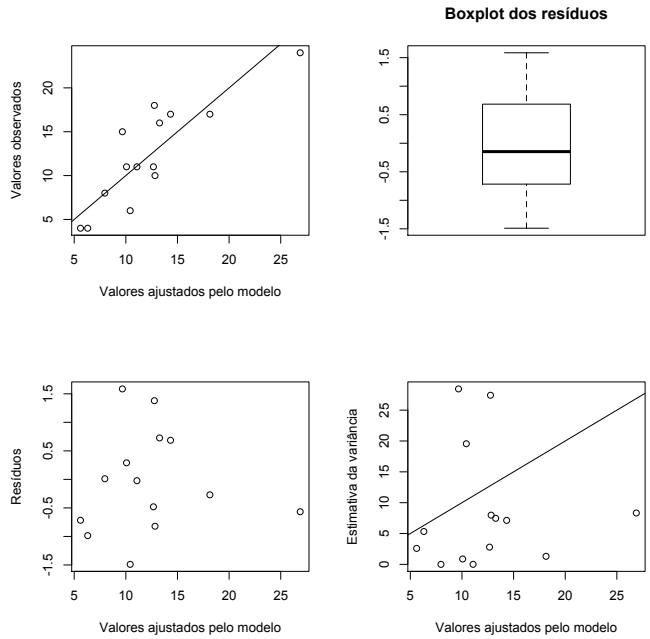


Figura 5.9: Gráficos de diagnóstico do novo modelo de Regressão de Poisson (Modelo 3).

A análise gráfica parece sugerir que o novo modelo de regressão (Modelo 3) faz um melhor ajustamento aos dados. Uma diferença notória neste modelo em relação ao modelo de regressão de Poisson considerado anteriormente (Tabela 5.2) assenta no facto de os resíduos não excederem o valor 1.6 (em módulo) e na circunstância de a estimativa da variância apresentar valores bastante mais baixos.

Voltamos novamente a referir que o objetivo central deste estudo procura determinar qual o efeito das variáveis ambientais sobre a abundância dos *Passeriformes* nos jardins da cidade do Porto. Com efeito, depois de termos escolhido um modelo para os dados em questão (Modelo 3) resta-nos averiguar qual o efeito das variáveis do modelo sobre o objetivo proposto. Para isso recorremos à interpretação do risco relativo (ver secção 2.4.1) das variáveis do modelo.
Comecemos por interpretar o efeito da variável ambiental *área de coberto vegetal* sobre o número esperado de *Passeriformes*.

$$RR = \frac{E(N_{pass}|log(avm)_c + 1)}{E(N_{pass}|log(avm)_c)} = exp(0.481) = 1.62$$

$$\Longrightarrow \quad E(N_{pass}|log(avm)_c + 1) = E(N_{pass}|log(avm)_c) + 0.62E(N_{pass}|log(avm)_c)$$

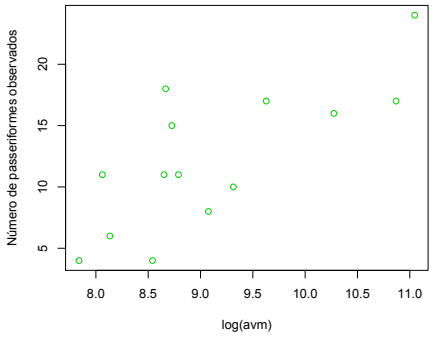


Figura 5.10: Número de *Passeriformes* observados de acordo com a variável $log(avm)$.

De acordo com o risco relativo, por cada unidade aumentada no $log(avm)_c$ e mantendo as restantes variáveis explicativas constantes, observa-se, em média, mais 62% de *Passeriformes*. Repare-se que aumentar uma unidade no $log(avm)_c$ significa multiplicar a *área de coberto vegetal* pela $e^1(\simeq 2.72)$, motivo pelo qual a percentagem de *Passeriformes* observados aumenta para mais de metade.

O risco relativo correspondente à variável *número de Columbiformes* é dado pela expressão:

$$\begin{aligned} RR &= \frac{E(N_{pass}|ncolumb+1)}{E(N_{pass}|ncolumb)} = exp(-0.013) = 0.987 \\ &\Longrightarrow \quad E(N_{pass}|ncolumb+1) = E(N_{pass}|ncolumb) - 0.013E(N_{pass}|ncolumb) \end{aligned}$$

Pela análise do risco relativo conclui-se que, por cada unidade aumentada no *ncolumb* e mantendo as restantes variáveis explicativas constantes, observa-se, em média, menos $1,3\%$ de *Passeriformes* nos jardins de acesso público da cidade do Porto. Apesar de o número de *Columbiformes* ter mostrado ser significativo, a presença de aves desta ordem nos jardins não influencia praticamente a abundância dos *Passeriformes*.

Para a totalidade dos dados, tentou-se corrigir a sobredispersão do modelo de regressão de Poisson (Modelo 1) através de um modelo de regressão Binomial Negativa (Modelo2). Contudo, este procedimento não resultou.
Uma outra possibilidade para corrigir a sobredispersão consiste na modelação de um modelo de regressão Quasi-Poisson.
À semelhança da modelação efetuada no modelo de regressão de Poisson, introduziu-se também a variável *ncolumb* no modelo de regressão Quasi-Poisson. Contudo, esta variável revelou não ser significativa.
O modelo de regressão Quasi-Poisson correspondente ao modelo 1 é:

- **Modelo 4:** $log(\mu) = \beta_0 + \beta_1 * log(avm)_c + \beta_2 * I_{agua}$,

com estimativas apresentadas na tabela 5.6.

Tabela 5.6: Output do Modelo de Regressão de Quasi-Poisson (Modelo 4).

Variável	Coef.	Erro Padrão	valor-p
constante	2.701	0.201	**1.35e-08**
$log(avm)_c$	0.373	0.125	**0.0112**
$I_{agua}(ref.0)$	-0.417	0.285	0.169
$\frac{\chi^2}{df} \simeq 1.9$ (c.p)			

O método de estimação presente neste modelo não é o método da máxima verosimilhança, pelo que não é possível obter-se um valor de AIC para este modelo. Comparando o output do modelo de regressão de Poisson (Tabela 5.2) com o output do modelo de regressão Quasi-Poisson (Tabela 5.6), constata-se de imediato que os parâmetros estimados são exatamente iguais. Contudo, os erros padrão são superiores no modelo de regressão Quasi-Poisson. Esta superioridade deve-se ao facto de este modelo ter em conta a heterogeneidade presente, e portanto, os erros padrão serem multiplicados pela raiz quadrada do parâmetro de dispersão (Ismail and Jemain, 2007). Neste caso, $\phi \simeq 1.9$ e portanto, $\sqrt{\phi} \simeq 1.36$. No modelo em causa, a variável *presença de água* (I_{agua}) era significativa no modelo de regressão de Poisson e deixou de o ser no modelo de regressão Quasi-Poisson. A não significância da variável I_{agua}, confirmada pelo teste das diferenças de desviâncias [1] (Zuur et al., 2009) conduz à sua remoção e, por conseguinte, o modelo escolhido é o Quasi-Poisson com equação dada por:

- **Modelo 5:** $log(\mu) = \beta_0 + \beta_1 * log(avm)_c$,

com o output apresentado na tabela 5.7.

Tabela 5.7: Output do Modelo de Regressão de Quasi-Poisson (Modelo 5).

Variável	Coef.	Erro Padrão	valor-p
constante	2.443	0.108	**8.16e-12**
$log(avm)_c$	0.263	0.099	**0.0198**
$\frac{\chi^2}{df} \simeq 1.95$ (c.p)			

No modelo em causa, a constante de proporcionalidade manteve-se praticamente igual, assumindo o valor de 1.95. Neste modelo, o número médio de *Passeriformes* num jardim onde a *área de coberto vegetal* é igual à média da amostra é de $e^{2.443} \simeq 12$. Da análise do risco relativo em relação à variável *área de coberto vegetal* sobre o número esperado de *Passeriformes* conclui-se que, por cada unidade aumentada no $\log(avm)_c$, observa-se, em média, mais 30% de *Passeriformes*. Este resultado está de acordo com as publicações de diversos autores, tais como, Daniels and Kirkpatrick (2006), Garden et al. (2007), Green and Baker (2003), Palomino and Carrascal (2007), Posa and Sodhi (2006) e Sandström et al. (2006) quando afirmam que a presença de vegetação aumenta a abundância de aves.

O modelo de regressão Quasi-Poisson (Modelo 5) revelou ser o modelo que melhor se ajusta aos dados, sem a necessidade de excluir observações.

[1] $\frac{D_2 - D_1}{\phi(p_1 - p_2)} \sim F_{p_1 - p_2, n - p_1}$

5.1.2 *Columbiformes*

A ordem dos *Columbiformes* inclui aves como as Pombas e as Rolas. Esta ordem inclui aves de tamanho médio, bastante pesadas, com asas pontiagudas e cauda bastante comprida (Bruun et al., 2002). Têm o bico curto e com cera, cabeça pequena e patas curtas. As aves alimentam-se no solo, consumindo sementes ou frutos. Possuem um voo rápido e resistente e, ao levantar voo produzem um barulho de asas de castanholas que serve de sinal de alarme. Estas aves podem beber com o bico imerso na água. Como refere Costa et al. (2011), os ninhos são instalados em árvores, em escarpas ou edifícios. As aves chocam 2 ovos e os filhotes são alimentados com um líquido especial, o "leite do pombo", produzido no papo.

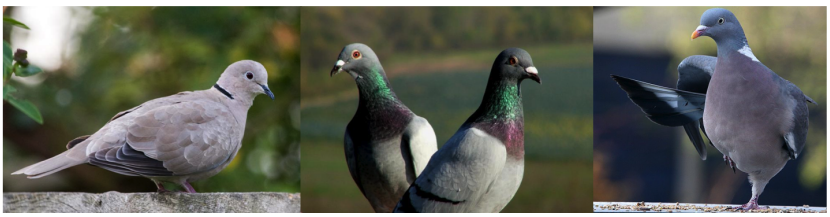


Figura 5.11: Aves da ordem *Columbiformes*.

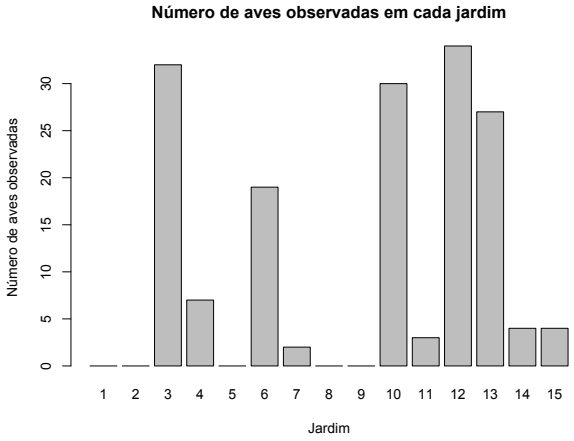


Figura 5.12: Número de *Columbiformes* em cada jardim.

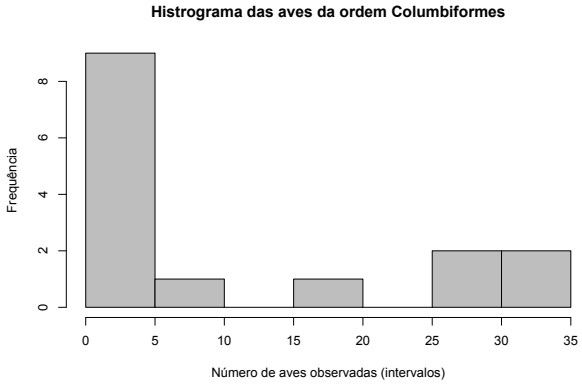


Figura 5.13: Histograma das aves da ordem *Columbiformes*.

Analisando a figura 5.12 é possível constatar que não se observou nenhuma ave da ordem *Columbiformes* em 5 jardins amostrados. Esta ausência significativa de *Columbiformes* leva-nos a afirmar que existe uma inflação de zeros nos dados em estudo (Figura 5.13). Neste contexto, observaram-se 33% de ausências de *Columbiformes*, ou seja, na nossa base de dados 5 observações são contagens nulas.

Dado que se pretende estudar o efeito de diversas variáveis ambientais sobre a abundância de *Columbiformes* nos jardins de acesso público da cidade do Porto, era expectável recorrer à regressão de Poisson uma vez que se está a lidar com dados de contagem. Contudo, devido ao excessivo número de zeros, os dados não podem ser ajustados a uma distribuição de Poisson. Isto porque, para uma distribuição de Poisson com 15 observações independentes e com média 10.80 (média da resposta – número esperado de *Columbiformes*), o numero de zeros esperados é $15 \times exp^{-10.80} = 3.06 \times 10^{-4}$. A solução passa por considerar modelos de contagem com um número excessivo de

zeros: modelos com zeros inflacionados e modelos com barreira.
Comece-se por realizar uma análise exploratória dos dados. A figura 5.14 ilustra como varia o número de *Columbiformes* em função das 7 variáveis ambientais em análise.

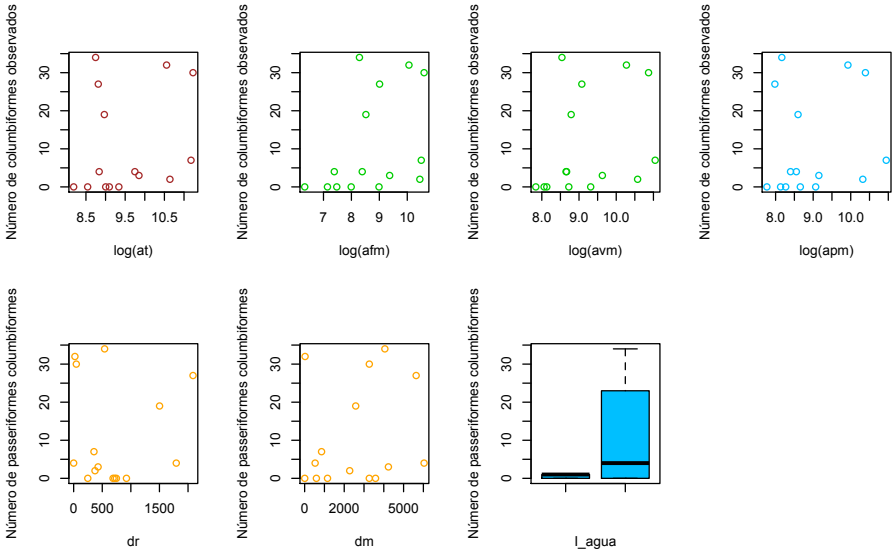


Figura 5.14: Análise exploratória das variáveis ambientais para os *Columbiformes*.

Os jardins em análise no estudo dos *Columbiformes* são coincidentes com os jardins estudados nos *Passeriformes* pelo que, consequentemente, houve também a necessidade de logaritmizar e dicotomizar as mesmas variáveis ambientais anteriormente referidas. Os gráficos apresentados na figura 5.14 espelham a não existência de uma tendência específica no número de *Columbiformes* à medida que se aumenta o valor de cada uma das variáveis ambientais. Para a variável *presença de água* (I_{agua}), a mediana do número de *Columbiformes* observados para os jardins sem água é relativamente mais baixa do que para os jardins com água. Contudo, das 5 observações nulas registadas, 2 delas foram registadas em jardins que não possuem água e as restantes 3 em jardins onde existe água.
Tal como no estudo dos *Passeriformes*, o modelo que melhor se ajusta aos dados dos *Columbiformes* foi determinado de forma heurística.
Apesar de o número de ausências de *Columbiformes* ser muito superior ao esperado para uma distribuição de Poisson, decidimos, ainda, ver como se comporta o melhor modelo de regressão de Poisson que se ajusta aos dados.
O modelo de regressão de Poisson escolhido é o seguinte

- **Modelo 6:** $log(\mu) = \beta_0 + \beta_1 * log(avm)_c + \beta_2 * I_{agua}$

Tabela 5.8: Ouput do Modelo de Regressão de Poisson (Modelo 6).

Variável	**Coef.**	**Erro Padrão**	**valor-p**
β_0	1.849	0.202	**<2e-16**
$\log(avm)_c$	0.220	0.085	**0.010**
$I_{agua}(ref.0)$	0.645	0.239	**0.007**
AIC:261.57		$\frac{\chi^2}{df} = 16.72$ (c.p)	

Tal como no estudo dos *Passeriformes*, a variável $\log(avm)_c$ foi centrada para que seja possível interpretar-se a constante e^{β_0}. Da análise das estimativas (Tabela 5.8) constata-se que ambas as variáveis ambientais presentes no modelo são estatisticamente significativas. Porém, o modelo possui uma constante de proporcionalidade de valor muito elevado. Esta constante de proporcionalidade diz respeito ao parâmetro de dispersão, como anteriormente referido. A elevada sobredispersão do modelo deve-se, sobretudo, ao excessivo número de zeros nos dados dos *Columbiformes*. Através dos gráficos de diagnóstico apresentados na figura 5.15 é possível retirar mais algumas ilações acerca do ajustamento do modelo em causa.

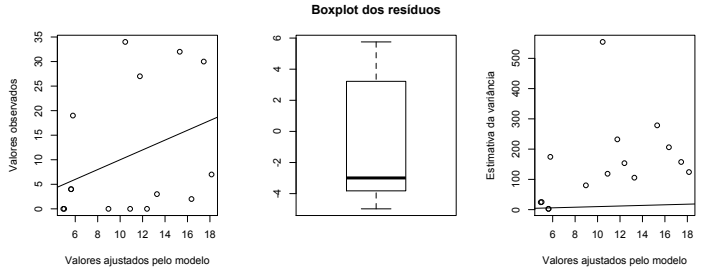


Figura 5.15: Gráficos de diagnóstico do modelo de regressão de Poisson (Modelo 6).

