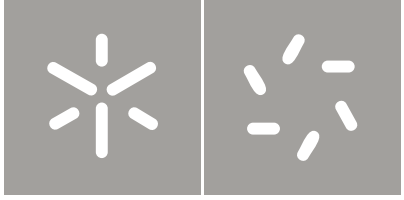
Universidade do Minho
Escola de Ciências

Jorge Helder Pereira dos Santos

Modelos Para Dados de Contagem com Excesso de Zeros

Outubro de 2013

Modelos Para Dados de Contagem com Excesso de Zeros
Jorge Helder Pereira dos Santos
UMinho | 2013



Universidade do Minho
Escola de Ciências

Jorge Helder Pereira dos Santos

Modelos Para Dados de Contagem com Excesso de Zeros

Tese de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora
Susana Margarida Ferreira de Sá Faria

Outubro de 2013

DECLARAÇÃO

Nome: Jorge Helder Pereira dos Santos

Correio electrónico: jorge.mfd@sapo.pt

Tlm.: 914382523

Número do Bilhete de Identidade:10050117

Título da dissertação:

Modelos Para Dados de Contagem com Excesso de Zeros

Ano de conclusão:2013

Orientadora:

Profª Doutora Susana Margarida Ferreira de Sá Faria

Designação do Mestrado:

Mestrado em Estatística

Escola: Escola de Ciências

Departamento: Departamento de Matemática e Aplicações

Guimarães, ___/___/_____

Assinatura: _____

≪Não é na ciência que está a felicidade, mas na aquisição da ciência.≫
(Edgar Allan Poe)

Agradecimentos

Gostaria de agradecer a algumas pessoas que de diferentes formas ajudaram a tornar possível este trabalho. Em particular gostaria de agradecer:

À Professora Doutora Susana Faria pela orientação, disponibilidade e partilha de conhecimentos, crucial na elaboração deste trabalho;

A todos os professores do Departamento de Matemática e Aplicações da Escola de Ciências;

À minha família e amigos pelo apoio, durante todo o meu percurso académico;

Resumo

Os modelos de regressão para dados de contagem são muito utilizados nas mais variadas áreas de estudo para a modelação de fenómenos. Estes modelos integram um quadro especial de metodologias devido ao facto de a variável resposta tomar apenas valores inteiros não negativos. A distribuição de Poisson é a mais conhecida, e a mais utilizada para modelar dados de contagem, no entanto sempre que existe sobredispersão, torna-se necessário recorrer a outras distribuições, nomeadamente à distribuição Binomial Negativa. Outro problema comum nos dados de contagem é o excesso de zeros na variável resposta. Os modelos de regressão de zeros inflacionados são amplamente usados para modelar esse tipo de dados. Estes modelos modelam as contagens como uma mistura de duas distribuições com dois processos subjacentes, um que trata do excesso de zeros modelado por uma massa pontual, e um outro que trata das contagens sendo modelado por uma distribuição de Poisson ou Binomial Negativa.

Neste trabalho pretendeu-se estudar os modelos de regressão para dados de contagem e a sua aplicação a dados bancários relativos a clientes a quem foi garantido crédito de consumo por um banco. Tem como principal objetivo estudar a relação do número de não pagamento da prestação do empréstimo de um cliente em função das caracteristicas do cliente e do contrato. Em particular, foram ajustados os modelos de regressão de Poisson, modelos de regressão Binomial Negativa, modelos de regressão de Poisson de zeros inflacionados e modelos de regressão binomial negativa de zeros inflacionados utilizando o algoritmo EM para obter as estimativas de máxima verosimilhança dos parâmetros.

Os resultados obtidos mostraram que os modelos de regressão de zeros inflacionados apresentam um melhor ajustamento, quando comparados com os modelos que não têm em consideração o excesso de zeros. Mostraram ainda que os modelos baseados na distribuição Binomial Negativa, são os mais adequados para modelar estes dados, em vez dos modelos baseados na distribuição de Poisson.

Palavras-chave: Modelo de regressão de Poisson; Modelo de regressão Binomial Negativa; Modelo de regressão de zeros inflacionados

Abstract

Regression models for count data are highly used in several areas of study for modelation of phenomena. These models feature a special methodological board that comes from the fact that the response variable just takes non-negative integer values. The Poisson distribution is the most recognized and most widely used to model count data, however when there is overdispersion, it becomes necessary the use other distributions, as so, including negative binomial distribution. Another common problem in count data, is the excess of zeros in the response variable. Zero inflated regression models are widely used to model this type of data. These models model the counts as a mixture of two distributions with two underlying processes, one that deals with excess of zeros modeled by a pontual mass, and another one that handles the counts by being modelated by a Poisson or Negative Binomial distributions.

In this work we intended to study regression models for count data and its application on bank data clients to whom it was granted consumption credit by a bank. Its main objective is to study the relationship of the number of non payment of the installment of a client depending on the characteristics of client and the contract. In particular, we fit the Poisson regression models, negative binomial regression models, zero inflated Poisson regression models and negative binomial regression models for zero inflated using the EM algorithm to obtain maximum likelihood estimates of the parameters.

The results showed that zero inflated regression models have a better fit compared with models that do not take into account the extra zeros. Also showed that models based on the negative binomial distribution, are more suitable for modeling this data instead of models based on Poisson distribution.

Palavras-chave: Poisson Regression; Negative Binomial Regression; Zero Inflated Model

Conteúdo

ii

1 Introdução **1**

1.1 Estrutura do Trabalho 2

2 Modelos Lineares Generalizados **5**

2.1 Família Exponencial 5

2.2 Componentes do Modelo Linear Generalizado 8

2.2.1 Componente aleatória 9

2.2.2 Componente estrutural ou sistemática 9

2.2.3 Função de ligação 9

2.3 Inferência 10

2.3.1 Função de Log-Verosimilhança 10

2.3.2 Estimação dos Parâmetros 12

2.3.2.1 Método iterativo de mínimos quadrados ponderados 13

2.3.3 Testes de Hipóteses 14

2.3.3.1 Teste de Wald 15

2.3.3.2 Teste de razão de verosimilhanças 15

2.4 Seleção e Validação de Modelos 16

2.4.1 Qualidade do Ajustamento 16

2.4.1.1 *Deviance* 16

2.4.1.2 Percentagem de *Deviance* Explicada 17

2.4.1.3 Estatística de Pearson generalizada 18

2.4.2 Análise de Resíduos 18

2.4.2.1 Resíduos de Pearson 19

2.4.2.2 Resíduo *Deviance* 19

2.4.3 Seleção de modelos 19

2.4.4 Sobredispersão 20

2.4.4.1 Quasi-verosimilhança 21

3 Modelos de Regressão para Dados de Contagem **23**
3.1 Modelo de Regressão de Poisson 25
3.2 Modelo de Regressão Binomial Negativa 27
3.3 Modelo de Regressão para Dados Inflacionados 29
3.3.1 Modelo de Regressão de Poisson de zeros inflacionados (ZIP) 30
3.3.2 Modelo de Regressão Binomial Negativo de zeros inflacionados (ZINB) . 32
3.3.3 Resíduos 34
3.3.4 Teste Vuong 34

4 Aplicação a dados reais **37**
4.1 Base de Dados 37
4.2 Análise descritiva 39
4.2.1 Variáveis Explicativas Quantitativas 39
4.2.2 Variáveis Explicativas Qualitativas 43
4.2.3 Associação entre a variável dependente e as variáveis explicativas 45
4.2.4 Teste Kruskal-Wallis 46
4.3 Seleção de Modelos 47

5 Conclusões e trabalho futuro **65**

6 Anexos **67**

Lista de Figuras

3.1 Função de probabilidade de uma variável de Poisson com zeros inflacionados para diferentes valores para π. 24
3.2 Função de probabilidade de uma variável Binomial Negativa com zeros inflacionados para diferentes valores para π. 25

4.1 Gráfico de Barras da variável *MesessemPagar* 39
4.2 Caixa com bigodes e gráfico de barras para a variável *Idade* 40
4.3 Caixa com bigodes e gráfico de barras para a variável *IdadeContrato* 40
4.4 Caixa com bigodes e histograma para a variável *MontanteContratado* 41
4.5 Caixa com bigodes e histograma para a variável *CapitalVincendo* 41
4.6 Caixa com bigodes e gráfico de barras para a variável *NMesesLC* 42
4.7 Caixa com bigodes e histograma para a variável *PrestacaoMensal* 42
4.8 Caixa com bigodes e gráfico de barras para a variável *NAnosCliente* 43
4.9 *Envelope plot* dos resíduos de Pearson do modelo Poisson 50
4.10 Gráficos dos resíduos do modelo de regressão de Poisson 51
4.11 *Envelope plot* dos resíduos de Pearson do modelo Binomial Negativa 53
4.12 Gráficos dos resíduos do modelo de regressão Binomial Negativa. 54
4.13 *Envelope plot* dos resíduos de Pearson do modelo ZIP 57
4.14 Gráfico dos resíduos de Pearson do modelo ZIP 57
4.15 *Envelope plot* dos resíduos de Pearson do modelo ZINB 60
4.16 Gráfico dos resíduos de Pearson do modelo ZINB 60