Os gráficos de diagnósticos (Figura 5.15) suscitam que o modelo faz um péssimo ajustamento aos dados. O valor dos resíduos é extremamente elevado (em módulo) e a variância excede claramente a média. Para além disso, o modelo em causa não prevê nenhum zero.
O modelo de regressão de Poisson confirmou o que previámos, ao indicar que os dados dos *Columbiformes* não conseguem ser ajustados a um modelo de regressão de Poisson.
Neste sentido, vamos proceder ao ajustamento dos dados a um modelo de regressão de Poisson com zeros inflacionados e a um modelo de regressão de Poisson com barreira. Como visto anteriormente, o primeiro modelo, faz distinção entre os zeros falsos e os zeros verdadeiros. Neste problema, os zeros falsos dizem respeito aos *Columbiformes* que não foram possíveis de se observar mas que permaneciam no jardim (poderiam estar camuflados, p. ex.) enquanto que os zeros verdadeiros correspondem aos *Columbiformes* que não se encontram, de facto, no jardim (migraram, p. ex.). Já o segundo modelo, apenas considera zeros *versus* não zeros, pelo que o processo de contagem inerente a este modelo é truncado em zero.
Tendo em conta o reduzido tamanho amostral, não se consideraram modelos com mais de duas variáveis explicativas.
O modelo de regressão de Poisson com zeros inflacionados escolhido é

- **Modelo:**
$$\begin{aligned} log(\mu) &= \beta_0 + \beta_1 * log(avm)_c + \beta_2 * I_{agua} \\ logit(\pi) &= \gamma_0 + \gamma_1 * log(afm)_c \end{aligned}$$

juntamente com a equação (3.1).
Já o modelo de regressão de Poisson com barreira é traduzido por

- **Modelo:**

$$\begin{aligned} log(\mu) &= \beta_0 + \beta_1 * log(avm)_c + \beta_2 * I_{agua} \\ logit(\pi) &= \gamma_0 + \gamma_1 * log(afm)_c \end{aligned}$$

juntamente com a equação (3.11).

Note-se que, para o modelo de regressão de Poisson com zeros inflacionados o π diz respeito à probabilidade de existir um zero falso, enquanto que para o modelo de regressão de Poisson com barreira o π corresponde à probabilidade de se observar uma contagem não nula. Para uma maior facilidade de leitura, intitule-se de Modelo ZIP o modelo de regressão de Poisson com zeros inflacionados e Modelo com barreira, o modelo de regressão de Poisson com barreira. A implementação do modelo ZIP no R é efetuada através da função *zeroinfl()* e o modelo com barreira é implementado através da rotina *hurdle()*, ambas pertencentes à biblioteca *pscl*.

Tabela 5.9: Output do Modelo de Regressão de Poisson com zeros inflacionados e com barreira.

		Modelo ZIP			**Modelo c/ barreira**		
	Variável	Coef.	Erro Padrão	valor-p	Coef.	Erro Padrão	valor-p
Modelo de contagem	*constante*	2.049	0.198	**<2e-16**	2.049	0.198	**<2e-16**
	$\log(avm)_c$	-0.289	0.095	**0.002**	-0.289	0.095	**0.002**
	$I_{agua}(ref.0)$	1.106	0.23	**<2.34e-06**	1.106	0.234	**<2.34e-06**
Modelo de zeros	*constante*	-1.256	0.864	0.146	1.256	0.865	0.146
	$\log(afm)_c$	-1.641	0.881	0.062	1.641	0.881	0.062
		AIC=151.40			AIC=151.40		

Focando a análise no Modelo Zip, é possível constatar que a variável *área de coberto vegetal* e *presença de água* são estatisticamente significativas para o modelo de contagem pelo que nos interessa perceber qual o efeito destas variáveis sobre o número esperado de *Columbiformes*. Comecemos por interpretar o risco relativo para a variável *área de coberto vegetal*.

$$RR = \frac{E(N_{columb}|log(avm)_c+1)}{E(N_{columb}|log(avm)_c)} = exp(-0.289) = 0.75$$

$$\Longrightarrow \quad E(N_{columb}|log(avm)_c+1) = E(N_{columb}|log(avm)_c) - 0.25E(N_{columb}|log(avm)_c)$$

De acordo com o risco relativo, à medida que se aumenta o $\log(avm)_c$ de uma unidade e mantendo as restantes variáveis explicativas constantes, observa-se, em média, menos 25 % de *Columbiformes*. Como já referido, aumentar uma unidade no $log(avm)_c$ significa multiplicar a *área de coberto vegetal* pela $e^1(\simeq 2.72)$. Este resultado não era, de todo, esperado, uma vez que contradiz claramente tudo o que vem reportado na literatura (Garden et al., 2007; Palomino and Carrascal, 2007; Sandström et al., 2006), contradição que não nos foi possível justificar nesta investigação.
O risco relativo correspondente à variável *presença de água* (I_{agua}) é expresso por

$$RR = \frac{E(N_{columb}|agua)}{E(N_{columb}|\overline{agua})} = exp(1.106) = 3.02$$

$$\Longrightarrow \quad E(N_{columb}|agua) = 3.02E(N_{columb}|\overline{agua})$$

Da análise do risco relativo conclui-se que nos jardins com água observam-se, em média, mais 3.02 *Columbiformes*, por oposição aos que não têm água. O resultado obtido coincide com a publicação de Brown and Freitas (2002) onde estes autores afirmam que a presença de água traduz-se num efeito positivo para a biodiversidade.

Tendo em conta a definição de odds ratio para que se possa inferir qual o efeito da variável *área de coberto fanerófito* sobre o número de zeros falsos (sucesso) do modelo, tem-se que

$$\begin{aligned} OR &= \frac{Odds(zerofalso|\log(afm)_c+1)}{Odds(zerofalso|\log(afm)_c)} = exp(-1641) = 0.19 \\ \Longrightarrow\ & Odds(zerofalso|\log(afm)_c+1) = Odds(zerofalso|\log(afm)_c) \\ & -0.81Odds(zerofalso|\log(afm)_c) \end{aligned}$$

O *odds* para a existência de zeros falsos diminui cerca de 81% com o acréscimo de uma unidade em $\log(afm)_c$ (isto é, um aumento de 2.72 m^2 na *área de coberto fanerófito*). O resultado obtido é razoável dado que um aumento da *área de coberto fanerófito* conduz a um aumento do número esperado de *Columbiformes* e, portanto, a uma redução do número de zeros; em particular, o número de zeros falsos também irá diminuir. Note-se, contudo, que o nível de significância é ligeiramente superior ao usualmente estabelecido, 0.05.
Para cada uma das observações, a probabilidade π de existir um zero falso é dada pelo comando *predict(modeloZIP,type="zero")* no R. A tabela 5.10 apresenta o valor dessa probabilidade para cada observação.

Tabela 5.10: Probabilidade de existir um zero falso para cada jardim.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.679	0.933	0.029	0.015	0.786	0.277	0.016	0.476
9	**10**	**11**	**12**	**13**	**14**	**15**	
0.151	0.012	0.087	0.359	0.146	0.322	0.711	

Já a probabilidade de existir um zero (verdadeiro ou falso) para cada observação, é dada pelo comando *predict(modeloZIP,type="prob")[,1]*. A tabela 5.11 mostra-nos qual o valor dessa probabilidade para cada jardim.

Tabela 5.11: Probabilidade de existir um zero para cada jardim (Modelo ZIP).

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.679	0.933	0.029	0.015	0.786	0.277	0.016	0.476
9	**10**	**11**	**12**	**13**	**14**	**15**	
0.151	0.012	0.087	0.359	0.146	0.322	0.711	

Para cada um dos jardins, comparou-se a probabilidade π de existir um zero falso com a probabilidade de existir um zero verdadeiro, dada por $(1-\pi)e^{(-\mu)}$. Desta comparação chegou-se à conclusão que todos os zeros previstos pelo modelo ZIP são falsos.
Nas tabelas apresentadas apenas consideramos 3 casas decimais mas uma análise mais cuidada aos números indicados revela algumas diferenças inferiores a 0.001.

A média prevista para a distribuição condicionada $Y|X=\mathtt{x}_i, i=1,...,15$ corresponde à equação (3.2) e é obtida no R através do comando *fitted(modeloZIP,type="response")*. A tabela 5.12 contém a estimativa deste valor para cada jardim.

Tabela 5.12: Média prevista para a distribuição condicionada do modelo ZIP.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
3.469	2.338	16.725	13.578	2.273	6.344	15.548	14.141
9	**10**	**11**	**12**	**13**	**14**	**15**	
19.341	14.327	18.988	18.262	20.825	6.163	2.644	

O resultado total do modelo de zeros com o modelo de contagens, isto é, o modelo ZIP, é apresentado na figura 5.16 onde se representa a função de probabilidade $Y|x = \mathbf{x}_i$ para $i = 1, ..., 15$.

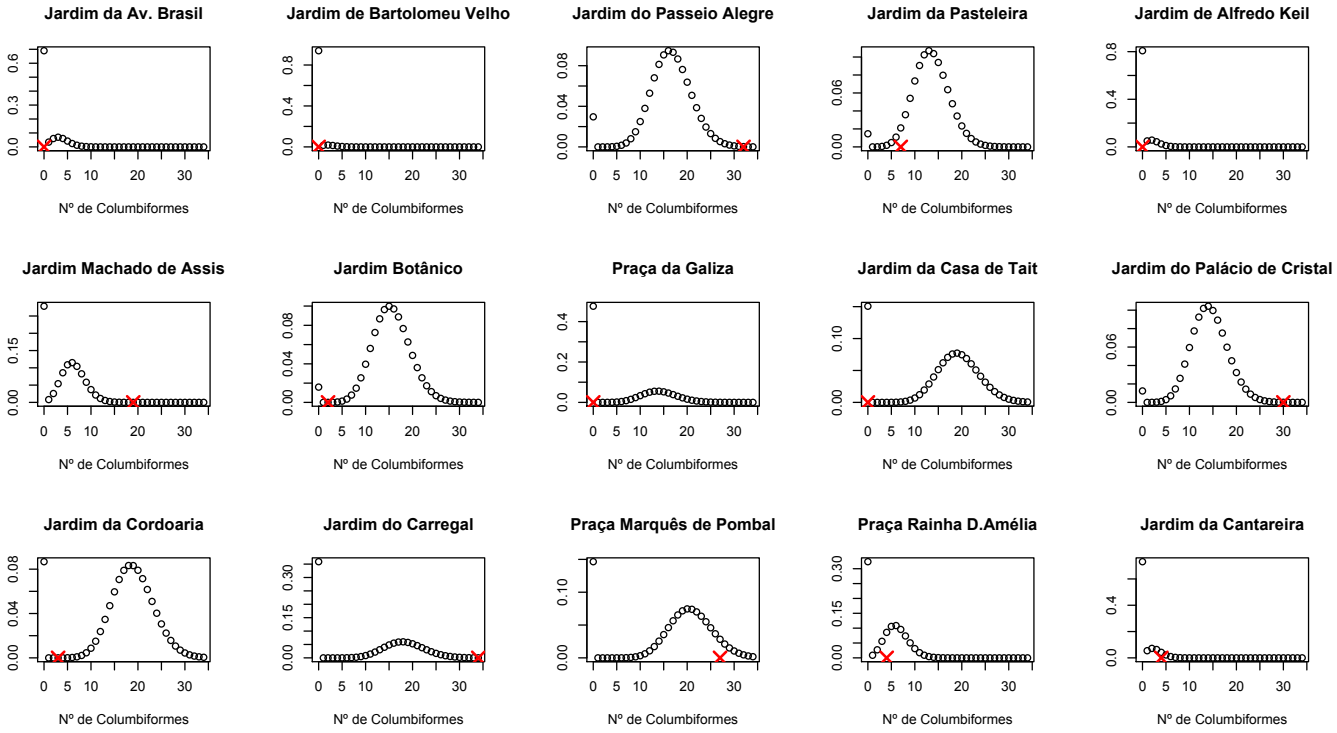


Figura 5.16: Função de probabilidade de $Y|X = \mathbf{x}_i \sim ZIP(\pi_i, \mu_i)$.

A indicação (X, a vermelho) desenhada nos gráficos da figura 5.16 indica os valores observados em cada jardim. É possível constatar que os zeros são bem previstos mas o modelo de contagem não parece estar a fazer um bom ajustamento. Eventualmente, precisaríamos de mais dados.
Centremo-nos agora na análise do modelo com barreira. Como é possível retirar da tabela 5.9, as estimativas dos parâmetros do modelo de contagem no modelo com barreira são essencialmente coincidentes com as do modelo ZIP. A igualdade das estimativas entre os dois modelos não é um resultado comum, uma vez que estes modelos modelam de forma diferente as contagens. Tal situação deve-se provavelmente ao facto de no modelo ZIP, todos os zeros previstos serem zeros falsos, ou seja, nenhum zero deriva de um processo de contagem, o que coincide com os pressupostos do modelo com barreira. Deste modo, as interpretações para os parâmetros de regressão são análogas às efetuadas para o modelo de contagem no modelo ZIP.
No caso do modelo de zeros, as estimativas dos parâmetros de regressão são iguais (em módulo) em ambos os modelos mas apresentam sinais inversos. Esta diferença de sinais assenta no facto de a probabilidade π assumir definições distintas em cada um dos modelos. Para o modelo com barreira, a componente barreira de zeros (π) descreve a probabilidade de se observar uma contagem

positiva enquanto que, para o modelo ZIP, a componente de inflação (π) prevê a probabilidade de se observar um zero falso (contagem nula que deriva de uma massa pontual em zero). Mais uma vez, realça-se o facto de que as estimativas dos parâmetros assumem o mesmo valor em módulo nos dois modelos, dado que o modelo ZIP apenas previu zeros falsos.
No modelo com barreira, e assumindo um erro de 6%, o odds para a existência de um número positivo de *Columbiformes* aumenta em 5.16 vezes sempre que o $\log(afm)_c$ aumenta de uma unidade. Isto está em consonância com o resultado obtido do modelo ZIP: um aumento na probabilidade de se observar uma contagem não nula implica uma diminuição da probabilidade de se observar um zero (uma vez que a soma é constante e igual a 1); em particular, uma diminuição da probabilidade de se observar um zero falso.
No modelo em questão, a probabilidade de existir um zero em cada um dos jardins é apresentada na tabela 5.13. A instrução no R é análoga à do modelo ZIP.

Tabela 5.13: Probabilidade de existir um zero para cada jardim (Modelo com barreira).

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.679	0.933	0.029	0.015	0.786	0.277	0.016	0.476
9	**10**	**11**	**12**	**13**	**14**	**15**	
0.151	0.012	0.087	0.359	0.146	0.322	0.711	

Como já era suposto, devido às interpretações anteriores, a probabilidade de se observar um zero em cada um dos jardins amostrados é igual nos dois modelos.
A média prevista para a distribuição condicionada $Y|x = \mathbf{x}_i, i = 1, ..., 15$ pode ser obtida no R da mesma forma que foi implementada para o modelo ZIP. A tabela 5.14 apresenta-nos o valor das médias previstas.

Tabela 5.14: Média prevista para a distribuição condicionada do modelo com barreira.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
3.469	2.338	16.725	13.578	2.273	6.344	15.548	14.141
9	**10**	**11**	**12**	**13**	**14**	**15**	
19.341	14.327	18.988	18.262	20.825	6.163	2.644	

Como é possível constatar-se, a média prevista no modelo com barreira é praticamente igual à do modelo ZIP, tabela 5.12 (na prática, os valores da média prevista pelo modelo com barreira são ligeiramente inferiores a partir da quarta casa decimal). De facto, os gráficos da figura 5.17 são muito semelhantes aos da figura 5.16.

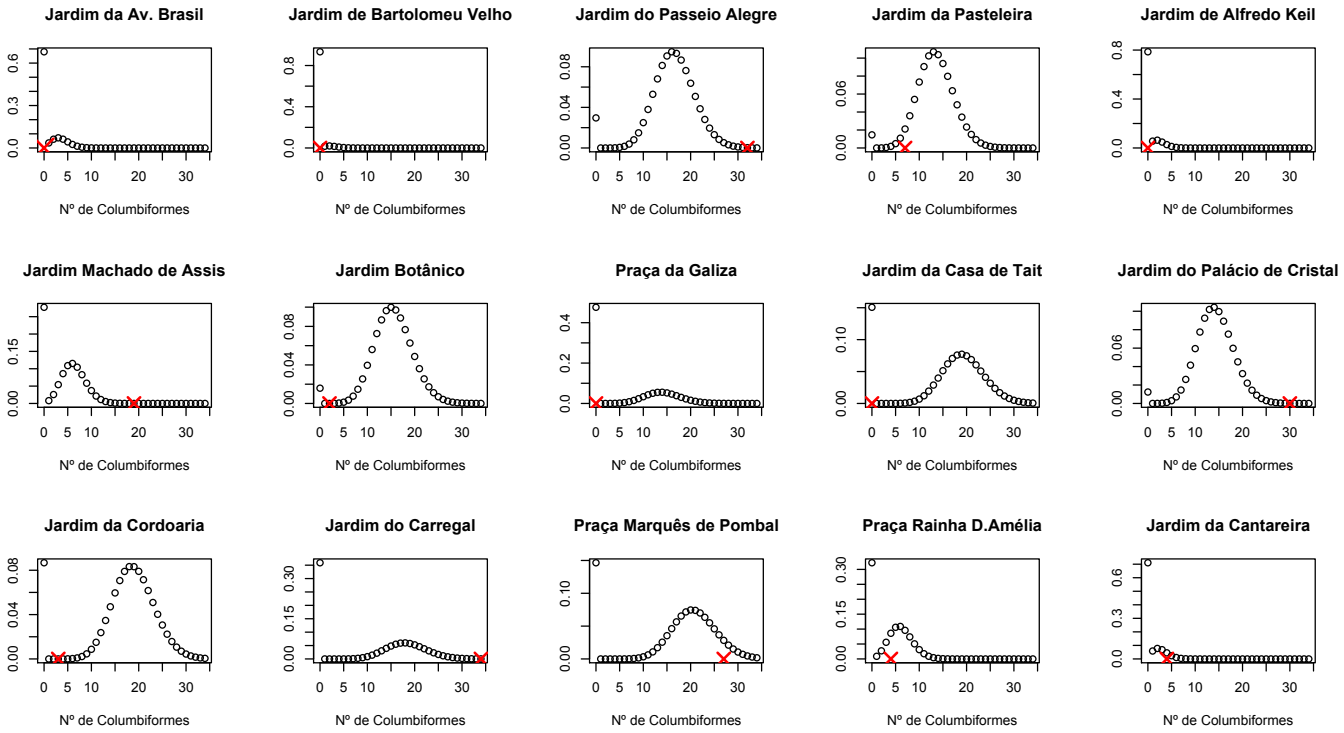


Figura 5.17: Função de probabilidade de $Y|X = \mathbf{x}_i \sim Barreira(\pi_i, \mu_i)$.

O modelo com barreira prevê os zeros corretamente, contudo, à semelhança do modelo ZIP, o modelo de contagem revela não estar a fazer um bom ajustamento. O valor do AIC de ambos os modelos apresenta igual valor.
Nesta situação, a escolha do modelo é indiferente, dado que ambos conduzem ao mesmo resultado. Face a esta situação, aconselha-se uma revisão ao processo de amostragem.

Note-se que ambos os modelos considerados apresentam sobredispersão ($\phi = 5.6$), o que nos remete para a sua modelação utilizando a distribuição binomial negativa. Dos vários modelos considerados com esta distribuição, nenhum apresentou variáveis explicativas significativas, mesmo considerando um nível de significância de 0.10.
Por este facto, e também por sabermos que a significância estatística dos coeficientes obtidos nos modelos de Poisson acima estudados pode ser espúria, os resultados obtidos devem ser interpretados com cuidado.