Lista de Tabelas

4.1 Informação sobre os dados. 38
4.2 Tabela de frequências do número de meses sem pagamento. 38
4.3 Tabela das medidas de tendência central e dispersão das variáveis quantitativas . 42
4.4 Tabela das medidas de assimetria e achatamento das variáveis quantitativas . . . 43
4.5 Tabela de frequências da variável *Sexo*. 43
4.6 Tabela de frequências da variável *Ordenado*. 44
4.7 Tabela de frequências da variável *EstadoCivil*. 44
4.8 Tabela de frequências da variável *Habilitacoes*. 44
4.9 Tabela de frequências da variável *Regiao*. 45
4.10 Tabela de frequências da variável *Profissao*. 45
4.11 Tabela de frequências da variável *SldMdSem.cat*. 45
4.12 Coeficiente de Correlação de *Spearman* das covariáveis quantitativas e a variável *MesessemPagar*. 46
4.13 Coeficiente de Correlação de *Spearman* das covariáveis quantitativas. 47
4.14 Teste Kruskal-Wallis para as covariáveis qualitativas e a variável *MesessemPagar*. 47
4.15 Estatísticas de ajustamento dos modelos de regressão de Poisson 49
4.16 Modelo de regressão de Poisson 49
4.17 Valores estimados pelo modelo de regressão de Poisson. 50
4.18 Estatísticas de ajustamento dos modelos de regressão Binomial Negativos 52
4.19 Modelo de regressão Binomial Negativa 52
4.20 Valores estimados pelo modelo de regressão Binomial Negativa. 53
4.21 Teste de razão de verosimilhanças entre o modelo de regressão de Poisson e o modelo de regressão Binomial Negativa. 54
4.22 Estatísticas de ajustamento dos modelos ZIP 55
4.23 Modelo de regressão de Poisson de zeros inflacionados 56
4.24 Valores estimados pelo modelo ZIP. 56
4.25 Teste Vuong entre modelo de regressão de Poisson e o modelo ZIP. 58
4.26 Estatísticas de ajustamento dos modelos ZINB 58
4.27 Modelo de regressão Binomial Negativa de zeros inflacionados 59
4.28 Valores estimados pelo modelo ZINB. 59

4.29 Teste Vuong entre modelo Binomial Negativa e o modelo ZINB. 61
4.30 Estatísticas de ajustamento dos modelos escolhidos 61
4.31 Teste Vuong entre os modelos de regressão. 61

6.1 Modelos de Regressão de Poisson com apenas uma variável explicativa. 68

Capítulo 1

Introdução

A Análise de Regressão é hoje uma das técnicas estatísticas mais usadas em todas as áreas da Ciência. Na Análise de Regressão pretende-se encontrar uma relação estocástica entre duas ou mais variáveis com o objetivo de explicar determinado fenómeno em estudo e nomeadamente prever a evolução desse fenómeno.

As bases para o aparecimento dos modelos de regressão encontram-se nos estudos realizados por Legendre e Gauss, no início do século XIX sobre o método dos mínimos quadrados, aplicados a observações astronómicas. O termo ”regressão” foi introduzido só mais tarde por Galton em 1885 e teve origem na observação de que filhos de pais mais altos do que a média tendiam a ser mais baixos do que os pais, e filhos de pais mais baixos do que a média tendiam a ser mais altos do que os pais, havendo assim uma tendência geral para ”regressar” aos valores médios da população.

Os modelos de regressão, podem ser usados para modelar a relação funcional entre duas ou mais variáveis. Mais precisamente, analisar a influência que uma ou mais variáveis (designadas por variáveis independentes ou explicativas) têm sobre uma variável de interesse (designada por variável resposta ou dependente). Estes modelos permitem ainda predizer o valor de uma variável dependente a partir de um conjunto de variáveis independentes.