Capítulo 6

Dados Longitudinais

Dados longitudinais surgem quando a resposta de cada uma das unidades experimentais é observada mais do que uma vez. Se pensarmos que as observações se efetuam ao longo do tempo, obtém-se uma sequência temporal de duas ou mais observações para cada indivíduo (Diggle et al., 2002). Estes dados podem ser obtidos de forma prospetiva, quando os indivíduos são seguidos ao longo do tempo, ou de forma retrospetiva quando múltiplas medições em cada indivíduo são extraídas do seu historial (Diggle et al., 2002). O objetivo primordial de um estudo longitudinal é o de caraterizar as alterações da variável resposta com o tempo. Além disso, estes estudos apresentam um papel fundamental nas mais diversas áreas de conhecimento na medida em que permitem, também, determinar se essas alterações estão relacionadas com um conjunto de fatores que não o tempo (Cabral and Gonçalves, 2011). Segundo Cameron and Trivedi (2008), este tipo de dados têm grande aplicação em diversos modelos de regressão, onde a variável resposta pode ser contínua ou discreta. Exemplos de dados longitudinais citados por Cameron and Trivedi (2008) incluem o número de patentes concedidas a cada uma das muitas empresas individuais durante vários anos, o número de acidentes em diversas regiões ao longo de diversos anos, assim como investigar as alterações de peso de um determinado indivíduo ao longo do tempo. Uma vantagem adicional na utilização destes dados reside no facto de estes permitirem aferir a heterogeneidade associada a cada indivíduo (Cameron and Trivedi, 2008). Esta circunstância implica que o conjunto de observações de um indivíduo ao longo do tempo tende a ser correlacionado e, por conseguinte, os dados longitudinais requerem métodos estatísticos específicos que tenham em conta a correlação existente, para que as inferências realizadas sejam válidas (Diggle et al., 2002). De forma sucinta, neste tipo de dados considera-se que as várias observações de um mesmo indivíduo são dependentes e que, observações de diferentes indivíduos assumem-se independentes.

Uma outra caraterística destes dados que discutiremos neste capítulo, tem subjacente a existência frequente de dados omissos, consequência de os indivíduos amostrados não possuirem o mesmo número de observações e de estas terem sido amostradas em diferentes ocasiões.

Diggle et al. (2002) consideraram três tipos de modelos para a análise de dados longitudinais: (i) o modelo marginal que usa Equações de Estimação Generalizadas[1], (ii) o modelo com efeitos aleatórios (Modelo Linear Generalizado Misto) e o (iii) modelo de transição. No presente trabalho apenas serão abordados os dois primeiros. A principal diferença entre estes modelos centra-se no objetivo da inferência que se pretende efetuar. O modelo marginal tem como propósito efetuar inferências sobre a população, enquanto que o modelo com efeitos aleatórios tem como objetivo realizar inferências sobre cada um dos indivíduos da população. Zeger et al. (1988) definem estes modelos como modelo *Population average (PA)* e modelo *Subject-specific(SS)*, respetivamente.

Neste capítulo será apresentada a teoria envolvente para modelos de regressão para dados longitu-

[1] GEE: **G**eneralized **E**stimating **E**quations.

dinais, assim como a aplicação dos dois modelos em estudo aos dados disponibilizados pelo projeto do CIBIO.

6.1 Dados Omissos em Estudos Longitudinais

Em estudos longitudinais é frequente a existência de dados omissos, decorrentes do facto de os indivíduos possuírem observações que não foram todas registadas nos mesmos instantes. Fitzmaurice et al. (2004) aludem que a presença de valores em falta tornam os dados não equilibrados e portanto, os métodos de análise devem ser capazes de trabalhar os dados não equilibrados sem terem de eliminar da análise os dados dos indivíduos que possuem valores omissos. Para além disto, estes autores referem também que quando estamos na presença de dados omissos, há perda de informação e redução na precisão com a qual as alterações na resposta média ao longo do tempo podem ser estimadas. Esta redução na precisão está diretamente relacionada com o número de valores omissos nos dados e pode também, até certo ponto, influenciar o modo como a análise lida com os dados em falta. Por fim, quando temos dados omissos, a validação de qualquer método de análise requer certas suposições acerca das razões das omissões, normalmente referidas como mecanismo de omissão de dados (*missing data mechanism*). Nestas circunstâncias deve-se ter em consideração as razões que levaram à omissão.
O mecanismo de omissão de dados é um modelo que descreve a probabilidade com que a resposta é ou não observada (*missing*) em determinada ocasião. Cabral and Gonçalves (2011) e Diggle et al. (2002) narram que Rubin (1976) e Little and Rubin (1987) classificaram o mecanismo de omissão dos dados em:

(i) omissão completamente aleatória (**MCAR**- *Missing Completely At Random*)

(ii) omissão aleatória (**MAR**- *Missing At Random*)

(iii) omissão não aleatória (**NMAR**- *Not Missing At Random*).

Neste estudo abordaremos apenas os dois primeiros mecanismos de omissão.
A definição dos mecanismos de omissão a apresentar fundamenta-se, sobretudo, na literatura produzida por Cabral and Gonçalves (2011) - capítulo 1, Diggle et al. (2002), capítulo 13 e Fitzmaurice et al. (2004), capítulo 4.
Os dados omissos são MCAR quando a probabilidade das respostas estarem omissas não está relacionada com o valor específico da mesma ou com o conjunto de respostas observadas. Por outras palavras, os dados longitudinais são MCAR quando o mecanismo de omissão é independente dos valores observados.
A título de exemplo, considere-se que se pretende amostrar qual o nível de colesterol dos adolescentes numa determinada escola do Porto. Aleatoriamente, um adolescente da amostra viu-se obrigado a mudar de escola por motivos de relocalização de emprego dos pais para outro distrito, e, portanto, não pode continuar no estudo. Este valor omisso nada teve a ver com o nível de colesterol da criança pelo que o mecanismo de omissão é classificado como MCAR.
Uma caraterística fundamental de MCAR é a dos dados observados poderem ser considerados uma amostra aleatória dos dados completos. Como consequência, os momentos (p. ex: média, variância e covariância) assim como as distribuições dos dados observados não diferem dos correspondentes momentos nem das distribuições para os dados completos.
Qualquer método de análise de dados longitudinais que leva a inferências válidas na ausência de dados omissos, conduze também a inferências válidas quando é aplicado a dados omissos MCAR. É de realçar que todos os métodos apresentados neste capítulo para a análise de dados longitudinais são válidos quando se considera o mecanismo de omissão MCAR. Deste modo, é possível afirmar

que o mecanismo de omissão MCAR revela ter importantes consequências na análise longitudinal.

Os dados omissos dizem-se *missing at random* (MAR) quando a probabilidade das respostas estarem omissas depende dos valores das respostas observadas, mas não se encontra relacionada com o valor omisso que deveria ser obtido. Por exemplo, quando um protocolo de estudo requer que um indivíduo seja removido da amostra caso o valor da resposta não pertença a um intervalo previamente especificado.
O facto do mecanismo MAR depender dos valores das respostas observadas implica que alguns métodos de análise de dados longitudinais deixem de produzir estimativas válidas caso a distribuição conjunta do vetor resposta não seja corretamente especificada, ou corretamente modelado o mecanismo de omissão.
Em contraste com o mecanismo MCAR, o mecanismo MAR não é permitido em todos os métodos de análise de dados longitudinais, como por exemplo o modelo marginal. Contudo, os métodos baseados na máxima verosimilhança conseguem produzir estimativas válidas quando o mecanismo de omissão é MAR.

6.2 Modelo Linear Generalizado Misto

Os modelos lineares generalizados mistos (GLMM[2]) são uma extensão dos modelos lineares generalizados (GLM) que permitem a inclusão de efeitos aleatórios no preditor linear. Estes modelos foram propostos por Breslow and Clayton (1993) e modelam os dados longitudinais com qualquer variável resposta que pertença à família exponencial.
O facto dos GLMM consentirem a introdução de efeitos aleatórios, permite que a estrutura de correlação entre observações de um mesmo indivíduo possa ser modelada (Hedeker, 2005).
Como mencionam Cabral and Gonçalves (2011), o objetivo dos modelos lineares generalizados mistos é descrever as alterações da resposta média de cada indivíduo e a relação destas alterações com as covariáveis de interesse. Este modelos pretendem, assim, inferir sobre o indivíduo e não sobre a população (Cabral and Gonçalves, 2011; Hedeker, 2005). Relativamente aos dados omissos, os GLMM assumem os mecanismos de omissão MCAR e MAR.
Ao nível computacional e ao nível de interpretação de alguns parâmetros estimados, estes modelos apresentam algumas complicações provenientes da estrutura não linear agregada à inclusão de efeitos aleatórios.
Nesta secção será abordada a estrutura do modelo linear generalizado misto recorrendo a um exemplo para dados de contagem e à interpretação dos parâmetros deste modelo. Por fim, é feita uma descrição da estimação dos parâmetros e dos métodos inferenciais associados a este modelo.

6.2.1 Estrutura do modelo linear generalizado misto

Seja y_{i,t_j} o valor da resposta observada para o indivíduo $i (i = 1, ..., n)$ no instante t_j $(j = 1, ..., T_i)$. y_{i,t_j} é a realização da variável resposta Y_{i,t_j} e

$$\mathtt{Y}_i = (Y_{i,t_1}, ..., Y_{i,t_{T_i}})^T$$

representa o vetor das variáveis resposta para o indivíduo i. O perfil de cada indivíduo é assim dado pelo vetor $\mathtt{y_i} = (y_{i,t_1}, ..., y_{i,t_{T_i}})^T$.

[2] **G**eneralized **L**inear **M**ixed **M**odels.

Seja[3]

$$\mathtt{x}_{i,t_j}^T = (t_j, x_{1,i,t_j}, ..., x_{p,i,t_j})$$

o vetor das covariáveis de interesse associado a cada Y_{i,t_j}, de dimensão $(p+1)$, a matriz das covariáveis, $\mathtt{X}_i$, de ordem $T_i \times (p+1)$ e o vetor dos coeficientes de regressão $\beta = (\beta_1, ..., \beta_{p+1})^T$, de dimensão $(p+1)$.
Por fim, defina-se

$$\mathtt{z}_{i,t_j}^T = (t_j, z_{1,i,t_j}, ..., z_{q,i,t_j})$$

como um vetor $(q+1)$ de covariáveis (usualmente um subconjunto de $\mathtt{x}_{i,t_j}$) associado a um vetor $(q+1)$ de efeitos aleatórios $\mathtt{b_i}$ (Cabral and Gonçalves, 2011).

Os modelos lineares generalizados podem ser especificados em 3 partes (Fitzmaurice et al., 2004):

1. A variável resposta $\mathtt{Y}_i = (Y_{i,t_1}, ..., Y_{i,t_{T_i}})^T$, condicionada pelo vetor dos efeitos aleatórios $\mathtt{b}_i$, segue uma distribuição pertencente à família exponencial com função densidade

$$f(y_{i,t_j}|\mathtt{b}_i) = exp\left\{\frac{\omega_{i,t_j}}{\phi}(y_{i,t_j}\theta_{i,t_j} - b(\theta_{i,t_j})) + c(y_{i,t_j}, \phi)\right\}.$$

O valor médio e a variância condicional são expressos, respetivamente, por:

$$\begin{aligned} E(Y_{i,t_j}|\mathtt{b_i}) &= \mu_{i,t_j} = b'(\theta_{i,t_j}), \\ \text{e} & \\ Var(Y_{i,t_j}|\mathtt{b_i}) &= v_{i,t_j} = b''(\theta_{i,t_j})\frac{\phi}{\omega_{i,t_j}}. \end{aligned}$$

Tal como nos GLM, assume-se que,

$$\begin{aligned} g(\mu_{i,t_j}) &= \mathtt{x}_{i,t_j}^T\beta + \mathtt{z}_{i,t_j}^T\mathtt{b}_i, \\ \text{e} & \\ v_{i,t_j} &= V(\mu_{i,t_j})\frac{\phi}{\omega_{i,t_j}}, \end{aligned}$$

onde $g(.)$ é uma função de ligação pré-especificada e V uma função de variância conhecida, ϕ é o parâmetro de dispersão e ω_{i,t_j} é um peso conhecido atribuído às observações da variável resposta. Uma vez fixados os efeitos aleatórios $\mathtt{b}_i$, as variáveis resposta assumem-se mutuamente independentes.

2. A média condicional da variável resposta Y_{i,t_j} depende dos efeitos fixos e dos efeitos aleatórios. Deste modo, o modelo linear generalizado misto pode ser escrito por (Cabral and Gonçalves, 2011; Fitzmaurice et al., 2004; Hedeker, 2005),

$$g\left\{E(Y_{i,t_j}|\mathtt{b}_i)\right\} = \eta_{i,t_j} = \mathtt{x}_{i,t_j}^T\beta + \mathtt{z}_{i,t_j}^T\mathtt{b}_i.$$

3. Por fim, resta-nos especificar a distribuição dos efeitos aleatórios $\mathtt{b}_i$. À partida, qualquer distribuição multivariada poderia ser considerada para $\mathtt{b}_i$. Contudo, na prática, assume-se que a distribuição comum aos efeitos aleatórios $\mathtt{b}_i$ é a Gaussiana multivariada com valor esperado zero e matriz de variância-covariância $\mathtt{D}$ (Diggle et al., 2002). Salienta-se o facto dos efeitos aleatórios $\mathtt{b}_i$ se assumirem independentes das covariáveis.

Face ao objetivo do estudo, apresentamos de seguida o exemplo ilustrativo do modelo linear generalizado misto para contagens.

[3]No caso do modelo conter um termo constante deve-se considerar $\tilde{\mathtt{x}}_i = (1, \mathtt{x}_i^T)^T$ e $\beta = (\beta_0, \beta_1, ..., \beta_{p+1})^T$, tal como nos modelos lineares generalizados.

Modelo linear generalizado misto para contagens

Considere-se que Y_{i,t_j}, a variável resposta do indivíduo i no instante t_j, segue uma distribuição de Poisson. Nesta situação tem-se:

1. As variáveis resposta Y_{i,t_j} condicionadas pelos efeitos aleatórios $\mathtt{b}_i$ são independentes e seguem uma distribuição de Poisson, com

$$E(Y_{i,t_j}|\mathtt{b}_i) = Var(Y_{i,t_j}|\mathtt{b}_i) \quad (\phi = 1).$$

2. A média μ_{i,t_j} está relacionada com a componente sistemática do modelo (que suponhamos incluir um termo constante) através da equação

$$\begin{aligned} log(\mu_{i,t_j}) &= (1, x_{i,t_j}^T)^T\beta + (1, z_{i,t_j}^T)^T\mathtt{b}_i \\ &= (\beta_0 + \mathtt{b}_{0i}) + (\beta_1 + \mathtt{b}_{1i})x_{1,i,t_j} + ... + (\beta_p + \mathtt{b}_{pi})x_{p,i,t_j} + (\beta_{p+1} + \mathtt{b}_{p+1,i})t_j, \end{aligned}$$

onde o log é a função de ligação usual quando estamos na presença de dados de contagem.

3. Por fim, assume-se que o vetor dos efeitos aleatórios, $\mathtt{b}_i$, segue uma distribuição Gaussiana de valor médio zero e matriz de variância-covariância $\mathtt{D}$, de dimensão $(p+2) \times (p+2)$.

6.2.2 Interpretação dos parâmetros do modelo

O facto de os modelos lineares generalizados mistos admitirem a inclusão de efeitos aleatórios nos modelos lineares generalizados permite, como já referido, modelar a estrutura de correlação entre observações de um mesmo indivíduo. A interpretação dos parâmetros fixos nos GLMM será assim afetada pelos efeitos aleatórios presentes no modelo, como explicam Cabral and Gonçalves (2011). Isto porque, nos GLMM (Cabral and Gonçalves, 2011; Costa, 2003; Fitzmaurice et al., 2004)

$$E(Y_{i,t_j}) = E\left\{E(Y_{i,t_j}|\mathtt{b}_i)\right\} = E\left[g^{-1}(\mathtt{x}_{i,t_j}^T\beta + \mathtt{z}_{i,t_j}^T\mathtt{b}_i)\right],$$

que, em geral, não pode ser simplificado devido à presença de funções não lineares em $g^{-1}(.)$ e, portanto,

$$E\left[g(Y_{i,t_j}|\mathtt{b}_i)\right] \neq g\left[E(Y_{i,t_j}|\mathtt{b}_i)\right]. \tag{6.1}$$

Note-se que no modelo linear a introdução de efeitos aleatórios não altera a interpretação dos parâmetros dos efeitos fixos, dado que a igualdade (6.1) é verificada. Existe apenas alteração na interpretação dos parâmetros dos efeitos fixos quando a função de ligação canónica dos GLMM não é linear (Fitzmaurice et al., 2004).
De acordo com Cabral and Gonçalves (2011), nos modelos lineares generalizados mistos podemos considerar as seguintes interpretações para os parâmetros fixos e para a componente aleatória do modelo:

(i) **Parte fixa**: As componentes β_j, $j = 0, 1, ..., p+1$, do vetor β têm uma interpretação em termos específicos do indivíduo. Elas representam a influência das covariáveis de interesse na alteração da resposta média esperada de um determinado indivíduo. Estes coeficientes, β, são designados por *coeficientes de regressão específicos do indivíduo* dado que a interpretação de β_j depende dos efeitos aleatórios do $i-$ésimo indivíduo assumirem um valor fixo.

(ii) **Parte aleatória:** Uma forma de interpretar as estimativas das variâncias dos efeitos aleatórios passa por considerar percentis dos efeitos aleatórios baseados na hipótese Gaussiana e, com base nos percentis obtidos calcular os limites de variação dos valores esperados. Outra maneira é considerar a representação gráfica dos perfis de cada indivíduo.

6.2.3 Inferência

Para que se possa aplicar a metodologia dos modelos lineares generalizados mistos a um conjunto de dados, há necessidade, após a formulação do modelo que se pensa adequado, de proceder à realização de inferências sobre esse modelo. A inferência nos GLMM é baseada na verosimilhança. Com efeito, não só o método da máxima verosimilhança é o método de eleição para estimar os parâmetros β e os parâmetros em D, como também os testes de hipóteses sobre os parâmetros do modelo são baseados na verosimilhança.

6.2.3.1 Estimação dos parâmetros do modelo

À semelhança do que acontece nos GLM, a estimação dos parâmetros nos modelos lineares generalizados mistos também recorre ao método da máxima verosimilhança. Nesta situação, a estimação de β e dos parâmetros em D é feita através da maximização da função de verosimilhança, obtida através da integração da função de verosimilhança sobre os efeitos aleatórios $\mathtt{b}_i$.
Como explicitaram Cabral and Gonçalves (2011) e Hedeker (2005), a contribuição do i-ésimo indivíduo para a função de verosimilhança é

$$L_i^A(\beta, \phi, D) = \int \prod_{t=1}^{T_i} f(y_{it}|\mathtt{b}_i, \beta, \phi) f(\mathtt{b_i}|D) d\mathtt{b}_i, \tag{6.2}$$

onde $\prod_{t=1}^{T_i} f(y_{it}|\mathtt{b}_i, \beta, \phi) = L_i^F(\beta, \phi, D)$ diz respeito à função verosimilhança do modelo com efeitos aleatórios e $f(\mathtt{b_i}|D)$ representa a distribuição dos efeitos aleatórios, assumida como sendo uma distribuição normal multivariada. O expoente A na função de verosimilhança indica que se trata da função de verosimilhança do modelo com efeitos aleatórios (Cabral and Gonçalves, 2011).
Já a função de verosimilhança para a amostra, contendo os n indivíduos é expressa por

$$\begin{aligned} L^A(\beta, \phi, D) &= \prod_{i=1}^{n} L_i(\beta, \phi, D) \\ &= \prod_{i=1}^{n} \int L_i^F(\beta, \phi, D) f(\mathtt{b}_i|D) d\mathtt{b}_i, \end{aligned} \tag{6.3}$$

onde o expoente F na função de verosimilhança indica que se trata da função de verosimilhança do modelo com efeitos fixos. Tomando o logaritmo da função verosimilhança, tem-se, naturalmente,

$$l^A(\beta, \phi, D) = \sum_{i=1}^{n} \log L_i^A(\beta, \phi, D). \tag{6.4}$$

Maximizando a função (6.4), obtém-se os estimadores de máxima verosimilhança dos parâmetros de regressão β e da matriz de variância-covariância dos efeitos aleatórios $\mathtt{b}_i$. Contudo, ao contrário do que acontece nos GLM, onde estes parâmetros são facilmente estimados através do método iterativo de mínimos quadrados reponderados, nos modelos lineares generalizados mistos este processo não funciona (Zuur et al., 2009). Neste tipo de modelos, a maximização da função log $-$verosimilhança é bastante complexa e o problema, geralmente, não tem uma solução analítica. Consequentemente, para que se possa inferir com base na verosimilhança, tem de se recorrer ao cálculo numérico. Como qualquer método, o cálculo numérico possui limitações, podendo tornar a obtenção das estimativas um processo computacionalmente intensivo, aumentando o seu tempo de resposta quando se está na presença de grandes amostras e/ou quando a estrutura do modelo é complexa (p .ex: a dimensão do espaço paramétrico é grande).

Como referem Cabral and Gonçalves (2011), de forma a resolver este problema têm sido propostas várias aproximações numéricas: (i) as que são baseadas na aproximação de dados (consideram o desenvolvimento da série de Taylor e inserem-se na metodologia do método da quasi-verosimilhança penalizada (PQL), proposto por Breslow and Clayton (1993) e o método da quasi−verosimilhança marginal apresentado por Goldenstein (1991)); (ii) as que são baseadas na aproximação da função a integrar (têm como base a aproximação de Laplace) e, por fim (iii) as que são baseadas na aproximação do integral (salienta-se, por exemplo, o método numérico da quadratura de Gauss-Hermite). Outros métodos têm sido implementados na tentativa de evitar os problemas de integração numérica, como por exemplo, a utilização de Modelos de Cadeias de Markov via métodos de Monte Carlo (Clayton, 1996).

6.2.3.2 Testes de hipóteses, análise de resíduos e seleção do modelo

Nos modelos lineares generalizados mistos, para comparar a significância estatística dos parâmetros de regressão β pode ser utilizado o teste de Wald, tal como nos GLM (ver secção 2.3.3).
Já quando se pretende comparar a estrutura fixa de dois modelos encaixados que possuam os mesmos efeitos aleatórios recorre-se, geralmente, a testes de razão de verosimilhança.
Nestes modelos a análise de resíduos é efetuada através de análises gráficas.
O critério de seleção de modelos, quando estes não são encaixados, passa por considerar o critério de informação de *Akaike* (AIC) ou o critério de informação Bayesiana (BIC).
Cabral and Gonçalves (2011) alertam para o facto de que, em qualquer modelo deve ter-se em atenção qual o método utilizado na maximização da função de verosimilhança, dado que esta tem de se basear nos dados e não na aproximação aos dados.
Caso se pretenda testar a presença de efeitos aleatórios, tem-se um problema fronteira (Cabral and Gonçalves, 2011) e, portanto, assume-se que a distribuição assintótica sob a hipótese nula para uma estatística de teste de razão de verosimilhanças é uma mistura de qui-quadrados. Contudo, a biblioteca *lme4* do R não está preparada para efetuar este teste.

6.3 Modelo Marginal: Equações de Estimação Generalizadas

O modelo marginal é um modelo utilizado para analisar dados longitudinais e é obtido através da extensão do modelo linear generalizado (GLM) do caso univariado para o multivariado. Este modelo foi proposto por Liang and Zeger (1986) e o termo "marginal"significa, neste contexto, que o valor médio da variável resposta depende apenas das covariáveis de interesse e não de efeitos aleatórios. Como Diggle et al. (2002) referem, este modelo tem a vantagem de o valor esperado da variável resposta e a correlação, proveniente de o mesmo indivíduo possuir várias observações e estas apresentarem dependência, serem modelados separadamente. O modelo marginal é apropriado quando se pretende fazer inferências acerca da população e não acerca do indivíduo.
A estimação dos parâmetros deste modelo é feita usando o método de equações de estimação generalizadas (método GEE), que se baseia na estimação por quasi−verosimilhança e, portanto, não necessita de especificar a distribuição para o vetor da variável resposta (Fitzmaurice et al., 2004).
Relativamente aos dados omissos, este modelo apenas admite o mecanismo de omissão de dados MCAR.

6.3.1 Estrutura do modelo marginal

Em conformidade com a secção 6.2.1, considere-se y_{i,t_j} o valor da resposta observada para o indivíduo $i (i = 1, ..., n)$ no instante t_j $(j = 1, ..., T_i)$. y_{i,t_j} é a realização da variável resposta Y_{i,t_j} e $\mathtt{Y}_i = (Y_{i,t_1}, ..., Y_{i,t_{T_i}})^T$ representa o vetor das variáveis resposta para o indivíduo i. O perfil de cada indivíduo é assim dado pelo vetor $\mathtt{y_i} = (y_{i,t_1}, ..., y_{i,t_{T_i}})^T$.
Seja $\mathtt{x}_{i,t_j}^T = (t_j, x_{1,i,t_j}, ..., x_{p,i,t_j})$ o vetor das covariáveis de interesse associado a cada Y_{i,t_j}, de dimensão $(p+1)$, a matriz das covariáveis, $\mathtt{X}_i$, de ordem $T_i \times (p+1)$ e o vetor dos coeficientes de regressão $\beta = (\beta_1, ..., \beta_{p+1})^T$ de dimensão $(p+1)$.

Assuma-se também que Y_{i,t_j} segue uma distribuição pertencente à família exponencial cuja densidade marginal é definida por:

$$f(y_{i,t_j}) = exp\left\{\frac{\omega_{i,t_j}}{\phi}(y_{i,t_j}\theta_{i,t_j} - b(\theta_{i,t_j})) + c(y_{i,t_j}, \phi)\right\}.$$

À semelhança do que é reportado na literatura, no decorrer deste trabalho, as especificações impostas ao modelo marginal serão sempre acompanhadas do termo "marginal", para que se realce o facto de que não se está a condicionar o modelo a efeitos aleatórios não observáveis.

Como mostram Cabral and Gonçalves (2011), Diggle et al. (2002), Fitzmaurice et al. (2004) e Zuur et al. (2009), o modelo marginal pode ser especificado em 3 partes:

1. O valor esperado marginal para cada variável resposta,

$$E(Y_{i,t_j}|x_{i,t_j}) = \mu_{i,t_j} = b'(\theta_{i,t_j}),$$

assume-se dependente das variáveis explicativas x_{i,t_j}, através de uma função de ligação

$$g(\mu_{i,t_j}) = \eta_{i,t_j} = \mathtt{x}_{i,t_j}^T\beta, \tag{6.5}$$

onde g(.) é uma função de ligação conhecida.

2. A variância marginal, condicionada pelas variáveis explicativas x_{i,t_j}, depende do valor médio marginal, de acordo com

$$Var(Y_{i,t_j}|x_{i,t_j}) = v_{i,t_j} = V(\mu_{i,t_j})\frac{\phi}{\omega_{i,t_j}} = b''(\theta_{i,t_j})\frac{\phi}{\omega_{i,t_j}},$$

onde V é uma função de variância conhecida, ϕ é o parâmetro de dispersão e ω_{i,t_j} é um peso conhecido atribuído às observações da variável resposta.

3. A correlação entre Y_{i,t_j} e Y_{i,t_k} é uma função do valor médio marginal e, possivelmente, de parâmetros adicionais, α. Esta função define-se como $corr(Y_{i,t_j}, Y_{i,t_k}) = \rho(\mu_{i,t_j}, \mu_{i,t_k})$, onde $\rho(.)$ é uma função conhecida.

Como é possível constatar, a extensão dos modelos lineares generalizados para o modelo marginal é feita na especificação 3, com a incorporação da correlação entre as respostas pertencentes ao mesmo indivíduo.

No modelo marginal, os parâmetros de regressão β descrevem as alterações da média da resposta da população ao longo do tempo e o modo como estas alterações estão relacionadas com as covariáveis (Fitzmaurice et al., 2004). Com efeito, a interpretação dos parâmetros é efetuada através do risco relativo.
A título de exemplo, considere-se o modelo marginal para contagens, de acordo com as 3 especificações já definidas.

Modelo marginal para contagens

Considere-se que Y_{i,t_j}, a variável resposta do indivíduo i na instante t_j, segue uma distribuição de Poisson. Nesta situação,

1. A variância marginal, condicionada pelas variáveis explicativas é

$$Var(Y_{i,t_j}|x_{i,t_j}) = \phi\mu_{i,t_j},$$

onde ϕ é o parâmetro de dispersão, que indica a existência de sobredispersão, caso $\phi > 1$.

2. A média μ_{i,t_j} está relacionada com a componente sistemática do modelo (que supomos incluir um termo constante) através da equação

$$log(\mu_{i,t_j}) = (1, x_{i,t_j}^T)^T\beta = \beta_0 + \beta_1 x_{1,i,t_j} + ... + \beta_p x_{p,i,t_j} + \beta_{p+1}t_j,$$

onde log é a função de ligação usualmente escolhida para um modelo de contagem.

3. A associação entre as medições de um mesmo indivíduo é dada por uma estrutura de correlação não definida,

$$corr(Y_{it}, Y_{ik}) = \alpha_{jk}, \quad j \neq k.$$

Contudo, podem ser consideradas outras estruturas de correlação, que serão posteriormente apresentadas na secção 6.3.2.1 .

6.3.2 Inferência

Como descrito na formulação do modelo marginal, este não necessita de especificar qualquer distribuição para a variável resposta. Sendo assim, toda a inferência neste modelo é baseada na função quasi−verosimilhança.

6.3.2.1 Estimação dos parâmetros do modelo: equações de estimação generalizadas

Dado que no modelo marginal não é necessário especificar a distribuição de $\mathtt{Y_i}$, métodos de estimação por máxima verosimilhança não podem ser utilizados e portanto, é necessário encontrar uma alternativa para a estimação dos parâmetros deste modelo.
O método GEE proposto por Liang and Zeger (1986) apresenta-se como alternativa. Este método é a extensão multivariada do método de estimação por quasi−verosimilhança apresentado anteriormente por Wedderburn (1974). A abordagem deste método é baseada no conceito de *estimação de equações* e permite a análise de respostas correlacionadas (Fitzmaurice et al., 2004). Por outras palavras, o método GEE envolve as mesmas especificações que o método de estimação por quasi−verosimilhança mas possui uma especificação adicional, a matriz de covariância do vetor resposta, $\mathtt{Y_i}$.
Cabral and Gonçalves (2011) referem que o método GEE para β apresenta a mesma forma das equações *score* para os modelos univariados, apenas difere a forma de escolher a matriz de variância-covariância.
Como escrito anteriormente, nos dados longitudinais, as observações de diferentes indivíduos

assumem-se independentes, enquanto que as observações sobre o mesmo indivíduo tendem a ser correlacionadas. Tendo em conta estas particularidades, o método GEE para β é expresso por

$$\mathtt{S}^*_\beta(\beta,\alpha) = \sum_{i=1}^{n} \mathtt{D}_i^T \mathtt{V}_i[(\alpha)]^{-1}(\mathtt{Y_i} - \mu_i) = 0, \tag{6.6}$$

onde $\mathtt{S}^*_\beta(\beta,\alpha)$ é a versão multivariada da função quasi−score, $\mathtt{D}_i$ é uma matriz $T_i \times (p+2)$ que contém as primeiras derivadas de μ_i em ordem aos parâmetros de regressão β ($\mathtt{D}_i = \frac{\partial \mu_i}{\partial \beta}$) e onde μ_i é o vetor do valor esperado de $\mathtt{Y}_i$. A matriz $\mathtt{V}_i(\alpha)$ é uma matriz diagonal $T_i \times T_i$ que para o $i-$ésimo indivíduo é dada por (Cabral and Gonçalves, 2011; Diggle et al., 2002; Fitzmaurice et al., 2004)

$$\mathtt{V}_i(\alpha) = \phi \left(\mathtt{A}_i^{1/2} \mathtt{R}_i(\alpha) \mathtt{A}_i^{1/2} \right), \tag{6.7}$$

conhecida como matriz de variância-covariância "corrente". Fitzmaurice et al. (2004) referem que a aplicação do termo "corrente"é utilizada porque se supõe inicialmente que esta matriz não esteja bem especificada em termos de variâncias do modelo e associações sobre o mesmo indivíduo. As matrizes $\mathtt{A}_i$, em (6.7), são matrizes diagonais com $Var(Y_{it}) = \phi V(\mu_{it})$ ao longo da diagonal − e portanto, $\mathtt{A}_i^{1/2}$ é uma matriz diagonal com os desvio padrão, $\sqrt{\phi V(\mu)_{it}}$, ao longo da diagonal − e, $\mathtt{R}_i(\alpha) = corr(\mathtt{Y}_i)$ é uma matriz diagonal $T_i \times T_i$ conhecida como matriz de correlação "corrente (*working correlation matrix*)". Como $\mathtt{R}_i(\alpha)$ representa a correlação entre as observações de um mesmo indivíduo, ajustada pelas covariáveis presentes no modelo, os valores que α pode assumir estão no intervalo [-1,+1] (Agranonik, 2009).
Cabral and Gonçalves (2011) e Fitzmaurice et al. (2004) mencionam que quando a matriz $\mathtt{R}_i(\alpha)$ está corretamente especificada, podemos considerar que $\mathtt{V}_i(\alpha) = Var(\mathtt{Y}_i)$. Cabral and Gonçalves (2011) aludem também que, caso a matriz $\mathtt{R}_i(\alpha)$ seja a matriz identidade, o método GEE reduz-se às equações de quasi−verosimilhança para um modelo linear generalizado que assume que as medições repetidas são independentes.
No método GEE, a estimação do vetor β depende de α e ϕ, que por sua vez dependem de β para serem estimados. Para que seja possível obter uma solução para o sistema de equações descrito em (6.6), é necessário recorrer ao processo iterativo proposto por Liang and Zeger (1986):

1. Calcular as estimativas iniciais de β, $\widehat{\beta}^{(0)}$ a partir do ajustamento a um modelo linear generalizado (isto é, assume-se independência sobre as observações).

2. Dado as estimativas de β, calcular α e ϕ. Estas estimativas são obtidas através dos resíduos de Pearson (Cabral and Gonçalves, 2011; Fahrmeir and Tutz, 2001; Fitzmaurice et al., 2004),

$$e_{i,t_j} = \frac{y_{i,t_j} - \widehat{\mu}_{i,t_j}}{\sqrt{Var(\widehat{\mu}_{i,t_j})}},$$

onde o parâmetro de dispersão ϕ é estimado através de

$$\widehat{\phi} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{i=1}^{T_i} e^2_{i,t_j}$$

e a estimação de α é obtida de forma equivalente, dependendo da estrutura de correlação definida. Exemplos de estruturas de correlação são propostas por Liang and Zeger (1986) e estão explicitadas na tabela 6.1.

Tabela 6.1: Estruturas de correlação propostas por Liang and Zeger (1986) (Fonte: Cabral and Gonçalves (2011)).

	$corr(Y_{i,t_j}, Y_{i,t_k})$	**Estimativa** α
Independente	0	–
Dependente	α	$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{T_i(T_i-1)\sum_{t\neq k} e_{i,t_j} e_{i,t_k}}$
AR(1)	$\alpha^{\lvert t_j - t_k\rvert}$	$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{T_i(T_i-1)\sum_{t<T_i-1} e_{i,t_j} e_{i,t_j+1}}$
Não Definida	$\alpha_{t_j t_k}$	$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} e_{i,t_j} e_{i,t_k}$

Posteriormente, calcula-se a estimativa da matriz de correlação corrente, $\mathtt{R}_i(\alpha)$ e por fim as variâncias estimadas, através de $\mathtt{V}_i(\alpha)$.

3. Atualizam-se as novas estimativas de β, de acordo com:

$$\widehat{\beta}^{(l+1)} = \widehat{\beta}^{(l)} + \left[\sum_{i=1}^{n} \widehat{\mathtt{D}}_i^T \left[\mathtt{V}_i(\widehat{\alpha})\right]^{-1} \widehat{\mathtt{D}}_i\right]^{-1} \times \left[\sum_{i=1}^{n} \widehat{\mathtt{D}}_i^T \left[\mathtt{V}_i(\widehat{\alpha})\right]^{-1} (y_i - \widehat{\mu}_i)\right]^{-1}. \tag{6.8}$$

4. Repete-se os passos 2 e 3 até o processo convergir.

Estruturas de Correlação

Especificar a matriz de correlação corrente, $\mathtt{R}_i(\alpha)$, de forma correta, aumenta a eficiência das estimativas dos parâmetros do modelo, o que se revela importante quando a correlação entre as observações sobre o mesmo indivíduo for alta.
Nesta secção serão apresentadas as estruturas de correlação propostas por Liang and Zeger (1986) para a matriz $\mathtt{R}_i(\alpha)$. A título de exemplo, fixe-se $T = 4$.

- **Estrutura Independente**

Neste caso, a matriz $\mathtt{R}_i(\alpha)$ é dada por

$$\mathtt{R}_i(\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

e esta é utilizada quando se assume independência entre as observações. Quando não se tem conhecimento prévio acerca da estrutura de correlação, esta estrutura é geralmente uma boa escolha.