Para dar resposta às situações em que a variável resposta não segue uma distribuição normal, os modelos lineares generalizados, que são uma extensão dos modelos de regressão linear, permitem incluir outras distribuições da variável dependente, desde que pertencentes à família exponencial.

Os dados de contagem são um tipo de dados muito frequentes nas mais diversas áreas de estudo. A natureza deste tipo de dados que assume apenas valores inteiros não negativos, correspondentes à ocorrência de um dado número de eventos durante um intervalo de tempo ou espaço, levaram ao aparecimento dos modelos de regressão para dados de

contagem.

O modelo de regressão de Poisson, que é construído com base na distribuição de Poisson é o modelo mais utilizado para este tipo de dados, mas um problema comum neste modelo surge quando a variância da resposta é superior ao seu valor médio, fenómeno designado por sobredispersão. Quando se verifica sobredispersão torna-se necessário recorrer-se ao modelo de regressão binomial negativa.

Nos dados de contagem é também muito comum existir excesso de zeros, o que pode levar a problemas de ajustamento no modelo de regressão de Poisson. O modelo de regressão binomial negativa, permite resolver o problema da sobredispersão, mas não resolve o problema do excesso de zeros. Para resolver este problema, surgem os modelos de regressão de zeros inflacionados e os modelos de regressão com barreira (que não serão aplicados neste trabalho), que foram desenvolvidos para ter em conta o excesso de zeros nos dados.

A gestão do risco de crédito é uma actividade fundamental e inerente à intermediação financeira. Para minimizar os riscos, as instituições financeiras analizam o crédito a ser concedido ao tomador de forma antecedente à operação (celebração do contrato). Ao conceder o crédito, uma preocupação importante está associada à possibilidade de que o cliente não cumpra com os compromissos assumidos.

Neste trabalho, será estudado uma amostra sobre dados bancários relativos a clientes a quem foi garantido crédito de consumo por um banco tendo como principal objetivo estudar a relação do número de não pagamentos da prestação do empréstimo de um cliente em função das caracteristicas do cliente e do contrato, aplicando os modelos de regressão para dados de contagem.

1.1 Estrutura do Trabalho

Este trabalho encontra-se dividido em cinco capítulos.

No Capítulo 2 é apresentada toda a base teórica relativa aos Modelos Lineares Generalizados, que fundamenta as metodologias utilizadas neste trabalho.

O Capítulo 3 tem por objetivo apresentar em detalhe os modelos de regressão para dados de contagem, nomeadamente o modelo de regressão de Poisson e o modelo de regressão Binomial Negativa. Este capítulo introduz ainda os modelos de regressão para dados de contagem com excesso de zeros, o modelo de regressão de Poisson de zeros inflacionados e o modelo de regressão binomial negativa de zeros inflacionados.

No Capítulo 4 apresentam-se os resultados da aplicação destes modelos a dados reais. Inicia-se o estudo com uma análise descritiva dos dados, efetuando-se seguidamente a

aplicação dos modelos de regressão para dados de contagem para selecionar o modelo que melhor se ajusta aos dados.

No Capítulo 5 são apresentadas as conclusões e algumas sugestões para trabalho futuro.

Capítulo 2

Modelos Lineares Generalizados

Um modelo estatístico é uma representação simplificada de alguns aspectos do mundo real, tomando a forma de uma equação, para descrever a relação entre várias variáveis. O modelo linear geral pretende modelar o efeito que uma ou mais variáveis (variáveis explicativas), medidas em indivíduos ou objectos, têm sobre uma variável de interesse (variável resposta) e surgiu no início do século XIX por Legendre e Gauss. Este modelo exprime a média da variável resposta como combinação linear das variáveis explicativas e é aplicado quando a variável resposta segue uma distribuição normal.

Apesar dos modelos lineares serem bastante úteis, só podem ser aplicados num determinado conjunto de situações. Assim, os modelos lineares generalizados surgem em 1972, por Nelder and Wedderburn [Nelder and Wedderburn, 1972] para dar resposta a situações em que a resposta é não normal. Por exemplo, quando os valores possiveis para a variável resposta toma valores binários, ou quando são valores provenientes de contagens, ou ainda, quando a variância da variável resposta depende da média, faz mais sentido usar-se este tipo de modelos.