- **Estrutura Dependente ou Permutável**

A matriz $\mathtt{R}_i(\alpha)$ é

$$\mathtt{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}.$$

A palavra "permutável"significa que se pode trocar a ordem de quaisquer duas observações sem alterar a estrutura de correlação. Isto porque, esta estrutura considera que a correlação entre observações de um mesmo indivíduo é a mesma.

- **Estrutura Autorregressiva de primeira ordem – AR(1)**

A matriz $\mathtt{R}_i(\alpha)$ é expressa por,

$$\mathtt{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}.$$

Esta estrutura expressa que a correlação entre observações sobre o mesmo indivíduo decresce com o tempo. Como explicam Zuur et al. (2009), a correlação entre observações separadas por uma unidade de tempo (esta unidade de tempo é dependente do problema; corresponde ao tempo entre o qual as observações são recolhidas) é representada por α e tende a ser mais similar do que a correlação entre observações separadas por duas (representada por α^2) ou mais unidades de tempo.

- **Estrutura Não Definida**

Nesta situação, a matriz $\mathtt{R}_i(\alpha)$ é

$$\mathtt{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 \end{pmatrix},$$

onde se assume que, entre cada observação sobre o mesmo indivíduo existe um valor de correlação diferente. Esta matriz requer especial atenção, na media em que, mesmo considerando um reduzido número de observações para cada indivíduo, implica um elevado número de parâmetros para estimar (Zuur et al., 2009).

6.3.2.2 Testes de hipóteses, análise de resíduos e seleção do modelo

Como o método GEE é baseado na estimação por quasi−verosimilhança, o teste da razão de verosimilhanças não pode ser aplicado. Contudo, tendo em atenção a distribuição assintótica dos estimadores GEE, o teste de Wald pode ser usado (Cabral and Gonçalves, 2011). Para a seleção do modelo, não é possível aplicar o critério de informação de Akaike face à estimação dos parâmetros não ser baseada na verosimilhança. Com o intuito de solucionar este problema, Pan (2001) criou um critério estatístico, semelhante ao AIC, mas considerando o facto de que as observações possam ser correlacionadas. Este critério é conhecido como *critério de quasi−verosimilhança sob o modelo de independência*, *QIQ* (*Quasi-likelihood under the Independence model Criterion*). Contudo, este critério não se encontra ainda implementado no *software* R.
Neste método, a análise de resíduos, efetuada através da análise gráfica, é um método de diagnóstico bastante popular mas de difícil análise.

6.4 Resultados da aplicação dos modelos longitudinais −Parte 2

Como tem vindo a ser referido, o objetivo da aplicação deste trabalho está inserido no projeto CIBIO e visa estudar o efeito de diferentes variáveis ambientais sobre a abundância de aves das

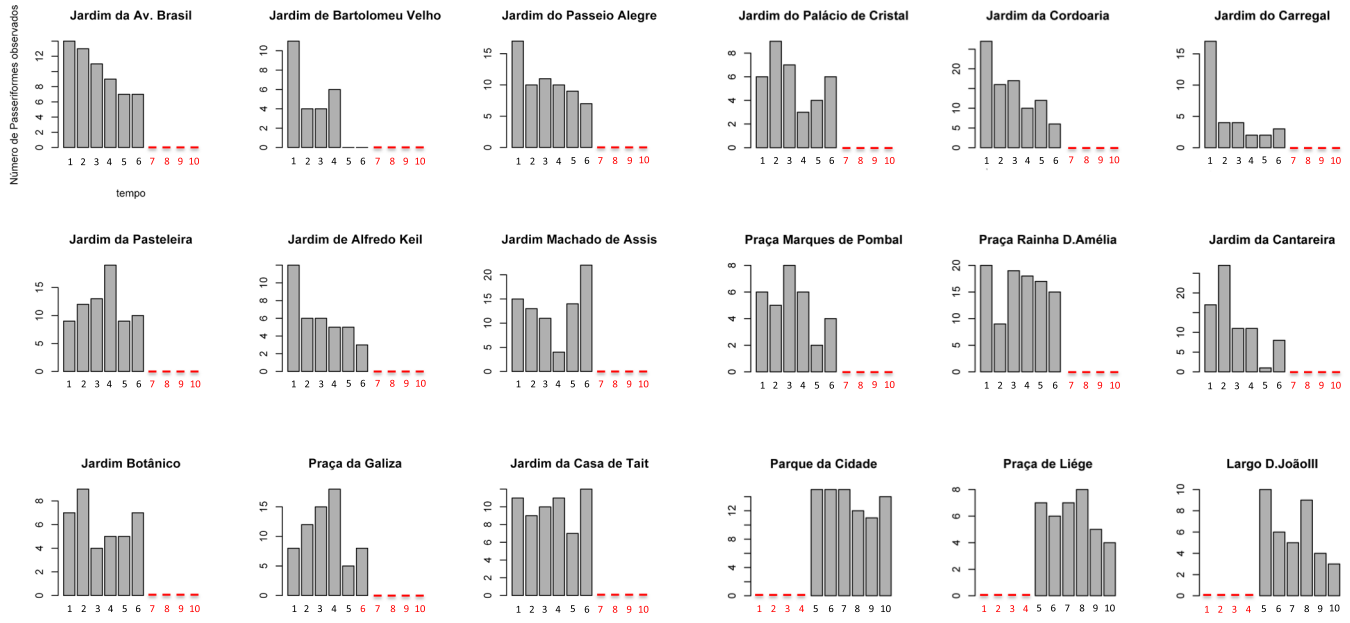
ordens *Columbiformes* e *Passeriformes* nos espaços verdes de acesso público da cidade do Porto. Neste secção, pretende-se analisar qual o comportamento das aves da ordem *Passeriformes* ao longo de vários meses tendo em conta diversas variáveis ambientais.
Para este estudo longitudinal consideraremos dois modelos: o modelo linear generalizado misto e o modelo marginal.
A base de dados e preparação da análise introduz esta secção, seguindo-se da análise exploratória aos dados. Por fim, os dados em estudo são modelados através dos modelos longitudinais abordados nesta dissertação.
A implementação dos diversos modelos na análise de dados longitudinais será efetuada no software R (R Development Core Team, 2012), onde o nível de significância é fixado em 0.05. Todas as bibliotecas necessárias para a implementação destes modelos serão devidamente mencionadas ao longo da análise efetuada nesta secção.

6.4.1 Base de dados e preparação da análise

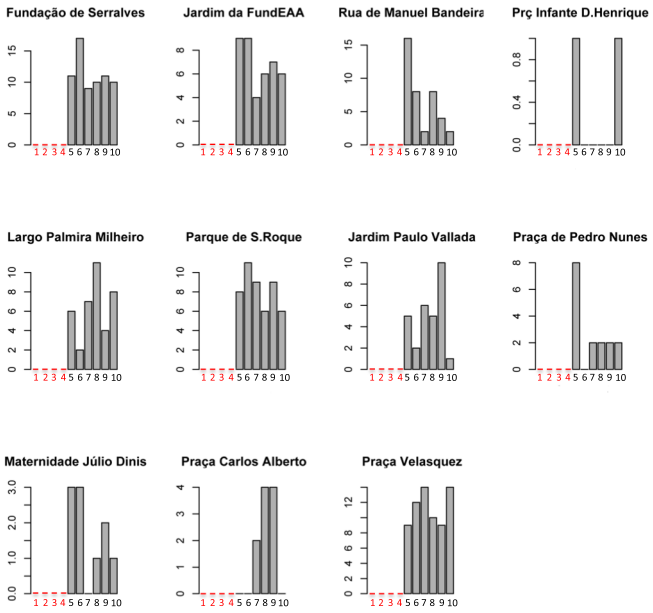
Em conformidade com o que foi mencionado na apresentação do projeto CIBIO (capítulo 4), os dados longitudinais em estudo foram amostrados em 10 meses distintos (aconselha-se a visualização do esquema apresentado na figura 4.3 para um perceção clara acerca dos dados em análise). Contudo, para cada jardim, existem 4 meses em que estes não são amostrados, pelo que as suas observações neste período vão ser omissas. Apesar da seleção dos jardins não ter sido aleatória, a variável resposta não teve qualquer influência sobre as observações omissas, isto é, o facto de terem sido amostrados jardins diferentes nas duas épocas (apenas existem dois meses comuns) foi opção do CIBIO e nada teve a ver com o número de *Passeriformes* observados. Deste modo, consideramos que o mecanismo de omissão presente nos dados longitudinais em estudo é MCAR. Alertamos ainda para o facto de estarmos cientes do mecanismo de seleção dos jardins não ser aleatório. Todavia, para que seja possível aplicar a metodologia referente aos dados longitudinais torna-se necessário assumirmos que os jardins em estudo foram escolhidos aleatoriamente.
Neste trabalho, cada um dos jardins corresponde a um indivíduo e o conjunto dos 29 jardins equivale à população.
Da base de dados facultada pelo CIBIO constam, no total, 29 jardins tendo sido 15 jardins amostrados na 1ª época e 14 na 2ª época, conforme já explicado. Para além disso, a base de dados possui 29 espécies de aves da ordem *Passeriformes* e 7 variáveis ambientais (Tabela 4.1).
Para cada jardim foi registado, em 6 instantes diferentes, o número de *Passeriformes* observados num período de 8 minutos (Rabaça, 1995). A amostragem foi efetuada pelas primeiras horas da manhã e numa área do jardim que a equipa do CIBIO considera privilegiada. O registo do número de aves observadas foi efetuado de acordo com a observação direta das aves ou pelo canto das mesmas (método dos transetos pontuais sem limite de distância). Devido à extensão de alguns dos jardins em estudo, revelou-se necessário considerar mais do que um ponto de localização para a amostragem. Para os jardins em causa, considerou-se ainda que o número de aves observadas da ordem *Passeriformes* é o número médio entre os diversos pontos de localização. Tal decorre do pressuposto de que os pontos de localização são independentes e, portanto, as aves que se observam num determinado ponto não são as mesmas que se observam num outro ponto de localização.
O facto de nos dados longitudinais existirem dois meses que possuem 29 observações levanta uma questão relevante acerca da causa de não se ter considerado um destes meses na análise efetuada no capítulo 5. Tal situação não ocorreu dado que isso implicaria a escolha de um deles, o que não nos pareceu razoável. Deste modo, mantivemos a análise inicial apenas com 15 jardins.

Análise exploratória
Antes de se ajustar os dados em estudo a um modelo, é necessário perceber como é que estes se

comportam. Comece-se por analisar qual o número de *Passeriformes* observados em cada um dos 29 jardins amostrados ao longo dos 10 tempos considerados.



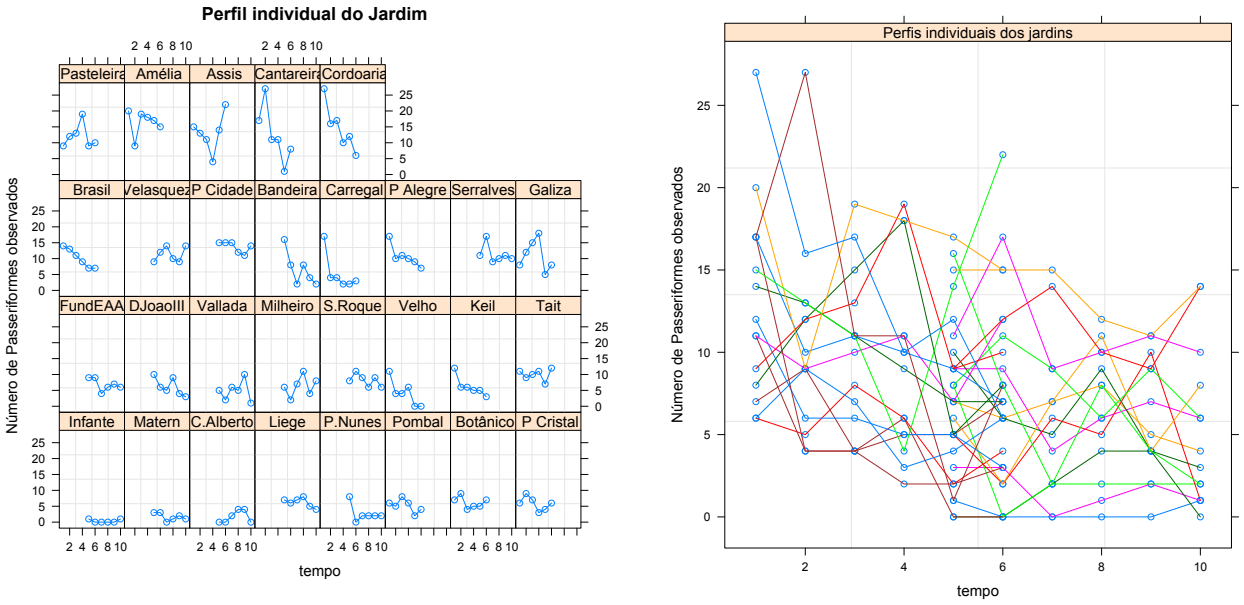
(a) (b)



(c)

Figura 6.1: Número de *Passeriformes* observados nos 29 Jardins.

O tracejado vermelho presente nas figuras diz respeito aos meses em que o respetivo jardim não foi amostrado, originando valores omissos. Este realce foi efetuado para que o leitor consiga interpretar facilmente a diferença entre o número de *Passeriformes* observados ser nulo e a existência de valores omissos nos jardins, num determinado mês.
À exceção de 5 jardins, registou-se sempre, pelo menos, uma ave da ordem *Passeriformes* nos meses em que foram amostrados. A figura 6.1 revela-nos claramente que o número de *Passeriformes* observados em cada um dos jardins ao longo do tempo é bastante distinto.
Através da função *groupedData()* da biblioteca *nlme* do software R, é possível traçar o perfil individual de cada um dos jardins. Esta função é um data frame que possui informação adicional acerca do agrupamento das observações e do papel de algumas variáveis.

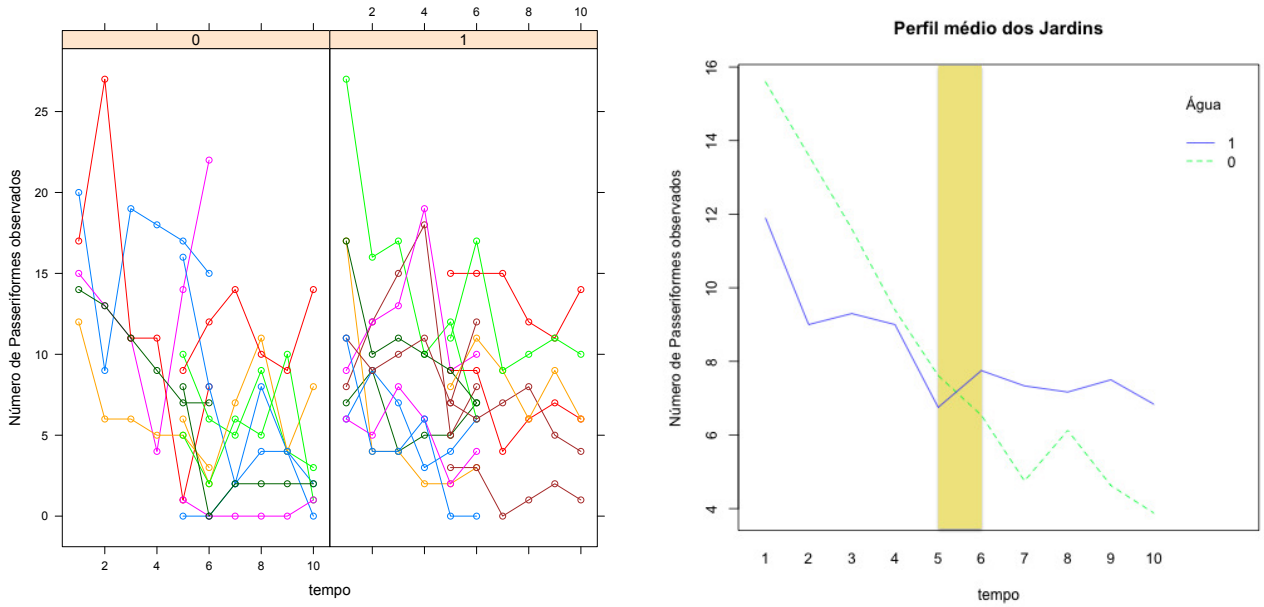


(a) (b)

Figura 6.2: Perfil dos jardins.

O perfil individual de cada jardim mostra ser completamente diferente de jardim para jardim (figura 6.2a). O facto de existirem dois meses onde o número de jardins amostrados é superior aos restantes (tempos 5 e 6), é observável na figura 6.2b com a deteção de um maior número de pontos nestes meses, que dizem respeito às observações de um determinado jardim. Devido à disparidade dos perfis, encontrar um modelo que se ajuste corretamente aos dados tornar-se-á uma tarefa árdua, uma vez que não existe uma tendência comum a todos os jardins, ao longo do tempo. Note-se que, mesmo que fossem consideradas apenas uma das duas épocas de amostragem (1^a época - do tempo 1 ao tempo 6 e 2^a época - do tempo 5 ao tempo 10), os jardins pertencentes a cada uma delas revelam também não possuir uma tendência comum entre eles (figura 6.2).
Como referido, os dados longitudinais em estudo resultam da amostragem efetuada em diferentes épocas, pelo que se torna relevante averiguar se as épocas contêm ou não jardins que sejam significativamente diferentes para cada uma das variáveis ambientais em análise.
Para que possamos inferir acerca da possível diferença entre os jardins das duas épocas, recorrer-se-á a testes de hipóteses. Neste estudo, consideramos que todas as variáveis ambientais em estudo são contínuas, à exceção da variável área de plano de água que foi dicotomizada, passando a ser

denominada de variável *presença de água*. A classe de referência da variável presença de água é o "0"e indica que não existe água no jardim em estudo. Nos jardins da 1ª época, 10 em 15 possuem água, enquanto que na 2ª época, possuem água, 6 dos 14 jardins.
Para as variáveis contínuas aplicar-se-á o teste de hipóteses de Mann-Whitney-Wilcoxon sob a hipótese nula de igualdade entre as medianas da variável ambiental em estudo para cada uma das épocas. O teste de hipóteses foi calculado no software R através da função *wilcox.test()*. Todas as variáveis ambientais, excluindo a variável *distância ao rio*, obtiveram um *valor-p* superior a 0.05. Logo, com um nível de significância de 0.05 não existe evidência estatística para rejeitar a hipótese nula e, portanto, nada se pode concluir acerca de H_1. A variável *distância ao rio* apresentou um *valor-p* de 0.02 pelo que é possível afirmar que se rejeita a hipótese nula com um nível de significância de 0.05. Isto significa que a mediana da variável distância ao rio é diferente nos jardins das duas épocas em análise.
Por fim, para a variável *presença de água* é aplicado o teste de hipóteses do qui-quadrado, com 1 grau de liberdade, recorrendo à função *chisq.test()*. Sob a hipótese nula de que a distribuição de água pelos jardins é igual nas duas épocas, o teste retornou um *valor-p* igual a 0.8619. Logo, com um nível de significância de 0.05 não se consegue rejeitar a hipótese nula e, por conseguinte, nada se pode concluir acerca da independência dos jardins.
É consensualmente aceite (Brown and Freitas, 2002) que a presença de água num jardim é um factor de grande influência sobre o número de *Passeriformes* observados.
Na figura 6.3a encontra-se representado o perfil dos jardins em estudo ao longo dos 10 tempos considerados, consoante a presença/ausência de água nos mesmos. No retângulo do "0"estão representados os perfis dos jardins amostrados que não possuem água e no retângulo "1", os perfis dos jardins que contam com a presença de água. A figura 6.3b mostra-nos qual o perfil médio dos jardins em função do factor água.

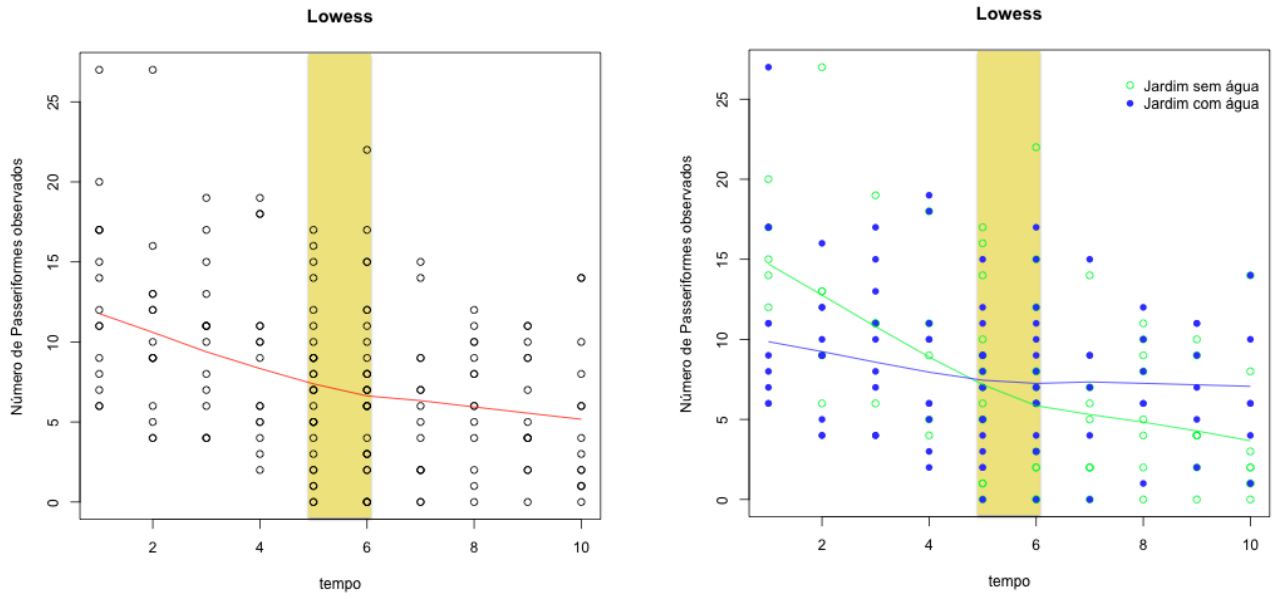


(a) (b)

Figura 6.3: Perfil dos jardins em função do fator água.

A figura 6.3a evidencia que não existe grande diferença no comportamento dos *Passeriformes* nos jardins com e sem água. Analisando estes jardins separadamente pode-se afirmar que ambos não

possuem qualquer tendência, uma vez que o número de *Passeriformes* observados varia bastante de jardim para jardim.
Na figura 6.3b o número de *Passeriformes* mostra ser superior nos jardins sem água até ao mês 5, invertendo-se esta tendência a partir do mês 6. Contudo, um ponto a ter em consideração neste gráfico é o facto de, até ao mês 6 termos 15 jardins da 1ª época e a partir do mês 5 termos 14 jardins da 2ª época, diferentes dos primeiros. Assim, para cada uma das épocas, não parece existir interação entre a existência de água e o tempo.
Face às análises apresentadas nesta secção, torna-se evidente a dificuldade em perceber qual o tipo de ajustamento que deve ser feito. Uma forma simples de termos uma ideia do tipo de ajustamento passa por considerar um método não paramétrico. Um dos vários métodos disponíveis no R é o *Robust Locally Weighted Regression - Lowess* (Cleveland, 1979). Este método é robusto a *outliers* e pode ser aplicado através da rotina *lowess()*. Ao contrário dos métodos paramétricos, que estimam o preditor linear globalmente, este método estima o preditor linear na vizinhança de cada ponto de interesse, isto é, faz uma regressão local. A figura 6.4 diz respeito ao ajustamento pelo método Lowess.



(a) (b)

Figura 6.4: Ajustamento pelo método Lowess.

Na figura 6.4a está representada a curva ajustada pelo método Lowess para todos os jardins em análise. Já na figura 6.4b encontram-se evidenciados o número de *Passeriformes* nos jardins que possuem (ou não) água, ao longo do tempo, tais como as respetivas curvas de ajustamento. Parece existir uma relação do tipo linear entre o número de *Passeriformes* e o tempo.

Concluída a análise exploratória, proceder-se-á à aplicação do modelo linear generalizado misto e do modelo marginal no ajustamento aos dados longitudinais em estudo. O objetivo da análise do projeto CIBIO visa, como citado anteriormente, determinar qual a abundância de *Passeriformes* nos jardins de acesso público da cidade do Porto. Isto significa, que se pretende inferir sobre os 29 jardins (população) e não sobre cada jardim especificamente. Face ao propósito do estudo, e de acordo com a metodologia apresentada no capítulo 6, o modelo mais indicado seria o modelo

marginal. Contudo, aplicar-se-á também o modelo linear generalizado misto uma vez que todos os jardins em estudo são independentes entre si e, por conseguinte, torna-se interessante descrever a alteração da média da resposta de cada jardim e a relação destas com as covariáveis.

6.4.1.1 Modelo linear generalizado misto

Como fundamentado na secção 6.2, o modelo linear generalizado misto é utilizado quando se pretende fazer inferências acerca do indivíduo e não da população.
Antes de ajustarmos os dados a um modelo linear generalizado misto com resposta que segue uma distribuição de Poisson, dado que estamos a lidar com dados de contagem, é importante antecipar quais os efeitos aleatórios que devem ser considerados. Com base na análise exploratória, foi possível constatar que o perfil de cada jardim comportou-se de forma distinta ao longo dos 10 meses amostrados (figura 6.2). Considere-se o ajustamento dos dados ao método não paramétrico Lowess, para cada um dos jardins.

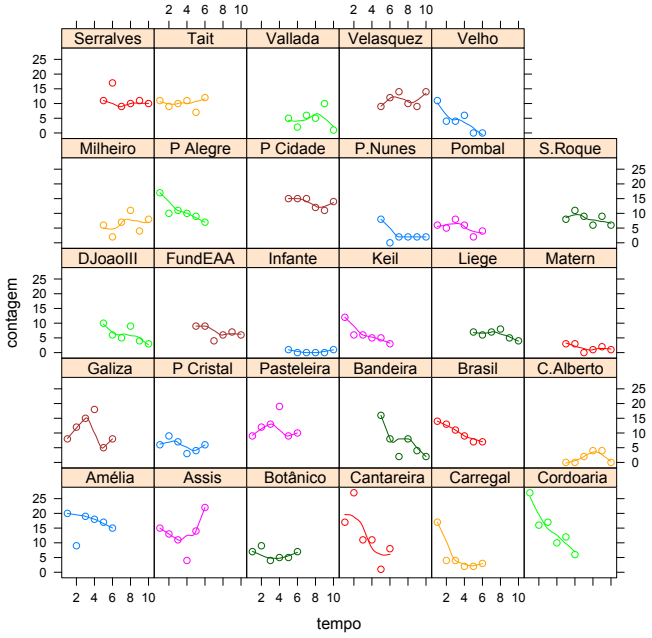


Figura 6.5: Curvas de regressão Lowess.

Como já visto na análise exploratória, detetamos um comportamento diferente para os vários jardins. As curvas de regressão Lowess apresentam ordenadas na origem e declives bastante diferentes em cada um deles, pelo que devemos incluir um efeito aleatório para o tempo.
Para o efeito, foram ajustados aos dados diversos modelos, com diferentes combinações de variáveis ambientais e efeitos aleatórios. O modelo eleito foi determinado de forma heurística através do critério de informação de Akaike. O modelo que faz um melhor ajustamento aos dados é expresso por,

- **Modelo:** $log(\mu_{i,t_j}) = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) * tempo_{j-1} + \beta_2 * \log(afm)_{c_i} + \beta_3 * I_{agua_i}$,

onde o índice i corresponde ao jardim i e o índice j ao tempo j. b_{0i} representa o efeito aleatório

na constante (o que significa que para cada jardim, a ordenada na origem pode ser diferente), b_{0i} representa o efeito aleatório no declive (cada jardim pode possuir um declive distinto ao longo do tempo), $\log(afm)_c$ diz respeito à covariável área de coberto fanerófito logaritmizada e centrada e I_{agua} é a variável área de plano de água dicotomizada à qual passamos a denominar por variável presença de água. À semelhança do estudo dos modelos lineares generalizados e dos modelos com um número excessivo de zeros, a covariável afm foi logaritmizada devido ao tamanho elevado da sua escala. Neste estudo, esta mesma covariável foi ainda centrada, uma vez que apresentava um elevado valor de correlação com a constante do modelo. Já a covariável *área de plano de água* foi dicotomizada devido à grande amplitude e valores nulos que apresenta. Para que seja possível interpretar a constante e^{β_0}, efetuamos uma translação na covariável *tempo*, pelo que consideramos a covariável $tempo_{j-1}$ no modelo. Deste modo, o mês de abril de 2011 passa a ser interpretado como o mês zero e assim sucessivamente para os restantes meses em estudo. Na ausência do efeito aleatório, a constante e^{β_0} representa o número esperado de *Passeriformes* no mês de abril de 2011, num jardim com área de coberto fanerófito igual à média da amostra e sem água.
A aplicação deste modelo no R foi executada através da função *glmer*() da biblioteca *lme*4. Esta função ajusta os dados a um modelo linear generalizado misto utilizando o método numérico de Laplace. No software R existem outras funções que podem ser utilizadas no ajustamento a um modelo linear generalizado misto, como por exemplo a função *glmmML()* da biblioteca *glmmML*, a função *glmmPQL()* da biblioteca *MASS* e a função *glmmadmb()* da biblioteca *glmmADMB*. A nossa escolha recaiu sobre a função *lmer()* por ser a mais utilizada nos diversos estudos revistos.
O output resultante do modelo escolhido é o apresentado na tabela 6.2.

Tabela 6.2: Ouput do modelo linear generalizado misto com resposta Poisson.

Efeitos Aleatórios			
	Variância	**Desvio padrão**	**Correlação**
b_{0i}	0.2032	0.4508	
b_{1i}	0.0100	0.1001	-0.540
Efeitos Fixos			
Variável	**Coef.**	**Erro padrão**	**valor-p**
constante	2.6301	0.1600	$< 2e-16$
tempo	-0.1248	0.0243	$2.84e-07$
$\log(afm)_c$	0.2617	0.0639	$4.14e-05$
I_{agua}	-0.3974	0.1955	0.0421
Correlação dos efeitos fixos			
	constante	*tempo*	$\log(afm_c)$
tempo	-0.515		
$\log(afm_c)$	0.323	- 0.010	
I_{agua}	-0.719	0.072	-0.515
AIC:319			

O modelo eleito apresenta um AIC de 319 e todas as variáveis presentes revelam ser estatisticamente significativas. O output do modelo mostra também que a correlação existente entre os efeitos fixos presentes no modelo não apresenta valores elevados. Interpretando os efeitos aleatórios do modelo, tem-se que a estimativa de variância de b_{0i} é de 0.2032 sinalizando a existência de variabilidade entre os jardins no número esperado de *Passeriformes* no mês de abril de 2011. A estimativa de variância para b_{1i} é de 0.01, o que significa que não existe grande variabilidade no declive dos jardins ao longo do tempo.
Neste modelo, o número esperado de *Passeriformes* no mês de abril de 2011, num jardim com área

de coberto fanerófito igual à média da amostra e sem água é de $e^{2.63010} \simeq 14$. Nos modelos lineares generalizados mistos a interpretação de risco relativo não faz sentido uma vez que o modelo em estudo não prevê valores para a população mas sim para o indivíduo. Nesta situação, para que seja possível aplicar o risco relativo, todos os jardins terão de ter os mesmos efeitos aleatórios b_0 e b_1, ou seja, na prática não podem existir efeitos aleatórios e, portanto, não se aplica o risco relativo neste modelo. A função $ranef()$ do R permite obter as estimativas dos efeitos aleatórios presentes no modelo em causa.

A figura 6.6 aclara os intervalos de confiança para os efeitos aleatórios. Da análise pormenorizada da figura, ressalta bastante sobreposição dos intervalos de confiança em qualquer um dos efeitos aleatórios mas também constatamos intervalos de confiança disjuntos, especialmente no que se refere à constante. Todavia, dado que o modelo em causa revelou ser o melhor modelo, em comparação com os diversos modelos que foram considerados, decidiu-se manter os efeitos aleatórios. A remoção dos mesmos não faria sentido uma vez que não se teria em conta a correlação presente nos dados.

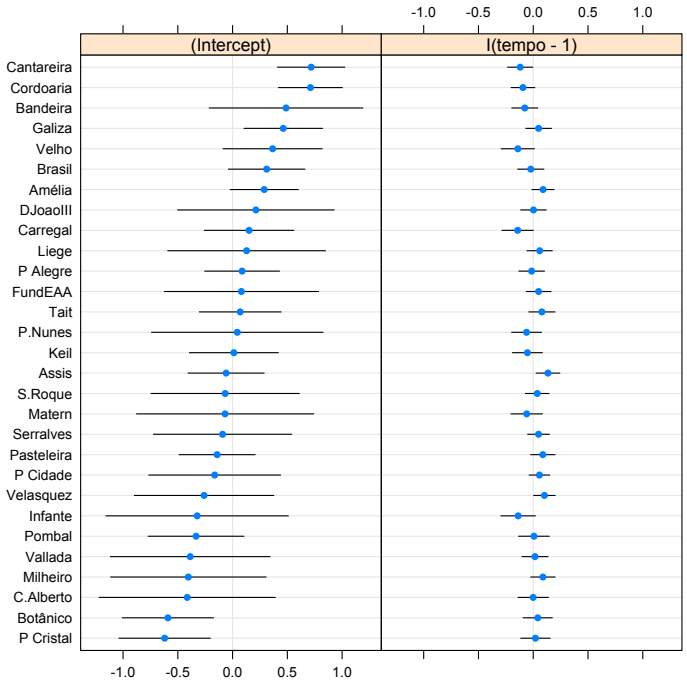
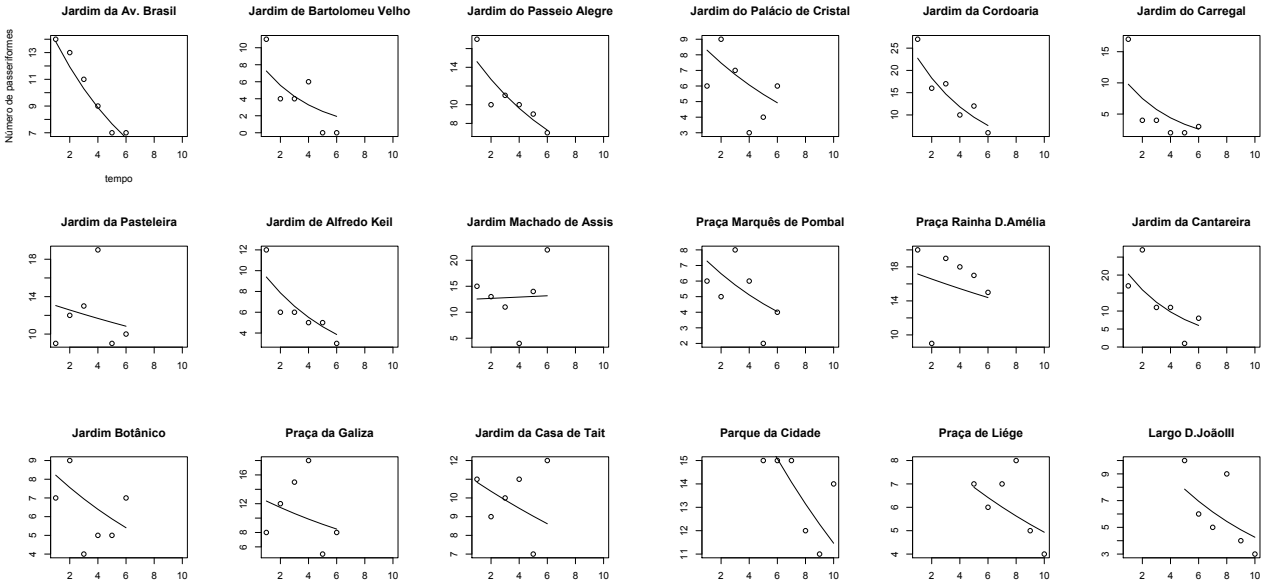
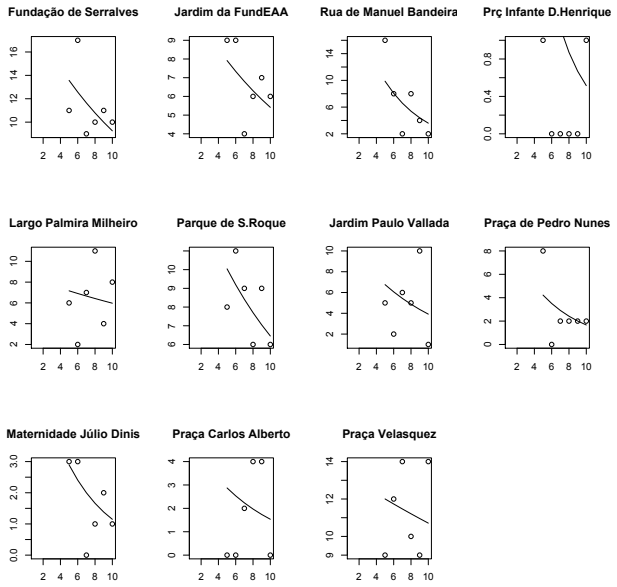


Figura 6.6: Intervalo de confiança para os efeitos aleatórios.