Estes modelos apresentam uma estrutura linear nas variáveis explicativas, e assumem que a distribuição condicional da variável resposta pertence à família exponencial, podendo eventualmente relacionar, de forma não linear a média da resposta com a estrutura linear das variáveis explicativas.

2.1 Família Exponencial

A variável aleatória Y tem distribuição pertencente à família exponencial de distribuições se a sua função de densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) é da forma

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

onde θ é o parâmetro de localização, ϕ é o parâmetro de dispersão ou parâmetro de escala e $a(.)$, $b(.)$ e $c(.)$ são funções reais conhecidas. A função $a(.)$ depende apenas do parâmetro de dispersão e é geralmente da forma $a(\phi) = \frac{\phi}{w}$, onde w é uma constante conhecida, a função $b(.)$ depende apenas do parâmetro θ e a função $c(.)$ depende apenas da variável aleatória Y e do parâmetro de dispersão ϕ.

Pode ser demostrado [McCullagh and Nelder, 1989] que se a variável aleatória Y com uma distribuição pertencente à família exponencial, então

$$\mathbb{E}(Y) = \mu = b'(\theta) \tag{2.2}$$

$$\text{Var}\,(Y) = \sigma^2 = a(\phi)b''(\theta) \tag{2.3}$$

onde $b'(\theta)$ e $b''(\theta)$ são a primeira e a segunda derivadas de $b(\theta)$, respectivamente. Assim, a variância de Y é o produto de duas funções, $b''(\theta)$ que depende apenas do parâmetro canónico θ que se designa por *função de variância* e que se representa por , $V(\mu)$, e outra, $a(\phi)$ que depende apenas do parâmetro de dispersão ϕ [Turkman and Silva, 2000].

São exemplos de distribuições que pertencem à família exponencial, as seguintes distribuições:

- 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

$$\begin{aligned} f(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\, exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= exp\left[\frac{\mu y - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right)\right] \end{aligned} \tag{2.4}$$

para $y \in \mathbb{R}$. Tem-se então que esta função é do tipo (2.1), com $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2}$, $a(\phi) = \sigma^2$, e $c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right)$

A média e a variância de Y são,

$$\mathbb{E}(Y) = b'(\theta) = \theta = \mu$$
$$\text{Var}(Y) = a(\phi)b''(\theta) = \sigma^2$$

A função de variância é $V(\mu) = 1$.

- Distribuição Binomial

Se $Y \sim B(n,\pi)$, onde n é o número de experiências de Bernoulli de um determinado acontecimento e π é a probabilidade de sucesso desse acontecimento em cada experiência. A função de probabilidade é dada por

$$\begin{aligned} f(y|n,\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= exp\left[y\,log\left(\frac{\pi}{1-\pi}\right) + n\,log(1-\pi) + log\binom{n}{y}\right] \end{aligned} \tag{2.5}$$

Obtém-se, $\theta = log\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = n\,log(1+exp(\theta))$, $a(\phi) = 1$ e $c(y,\phi) = log\binom{n}{y}$.

A média e a variância de Y são representadas por,

$$\mathbb{E}(Y) = b'(\theta) = n\pi$$
$$\text{Var}(Y) = a(\phi)b''(\theta) = n\pi(1-\pi)$$

A função de variância é $V(\mu) = n\pi(1-\pi)$

- Distribuição Poisson

Considerando que Y segue uma distribuição de Poisson, com parâmetro μ, $P(\mu)$, a função de probabilidade de Y é dada por

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

Neste caso, $\theta = log(\mu)$, $b(\theta) = exp(\theta)$, $a(\phi) = 1$ e $c(y,\phi) = -log(y!)$.

A média e a variância de Y são, respectivamente

$$\mathbb{E}(Y) = b'(\theta) = exp(\theta) = \mu$$
$$\text{Var}(Y) = a(\phi)b''(\theta) = exp(\theta) = \mu$$

A função de variância $V(\mu) = \mu$

- Distribuição 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 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 densidade de probabilidade de Y é dada por

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

Neste caso, a distribuição Binomial Negativa está escrita na forma canónica, onde $\theta = log(1-p)$, $b(\theta) = -k\, log(p)$, $a(\phi) = 1$, e $c(y,\phi) = log\binom{y+k-1}{k-1}$.