A figura 6.7 mostra-nos o número de *Passeriformes* observados e previstos pelo modelo, para cada jardim.



(a) (b)



(c)

Figura 6.7: Número de *Passeriformes* observados (pontos) e previstos pelo modelo de efeitos aleatórios (curvas).

Da análise da figura 6.7 diríamos que os jardins da Avenida Brasil e do Passeio Alegre parecem ser os jardins onde o ajustamento revelou ser melhor. Interpretando os valores ajustados, observam-se curvas decrescentes em todos os jardins, à exceção do jardim Machado de Assis.

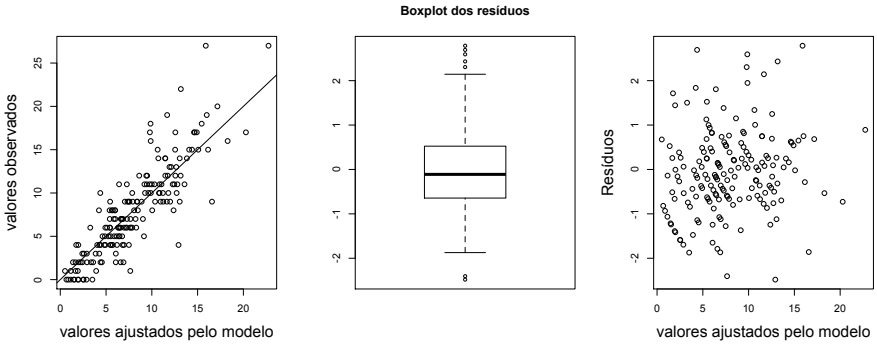


Figura 6.8: Gráficos de diagnóstico do modelo linear generalizado misto.

Os gráficos de diagnóstico apontam para a existência de alguns outliers moderados no modelo, contudo, os resíduos não assumem valores muito elevados.
A função *glmer()* não permite efetuar o ajustamento dos dados a modelos de regressão Binomial Negativa e Quasi-Poisson com efeitos aleatórios.
Das restantes funções disponibilizadas pelo R para ajustar os modelos em causa, vários investigadores recomendam o uso da função da função *glmmadmb* para ajustar um GLMM com distribuição Binomial Negativa. Neste estudo, executou-se essa instrução mas a convergência do método falhou, pelo que não foi possível obter nenhum resultado. Já o ajustamento de um GLMM com distribuição Quasi-Poisson pôde ser efetuado utilizando a função *glmmPQL*. Obtiveram-se estimativas para os coeficientes semelhantes às do GLMM para uma distribuição de Poisson com, surpreendentemente, valores de erro padrão também muito semelhantes. Este facto leva-nos a crer que o modelo de Poisson com efeitos aleatórios está, neste caso, a ser o mais adequado.

6.4.1.2 Modelo marginal

Para que seja possível aplicar o modelo marginal aos dados longitudinais em análise é necessário definir qual a estrutura de correlação a usar. Devido às caraterísticas dos dados, decidiu-se eleger a estrutura de correlação autorregressiva de primeira ordem (AR(1)). Esta estrutura define que a correlação entre observações de um mesmo jardim decresce com o tempo, o que faz sentido para os dados em análise uma vez que, por exemplo, o número de *Passeriformes* existentes no mês de abril de 2011 está mais relacionado com o número de *Passeriformes* existentes no mês de junho de 2011 do que nos meses seguintes. A estrutura de correlação autogressiva de primeira ordem foi também a escolhida por Zuur et al. (2009) em estudos acerca da abundância de aves e do número de chamamentos efetuados por aves bebés na ausência dos progenitores num determinado período de tempo.
Uma outra estrutura de correlação que nos parece ser adequada é a estrutura de correlação não definida. Contudo, os dados em análise são manifestamente insuficientes para a estimação de tantos parâmetros.
No modelo marginal a estimação dos parâmetros não é efetuada de acordo com a máxima verosimilhança e, portanto, não se consegue estimar um valor de AIC para os modelos. Deste modo, o melhor modelo obtido resultou da aplicação do teste de Wald, univariado ou multivariado, sobre as diversas variáveis (Fitzmaurice et al., 2004; Zuur et al., 2009).

O modelo eleito apresenta-se da seguinte forma:

- $log(\mu_{i,t_j}) = \beta_0 + \beta_1 * tempo_{j-1} + \beta_2 * \log(afm)_{c_i} + \beta_3 * I_{agua_i}$,

com uma matriz de correlação do tipo AR(1). À semelhança do modelo linear generalizado misto, a variável tempo foi translacionada e a variável $\log(afm)$ foi centrada para que se conseguisse obter uma interpretação para a constante e^{β_0}.
A tabela 6.3 apresenta o output do modelo marginal eleito para os dados em análise.

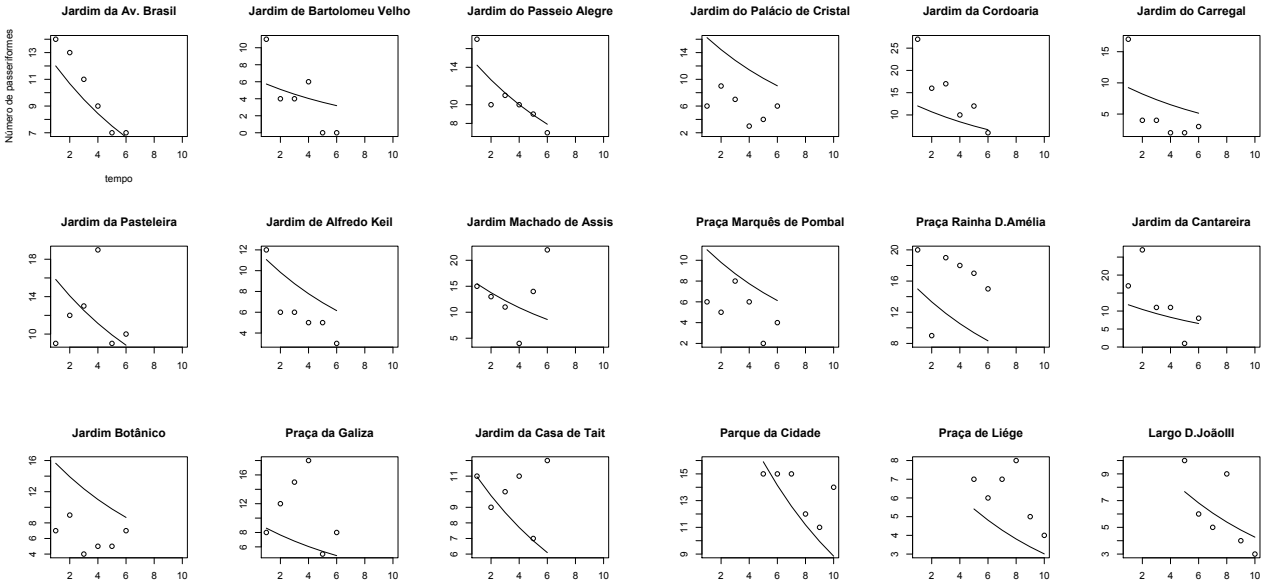
Tabela 6.3: Output do modelo marginal com resposta Poisson.

Variável	**Coef.**	**Erro Padrão**	**valor-p**
constante	2.7782	0.1287	<2e-16
tempo	-0.1174	0.0226	2.05e-07
$\log(afm)_c$	0.2435	0.0461	1.30e-07
$I_{agua}(ref.0)$	-0.4593	0.1503	0.0022
	$\widehat{\phi}$	**Erro Padrão**	
dispersão(ϕ)	2.269	0.2989	
Parâmetro de correlação	$\widehat{\alpha}$	**Erro Padrão**	
α	0.592	0.0818	

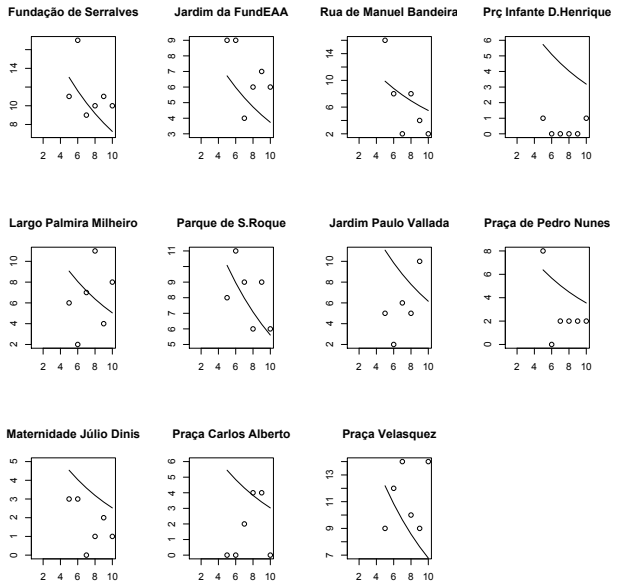
O output do modelo revela que todas as variáveis presentes são estatisticamente significativas. A correlação existente entre o número de *Passeriformes* em dois instantes de tempo consecutivos é estimada em 0.592; caso o intervalo de tempo seja de duas unidades, a correlação passará a ser 0.592^2=0.35 e assim sucessivamente para os restantes intervalos. O parâmetro de dispersão ϕ apresenta um valor de 2.27 o que indica que existe alguma sobredispersão no modelo. Neste mesmo modelo, o número esperado de *Passeriformes* no mês de abril de 2011, num jardim com área de coberto fanerófito igual à média da amostra e sem água é de $e^{2.7782} \simeq 16$. O modelo marginal ajustado aos dados longitudinais pode ser expresso por:

$$\begin{aligned} E(Y_{i,t_j}|x_{i,t_j}) &= \mu_{i,t_j} = e^{2.7782-0.1174*tempo_{j-1}+0.2435*\log(afm_c)_i-0.4593*I_{agua_i}} \\ Var(Y_{i,t_j}|x_{i,t_j}) &= 2.27 * \mu_{i,t_j} \\ corr(Y_{i,t_j}, Y_{i,t_k}) &= 0.592^{|t_j-t_k|}. \end{aligned}$$

De facto, o modelo aqui apresentado é um modelo Quasi-Poisson, onde a variável resposta segue uma distribuição Quasi-Poisson com função de probabilidade igual à da distribuição de Poisson e com média μ_{i,t_j} e variância de 2.27*μ_{i,t_j}.
Até ao momento, nenhuma das funções do R que tem em conta a estimação de equações generalizadas, *gee*, *geeglm* e *yags* inclui rotinas que considerem a distribuição Binomial Negativa. Existem, no entanto, softwares como o SAS e o Stata onde o método GEE para esta distribuição está já implementado (Hilbe, 2011).



(a) (b)



(c)

Figura 6.9: Número de *Passeriformes* observados (pontos) e previstos pelo modelo marginal (curvas).

A análise comparativa entre a figura 6.7 e a figura 6.9 evidencia uma ligeira superioridade do modelo linear generalizado misto com resposta Poisson. Estando os resultados próximos, parece-nos razoável admitir que o modelo aqui reproduzido tem um desempenho bastante satisfatório. Na figura 6.10 estão representados alguns gráficos de diagnóstico do modelo em estudo.

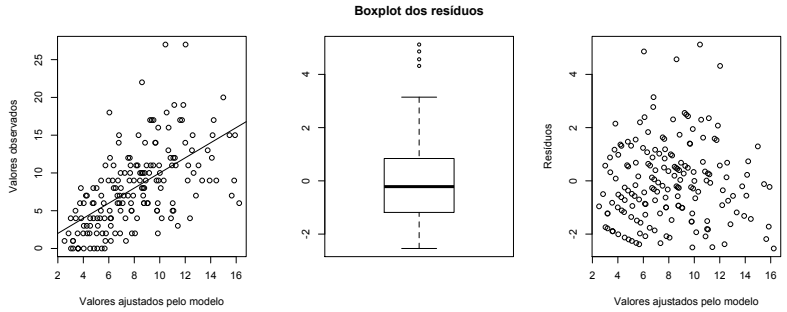


Figura 6.10: Gráficos de diagnóstico do modelo marginal.

Existem, de facto, alguns outliers no modelo marginal eleito e observações com resíduos ligeiramente elevados. Contudo, o modelo marginal escolhido parece fazer um ajustamento minimamente razoável aos dados em estudo.
Resta-nos averiguar qual o efeito das variáveis do modelo sobre o número esperado de *Passeriformes* nos jardins públicos da cidade do Porto.
Para o estudo em análise, é importante perceber-se como se comporta a abundância de *Passeriformes* em função das variáveis ambientais presentes no modelo eleito. Para tal, recorremos à interpretação do risco relativo.
Comece-se por interpretar o efeito da variável *tempo*.

$$\begin{aligned} RR &= \frac{E(N_{pass}|tempo+1)}{E(N_{pass}|tempo)} = exp(-0.1174) = 0.89 \\ &\Longrightarrow \quad E(N_{pass}|tempo+1) = E(N_{pass}|tempo) - 0.11E(N_{pass}|tempo) \end{aligned}$$

De acordo com o risco relativo, por cada unidade aumentada no *tempo* observa-se, em média, menos 11% de *Passeriformes* nos jardins públicos da cidade do Porto. Este resultado já era expectável depois da análise da figura 6.9 revelar que o modelo previu que a maioria dos jardins obteve curvas decrescentes.

O risco relativo para o variável *área de coberto fanerófito* é expresso por:

$$\begin{aligned} RR &= \frac{E(N_{pass}|log(afm)_c+1)}{E(N_{pass}|log(afm)_c)} = exp(0.2435) = 1.28 \\ &\Longrightarrow \quad E(N_{pass}|log(afm)_c+1) = E(N_{pass}|log(afm)_c) + 0.28E(N_{pass}|log(afm)_c) \end{aligned}$$

Pela análise do risco relativo conclui-se que, por cada unidade aumentada no $\log(afm)_c$, o número médio de *Passeriformes* observados aumenta 28%. Tenha-se em atenção que aumentar uma unidade no $\log(afm)_c$ significa multiplicar a *área de coberto fanerófito* pela exponencial de $1(\simeq 2.72)$. O resultado obtido é concordante com os estudos de Daniels and Kirkpatrick (2006), Green and Baker (2003) e Posa and Sodhi (2006) que afirmam que a presença de árvores e arbustos aumentam a presença de espécies de aves.

Por fim, resta interpretar o risco relativo para a variável *presença de água*:

$$\begin{aligned} RR &= \frac{E(N_{pass}|agua)}{E(N_{pass}|\overline{agua})} = exp(-0.4593) = 0.63 \\ &\Longrightarrow \quad E(N_{pass}|agua) = E(N_{pass}|\overline{agua}) - 0.37E(N_{pass}|\overline{agua}) \end{aligned}$$

O resultado do risco relativo revela que existem, em média, menos 37% de *Passeriformes* nos jardins que possuem água. Este resultado contradiz o que é reportado na literatura (Brown and Freitas, 2002) e o senso comum, de que a água influencia positivamente a presença de aves num espaço verde. Do ponto de vista biológico, o resultado obtido pode estar inflacionado por outras variáveis que não estão presentes no modelo e que podem condicionar a presença de *Passeriformes*, assim como pelo facto de se terem agrupado as espécies em ordens. Os jardins sem água, ao oferecerem menores condições para a maior parte das espécies, vão potenciar uma maior abundância das espécies mais resistentes e generalistas, como os pardais, ao diminuir a competição interespecífica. Este argumento corrobora com o facto de nos jardins sem água que foram amostrados terem sido detetados muitos pardais, o que nos leva a crer que a elevada abundância desta espécie pode ter influenciado as contagens. Esta situação decorre de termos agrupado as espécies em ordens, que podem incluir espécies com requisitos muito diferentes.

Capítulo 7

Conclusões e trabalho futuro

O propósito deste trabalho centrou-se no estudo de modelos de regressão para dados de contagem e na aplicação desses modelos a dados sobre abundância de aves. Mais concretamente, o CIBIO-UP colocou-nos o problema da identificação de variáveis ambientais que sejam determinantes para o número de *Columbiformes* e *Passeriformes* existentes nos espaços verdes públicos da cidade do Porto. As variáveis amostradas foram a *área total dos jardins*, a *área de coberto fanerófito*, a *área permeável*, a *área de plano de água*, a *distância ao mar* e a *distância ao rio*.

Os dados recolhidos sobre o número de *Passeriformes* durante o mês de junho de 2011 foram interpretados através de um modelo de regressão Quasi-Poisson. Nessa análise, a única variável que mostrou ter um efeito significativo sobre a resposta foi a *área de coberto vegetal*, influenciando positivamente o número esperado de *Passeriformes*. Este resultado corrobora aliás, com as conclusões de um artigo de Farinha-Marques et al. (2011b). Um outro modelo que também se revelou satisfatório foi o modelo de regressão de Poisson, após a eliminação de um dos jardins amostrados. De entre os efeitos significativos, identificou-se um efeito positivo da *área de coberto vegetal* e um efeito negativo da *presença de aves da ordem Columbiformes* sobre o número esperado de Passeriformes.

Os dados recolhidos sobre o número de *Columbiformes* durante o mesmo período de tempo apresentaram um excesso de zeros na distribuição de frequências pelo que foram analisados por modelos que tivessem esse facto em consideração. Os modelos que apresentaram uma melhor qualidade de ajustamento foram o modelo de regressão de Poisson com zeros inflacionados e o modelo de regressão de Poisson com barreira, ambos selecionando os mesmo efeitos significativos. Na componente de contagem dos dois modelos, os coeficientes foram muito semelhantes, resultado que não é estranho uma vez que todos os zeros foram classificados no modelo ZIP como zeros falsos. As variáveis explicativas significativas foram a *área de coberto vegetal*, com um efeito negativo no número esperado de *Columbiformes* e a *presença de água*, com um efeito positivo. Na componente que modela as presenças *versus* ausências, as estimativas dos efeitos significativos (para um nível de significância de 0.06) foram também semelhantes nos dois modelos. Nessas estimativas, observaram-se sinais contrários decorrentes da definição de π em cada um dos modelos. O modelo ZIP evidenciou a existência de uma associação negativa entre a presença de zeros falsos e a *área de coberto fanerófito*; o modelo de Poisson com barreira estimou uma associação positiva entre a presença de *Columbiformes* e essa mesma variável.
Face aos resultados obtidos, a escolha entre o modelo ZIP e o modelo de regressão de Poisson com barreia para a modelação do número esperado de *Columbiformes* pareceu ser uma questão irrelevante dado que ambos conduzem ao mesmo resultado.