A média e a variância são expressas, respectivamente, por

$$\mathbb{E}(Y) = b'(\theta) = \frac{k(1-p)}{p}$$
$$\text{Var}(Y) = a(\phi)b''(\theta) = \frac{k(1-p)}{p^2}$$

A função de variância é $V(\mu) = \frac{k(1-p)}{p^2}$

2.2 Componentes do Modelo Linear Generalizado

Os modelos lineares generalizados são caracterizados pelos seguintes componentes:

- Componente aleatória - que identifica a variável aleatória resposta Y e especifica uma distribuição para Y pertencente à família exponencial;
- Componente estrutural ou sistemática - especifica as variáveis explicativas, ou covariáveis do modelo, e considera uma combinação linear dessas variáveis;
- Função de ligação - estabelece a ligação entre as componentes aleatória e estrutural [Agresti, 2007], [McCullagh and Nelder, 1989].

2.2.1 Componente aleatória

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

$$\mathbb{E}(Y_i|\mathsf{x}_i) = \mu_i = b'(\theta_i), \quad i = 1, ..., n.$$

2.2.2 Componente estrutural ou sistemática

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

$$\eta = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.8}$$

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

$$\eta_i = z_i^T \beta, \quad i = 1, \ldots, n,$$

onde β é o vector dos coeficientes de regressão e

$$z_i = (1, \mathsf{x}_i^T)^T$$

Em notação matricial,

$$\eta = Z\beta,$$

onde Z é a matriz de especificação de dimensão $n \times (p+1)$, ou seja, é a matriz cuja primeira coluna é formada apenas por 1‘s e cujas restantes colunas são constituídas pelos vetores coluna 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 é uma função g, monótona e diferenciável, que relaciona o preditor linear η, com o valor esperado da variável resposta,

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

A função de ligação $g(.)$ estabelece a ligação entre a componente aleatória e a componente estrutural do modelo. Quando a função de ligação torna o preditor linear η igual

ao parâmetro canónico θ, diz-se que a função de ligação é a função de ligação canónica.

2.3 Inferência

Nos modelos lineares generalizados, após a formulação do modelo que se pensa apropriado há necessidade de proceder à realização de inferências sobre esse modelo. Os métodos de inferência são fundamentalmente baseados na função de verosimilhança.

2.3.1 Função de Log-Verosimilhança

Assumindo que a variável aleatória Y tem distribuição da família exponencial, a função de verosimilhança do modelo, em função de β é dada por [Turkman and Silva, 2000],

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f(y_i|\theta_i,\phi) \\ &= \prod_{i=1}^{n} exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \\ &= exp\left\{\sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)\right\} \end{aligned} \tag{2.9}$$

O logaritmo da função de verosimilhança (que passaremos a designar por log-verosimilhança) é dado por

$$\begin{aligned} ln(L(\beta)) &= \ell(\beta) \\ &= \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \\ &= \sum_{i=1}^{n}\ell_i(\beta) \end{aligned} \tag{2.10}$$

onde ℓ_i é a contribuição de cada observação y_i para a verosimilhança [Turkman and Silva, 2000].

Os estimadores de máxima verosimilhança para β são obtidos como solução do sistema de equações de verosimilhança:

$$\frac{\partial\ell(\beta)}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial\ell_i(\beta)}{\partial\beta_j} = 0, \quad j = 0,1,\ldots,p. \tag{2.11}$$

Para obtermos estas equações, escrevemos [McCullagh and Nelder, 1989],

$$\frac{\partial \ell_i(\beta)}{\partial \beta_j} = \frac{\partial \ell_i(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i(\beta)}{\partial \beta_j}, \quad j = 0, 1, \ldots, p. \tag{2.12}$$

Tendo em conta a função log-verosimilhança, e sabendo que $b'(\theta_i) = \mu_i$ vem

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

e como $\text{Var}(Y_i) = a(\phi)b''(\theta_i)$ então

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

e como $\eta_i = z_i^T \beta$ vem

$$\frac{\partial \eta_i(\beta)}{\partial \beta_j} = z_{ij}$$

assim

$$\frac{\partial \ell_i(\beta)}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)}\frac{a(\phi)}{\text{Var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}z_{ij} \tag{2.13}$$

e as equações de verosimilhança para β são dadas por

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\text{Var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}z_{ij} = 0 \quad j = 1, \ldots, p. \tag{2.14}$$

A primeira derivada da função log-verosimilhança em ordem a β é designada por função *score* e é obtida por

$$s(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta) \tag{2.15}$$

onde $s_i(\beta)$ é o vetor de componentes $\frac{\partial \ell_i(\beta)}{\partial \beta_j}$ obtidas na equação (2.13) .