Devido a alguma sobredispersão presente nos modelos, consideraram-se modelos de regressão com um número excessivo de zeros que usassem a distribuição Binomial Negativa na componente de contagem. Para o conjunto de variáveis ambientais em análise, não foi identificado qualquer efeito significativo, mesmo para um nível significância de 0.10.

Os dados recolhidos sobre o número de *Passeriformes* ao longo dos 10 meses do estudo do CIBIO-UP foram interpretados através de um modelo de regressão de Poisson com um efeito aleatório na variável *tempo*. Foi obtida significância estatística para os coeficientes correspondentes às variáveis *área de coberto fanerófito*, *tempo* e *presença de água*. Foram também considerados modelos de regressão Binomial Negativa e Quasi-Poisson com efeitos mistos; para a primeira distribuição não foi possível obter convergência do algoritmo numérico de estimação dos parâmetros, enquanto que os resultados obtidos para a distribuição Quasi-Poisson foram muito semelhantes aos da distribuição de Poisson.
Para o estudo de dados longitudinais acerca dos *Passeriformes* foi ainda ajustado o modelo marginal de Poisson com matriz de correlação autorregressiva de ordem 1, AR(1). A análise revelou que o número esperado de *Passeriformes* diminui à medida que se avança no *tempo* e que as variáveis *área de coberto fanerófito* e *presença de água* têm associações positiva e negativa, respetivamente com esse número.

No nosso entender, as principais limitações deste estudo são o processo de amostragem e o reduzido tamanho amostral. Quando existe mais do que um ponto de localização no mesmo jardim, o processo de amostragem consiste da amostragem de um ponto durante 8 minutos; passado esse tempo, o mesmo observador desloca-se para o próximo ponto de localização pré-definido e procede da mesma forma. Ora, o facto de os diversos pontos de localização não serem amostrados em simultâneo pode originar que os indivíduos a amostrar se consigam deslocar até outro ponto e serem amostrados novamente. Face a esta situação, é ainda dúbia, a escolha mais acertada acerca das medidas estatísticas a considerar entre os vários pontos de localização, isto é, número máximo ou médio entre os indivíduos amostrados.
O facto de os dados registarem apenas observações de 29 jardins (ou mesmo 15) condicionou muito o número de variáveis explicativas nos modelos considerados. Uma alternativa às análises apresentadas teria sido considerar modelos não paramétricos.
No caso dos dados longitudinais, o facto de os jardins amostrados serem diferentes nas duas épocas, inviabilizou que se pudesse averiguar a existência de alterações no número médio de *Passeriformes* do ano 2011 para o ano 2012.
O facto de se terem agrupado as espécies em ordens levou a que se perdessem algumas caraterísticas específicas das espécies.

Como trabalho futuro, gostaríamos de estudar modelos de regressão com um número excessivo de zeros para dados longitudinais. Tais modelos encontram-se apenas parcialmente implementados no R, tanto quanto sabemos, apesar de existirem já várias publicações sobre o assunto. Uma breve incursão na revisão da literatura mostrou que Hall (2000) foi o primeiro a incluir efeitos aleatórios nos modelos de regressão de Poisson com zeros inflacionados, ainda que apenas tenha considerado efeitos aleatórios no modelo de contagem. Também, Yau and Lee (2001) propuseram a introdução de efeitos aleatórios independentes nos dois processos do modelo com barreira. Já Min and Agresti (2005) estenderam os modelos apresentados por Yau and Lee (2001) ao considerarem que os efeitos aleatórios nas duas componentes do modelo com barreira podem ser correlacionados. Dobbie and Welsh (2001) aplicaram o modelo marginal recorrendo ao método de estimação de equações generalizadas nas duas componentes do modelo com barreira. A implementação computacional da análise de dados longitudinais com um número excessivo de zeros na linguagem R é ainda

bastante escassa. No R, é possível efetuar ajustamentos a modelos de regressão de Poisson com zeros inflacionados e com efeitos aleatórios apenas no processo de contagem, através da função *glmmadmb()* da biblioteca *glmmADMB*. Para além disso, esta função só permite que o modelo de inflação de zeros consista de uma constante. A inclusão de efeitos aleatórios nas duas componentes do modelo com barreira é permitida na função *glmmadmb()*. A biblioteca *MCMCglmm* permite ajustar modelos lineares generalizados mistos através do método de Monte Carlo via Cadeias de Markov. A rotina *MCMCglmm()* dessa mesma biblioteca consente a inclusão de efeitos aleatórios em dados com inflação de zeros.

Referências

Agranonik, M. (2009). Equações de estimação generalizadas (gee): Aplicação em estudo sobre mortalidade neonatal em gemelares de porto alegre, rs(1995-2007). Master's thesis, Faculdade de Medicina - Universidade Federal do Rio Grande do Sul, Porto Alegre.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., New Jersey, second edition.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Bermudez, P. and Bispo, R. (2012). Statistical methodologies applied to environmental and ecological problems. In *Book of Abstracts of the Joint Meeting of y-BIS and jSPE*, page 78, Campus da Caparica, Lisbon.

Böhning, D., Dietz, E., and Schlattmann, P. (1997). Zero-inflated count models and their applications in public health and social science. In *Applications of Latent Trait and Latent Class Models in the Social Sciences*, pages 333–344.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Brooks, S. (1999). Bayesian analysis of animal abundance data via mcmc. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 6 - Proceedings of the Sixth Valencia International Meeting*, pages 723–732. Oxford University Press.

Brown, K. and Freitas, A. (2002). Butterfly communities of urban forest fragments in campinas, são paulo, brazil: structure, instability, environmental correlates, and conservation. *Journal of insect conservation*, 6:217–231.

Bruun, B., Delin, H., and Svensson, L. (2002). *Aves de Portugal e Europa*. Fapas - Fundo para a Protecção dos Animais Selvagens, third edition.

Cabral, M. S. and Gonçalves, M. H. (2011). *Análise de Dados Longitudinais*. Sociedade Portuguesa de Estatística.

Cameron, A. and Trivedi, P. (1986). Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1(1):29–53.

Cameron, A. and Trivedi, P. (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York, first edition.

Cameron, A. C. and Trivedi, P. K. (2008). *Regression analysis of count data*. Cambridge University Press, New York, seventh edition.

Catry, P., Costa, H., Elias, G., and R.Matias (2010). *Aves de Portugal. Ornitologia do território continental*. Assírio & Alvim, Lisboa.

Cheung, Y. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Satistics in Medicine*, 21(10).

Clayton, D. (1996). Generalized linear mixed models. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice*, pages 275–301. Chapman & Hall, London.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cohen, A., Peters, H., and Foote, L. (1960). Calling behavior of mourning doves in two midwest life zones. *Journal of Wildlife Management*, (203-212).

Consul, P. (1989). *Generealized Poisson distribution: properties and applications*. Marcel Dekker, New York.

Costa, H., Juana, E., and J.Varela (2011). *Aves de Portugal: Incluindo os arquipélagos dos Açores, Madeira e das Selvagens*. Lynx e Spea.

Costa, S. C. (2003). *Modelos Lineares Generalizados Mistos para Dados Longitudinais*. PhD thesis, Escola Superior de Agricultura Luiz de Queiroz - Universidade de São Paulo, São Paulo.

Crooks, K. R., Suarez, A. V., and Bolger, D. T. (2001). Extinction and colonization of birds on habitat islands. *Conservation Biology*, 15(1):159–175.

Cyr, A. (1981). Limitation and variability in hearing ability in censuring birds. *Studies in Avian Biology*, 6:327–333.

Dalrymple, M. L., Hudson, I., and Ford, R. (2003). Finite mixture, zero-inflated poisson and hurdle models with application to sids. *Computational Statistics & Data Analysis*, 41:491–504.

Daniels, G. and Kirkpatrick, J. (2006). Does variation in garden characteristics influence the conservation of birds in suburbia? *Biological conservation*, 133:326–335.

Dawson, D. (1981). Counting birds for a relative measure (index) of density. *Studies in Avian Biology*, 6:12–16.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press Inc, New York, second edition.

Dobbie, M. and Welsh, A. (2001). Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics*, 43(4):431–444.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, USA, second edition.

Edwards, D. P., Hodgson, J. A., Hamer, K. C., Mitchell, S. L., Ahmad, A. H., Cornell, S. J., and Wilcove, D. S. (2010). Wildlife-friendly oil palm plantations fail to protect biodiversity effectively. *Conservation Letters*, 3:236–242.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, second edition.

Famoye, F. and Singh, K. P. (2006). Zero-inflated poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4:117–130.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, Boca Raton.

Farinha-Marques, P., Fernandes, C., Lameiras, J., Silva, S., Leal, I., and Guilherme, F. (2011a). *Estudo da Relação entre a Morfologia do Espaço Público e a Diversidade de Flora e Fauna na cidade do Porto. Livro I - Metodologia para a selecção das áreas de estudo*. CIBIO-UP, Porto.

Farinha-Marques, P., Lameiras, J. M., Fernandes, C., Silva, S., and Guilherme, F. (2011b). Urban biodiversity: a review of current concepts and contributions to multidisciplinary approaches. *Innovation: The European Journal of Social Science Research*, 24(3):247–274.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., New York.

Forcey, G., Thogmartin, W., Linz, G., Bleier, W., and McKann, P. (2011). Land use and climate influences on waterbirds in the prairie potholes. *Journal of Biogeography*, 38:1694–1707.

Gaio, A. R. (2011-2012). Apontamentos escritos da disciplina eace: Mestrado em engenharia matemática.

Garden, J. G., Mcalpine, C. A., Possingham, H. P., and Jones, D. N. (2007). Habitat structure is more important than vegetation composition for local-level management of native terrestrial reptile and small mammal species living in urban remnants: A case study from brisbane, australia. *Austral Ecology*, 32:669–685.

Goldenstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78:45–51.

Grandell, J. (1997). *Mixed Poisson Processes*. Chapman & Hall, Great Britain, first edition.

Green, D. and Baker, M. (2003). Urbanization impacts on habitat and bird communities in a sonoran desert ecosystem. *Landscape and urban planning*, 63:225–239.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. Technical Report EC -94-10, New York University, Leonard N. Stern School of Business, Department of Economics.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56:1030–1039.

Hedeker, D. (2005). Generalized linear mixed models. In Everitt, B. S. and Howell, D., editors, *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Inc, New York.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press, United Kingdom, second edition.

Hoef, J. M. V. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*, volume 1. John Wiley & Sons, Inc., second edition.

Ismail, N. and Jemain, A. A. (2007). Handling overdispersion with negative binomial and generalized poisson regression models. *Casualty Actuarial Society Forum*, Winter:103–158.

Januário, F., Gaio, R., Fernandes, C., and Farinha-Marques, P. (2012). Abundância relativa das aves das ordens columbiformes e passeriformes nos espaços verdes da cidade do porto. In *Livro de Resumos das XIX Jornadas de Classificação e Análise de Dados (JOCLAD)*, pages 124–127. Associação Portuguesa de Classificação e Análise de Dados (CLAD) e Instituto Politécnico de Tomar.

Joseph, L., Elkin, C., Martin, T., and Possingham, H. (2009). Modeling abundance using n-mixture models: the importance of considering ecological mechanisms. *Ecological Applications*, 19(3):631–642.

Kepler, C. and Scott, J. (1981). Reducing bird count variability by training observers. *Studies in Avian Biology*, 6(366-371).

Kleinbaum, D. G. and Klein, M. (2002). *Logistic Regression: A Self-Learning Text*. Springer-Verlag, New York, second edition.

Kleinschmidt, I., Sharp, B., Clarke, G., Curtis, B., and Fraser, C. (2001). Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in kwazulu natal, south africa. *American Journal of Epidemiology*, 153(12):1213–1221.

Kéry, M. (2008). Estimating abundance from bird counts: Binomial mixture models uncover complex covariate relationships. *The Auk*, 125(2):336–345.

Lam, K., Xue, H., and Cheung, Y. B. (2006). Semiparametric analysis of zero-inflated count data. *Biometrics*, 62(4):996–1003.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *the canadian journal of statistics*, 15(3).

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York, second edition.

Lewsey, J. and Thomson, W. (2004). The utility of the zero-inflated poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal dmf data examining the effect of socio-economic status. *Community Dent Oral Epidemiol*, 32(3):183–189.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Linsdale, J. (1924). A method of showing relative frequency of occurrence of birds. *Condor*, 30:180–184.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing data*. John Wiley & Sons, Inc., New York.

Luther, D., Hilty, J., Weiss, J., Cornwall, C., Wipf, M., and Ballard, G. (2008). Assessing the impact of local habitat variables and landscape context riparian birds in agricultural, urbanized, and native landscapes. *Biodivers Conserv*, 17:1923–1935.

Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A., and Possingham, H. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8(11):1235–1246.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall.

McDowell, A. (2003). From the help desk: hurdle models. *The Stata Journal*, 3(2):178–184.

McNeney, B. and Petkau, J. (1994). Overdispersed poisson regression models for studies of air pollution and human health. *Canadian Journal of Statistics*, 22(4):421–440.

Melles, S. (2005). Urban bird diversity as an indicator of social diversity and economic inequality in vancouver, british columbia. *Urban Habitats*, 3(1):25–48.

Miller, J. M. (2007). *Comparing Poisson, Hurdle, and Zip Model fit under varying degrees of skew and Zero-Inflation*. PhD thesis, University of Florida, Florida.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5:1–19.

Mörtberg, U. (2001). Resident bird species in urban forest remnants; landscape and habitat perspectives. *Landscape Ecology*, 16:193–203.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society*.

Noro, M., Hara, F., and Hagiwara, T. (2010). Analysis of deer ecology and landscape features as factors contributing to deer-vehicle collisions in hokkaido. In *Book of Abstracts of Transportation Research Board 89th Annual Meeting*. Transportation Research Board 89th Annual Meeting.

Nur, N., Jones, S., and Geupel, G. (1999). *A statistical guide to data analysis of avian monitoring programs*. U.S Department of the Interior, Fish and Wildlife Service, BTP-R6001-1999, Washington, D.C.

Oliveira, C. (2007). Métodos de captura e recaptura para a estimação da abundância de uma população: Aplicação da metodologia bootstrap. Master's thesis, Faculdade de Ciêcnias - Universidade de Lisboa.

Palomino, D. and Carrascal, L. (2007). Threshold distances to nearby cities and roads influence the bird community of a mosaic landscape. *Biological conservation*, 140:100–109.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 7(1):120–125.

Pardoe, I. and Durham, C. A. (2003). Model choice applied to consumer preferences. In *Proceedings of the 2003 Joint Statistical Meetings*, Alexandria. American Statistical Association.

Pope, S. E., Fahrig, L., and Merriam, H. G. (2000). Landscape complementation and metapopulation effects on leopard frog populations. *Ecology*, 81(9):2498–2508.

Posa, M. and Sodhi, N. (2006). Effects of an anthropogenic land use on forest birds and butterflies in subic bay, philippines. *Biological conservation*, 129:256–270.

Potts, J. M. and Elith, J. (2006). Comparing species abundance models. *Ecological Modelling*, 199(2):153–163.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, retrieved from `http://www.R-project.org`.

Rabaça, J. (1995). *Método de Censos de Aves: Aspectos Gerais, Pressupostos Principios de Aplicação*. Sociedade Portuguesa para o Estudo das Aves, Lisboa.

Ramsey, F. and Scott, J. (1981). Tests of hearing ability. *Studies in Avian Biology*, 6:341–345.

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

Ridout, M., Hinde, J., and Demétrio, C. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1):219–223.

Rodríguez, G. (2007). Lecture notes on generalized linear models. Retrieved October 11, 2011, from `http://data.princeton.edu/wws509/notes/`.

Rosenstock, S., Anderson, D., Giesen, K., Leukering, T., and Carter, M. (2002). Landbird counting techniques: Current practices and an alternative. *The Auk*, 119(1):46–53.

Royle, J., Nichols, J., and Kéry, M. (2005). Modelling occurrence and abundance of species when detection is imperfect. *OIKOS*, 110:353–359.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63:81–92.

Sandström, U., Angelstam, P., and Mikusinski, G. (2006). Ecological diversity of birds in relation to the structure of urban green space. *Landscape and urban planning*, 77:39–53.

Sauer, J., Peterjohn, B., and Link, W. (1994). Observer differences in the north american breeding bird survey. *The Auk*, 11:50–62.

Sayre, M., Baskett, T., and Sadler, K. (1978). Reappraising factors affecting mourning dove perch cooing. *Journal of Wildlife Management*, 45:428–434.

Seavy, N., Quader, S., Alexander, J., and Ralph, C. J. (2005). Generelized linear models and point count data: Statistical considerations for the design and analysis of monitoring studies. Technical report, USDA Forest Service, PSW-GTR-191.

Slymen, D. J., Ayala, G. X., Arredondo, E. M., and Elder, J. P. (2006). A demonstration of modeling count data with an application to physical activity. *Epidemiologic Perspectives & Innovations*, 3:1–9.

Tadano, Y. S., Ugaya, C. M. L., and Franco, A. T. (2009). Método de regressão de poisson: metodologia para a avaliação do impacto da poluição atmosférica na saúde populacional. *Ambiente & Sociedade*, 12(2):242–255.

Turkman, M. A. and Silva, G. L. (2000). *Modelos Lineares Generalizados da Teoria à Prática*. Edições SPE, Lisboa.

Tvardíková, K. (2010). Bird abundances in primary and secondary growths in papua new guinea: a preliminary assessment. *Tropical Conservation Science*, 3(4):373–388.

United Nations (1992). *Convention on Biological Diversity*. United Nations, Rio de Janeiro.

Vahtera, J., Kivimäki, M., Väänänen, A., Linna, A., Pentti, J., Helenius, H., and Elovainio, M. (2006). Sex differences in health effects of family death or ilness: are women more vulnerable than men? *Psychosom Med*, 68(2):283–291.

Verner, J. (1985). Assessment of counting tecniques. *Current Ornothology*, 2:247–302.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–477.

Welsh, A. H., Cunningham, R. B., Donnely, C. F., and Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling*, pages 297–308.

Wenger, S. and Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, 89(10):2953–2959.

White, G. and Bennetts, R. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology*, 77(8):2549–2557.

White, J. G., Antos, M. J., Fitzsimons, J., and Palmer, G. C. (2005). Non-uniform bird assemblages in urban environments: the influence of streetscape vegetation. *Landscape and urban planning*, 71:123–135.

White, K. A. (1942). Frequency of occurence of summer birds at the university of michigan biological station. *Wilson Bulletin*, 54:204–210.

Wilson, D. and Bart, J. (1985). Reliability of singin bird surveys: Effects of song phenology during breeding season. *Condor*, 87:69–73.

Yau, K. and Lee, A. (2001). Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, 20:2907–2920.

Zeger, S., Liang, K.-Y., and Albert, P. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27(8).

Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer Science+Business Media, New York, first edition.