A matriz de covariância da função *score*,

$$I(\beta) = \mathbb{E}\left[-\frac{\partial s(\beta)}{\partial \beta}\right] \tag{2.16}$$

é conhecida como *matriz de informação de Fisher*, e para ser obtida é necessário considerar-se o valor esperado das segundas derivadas da função log-verosimilhança em ordem a β,

$$
\begin{aligned}
-\mathrm{IE}\left(\frac{\partial^2 \ell_i}{\partial\beta_j\,\partial\beta_k}\right) &= \mathrm{IE}\left(\frac{\partial \ell_i}{\partial\beta_j}\frac{\partial \ell_i}{\partial\beta_k}\right) \\
&= \mathrm{IE}\left[\left(\frac{(Y_i-\mu_i)z_{ij}}{\mathrm{Var}\,(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right)\left(\frac{(Y_i-\mu_i)z_{ik}}{\mathrm{Var}\,(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right)\right] \\
&= \mathrm{IE}\left[\frac{(Y_i-\mu_i)^2 z_{ij}\,z_{ik}}{(\mathrm{Var}\,(Y_i))^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right] \\
&= \frac{z_{ij}\,z_{ik}}{\mathrm{Var}\,(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2
\end{aligned}
\tag{2.17}
$$

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

$$
I(\beta)_{j,k} = -\sum_{i=1}^{n}\mathrm{IE}\left(\frac{\partial^2\ell_i}{\partial\beta_j\,\partial\beta_k}\right) = \sum_{i=1}^{n}\frac{z_{ij}\,z_{ik}}{\mathrm{Var}\,(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{2.18}
$$

Na forma matricial temos

$$
I(\beta) = Z^T W Z \tag{2.19}
$$

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

$$
\begin{aligned}
\varpi_i = \frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{\mathrm{Var}\,(Y_i)} &= \frac{w_i\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{\phi b''(\theta)} \\
&= \frac{w_i}{\phi b''(\theta)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2} \\
&= \frac{w_i}{\phi V(\mu_i)}.
\end{aligned}
\tag{2.20}
$$

2.3.2 Estimação dos Parâmetros

Os estimadores de máxima verosimilhança de β são obtidos como soluções das equações de verosimilhança, no entanto, estas soluções podem não corresponder necessariamente a um máximo global da função log-verosimilhança, $\ell(\beta)$. Contudo, em muitos modelos a função $\ell(\beta)$ é côncava, e por isso, o único máximo local coincide com o máximo global, tornando único o estimador de máxima verosimilhança.

Partindo do princípio que existe solução e que ela é única, persiste ainda um problema com o cálculo das estimativas de máxima verosimilhança. As equações de verosimilhança geralmente não têm solução analítica, uma vez que são equações não lineares, o que implica que sejam usados para a sua resolução métodos numéricos. O método número mais utilizado para resolver estas equações é o método iterativo dos mínimos quadrados

ponderados.

2.3.2.1 Método iterativo de mínimos quadrados ponderados

O método iterativo de mínimos quadrados ponderados é baseado no método de *scores* de Fisher.

Seja $\widehat{\beta}^{(0)}$ uma estimativa inicial para β. O método de *scores* de Fisher define sucessivas iterações através da relação

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + \left[I(\widehat{\beta}^{(k)})\right]^{-1} s(\widehat{\beta}^{(k)}), \quad k = 0,1,2,\ldots \tag{2.21}$$

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

Escrevendo a expressão (2.21) na seguinte forma

$$\left[I(\widehat{\beta}^{(k)})\right]\widehat{\beta}^{(k+1)} = \left[I(\widehat{\beta}^{(k)})\right]\widehat{\beta}^{(k)} + s(\widehat{\beta}^{(k)}), \quad k = 0,1,\ldots \tag{2.22}$$

e como o lado direito da equação (2.22) é um vector com elemento genérico de ordem l dado por

$$\sum_{j=1}^{p}\left[\sum_{i=1}^{n}\frac{z_{ij}z_{il}}{\mathrm{Var}\,(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right]\widehat{\beta}_j^{(k)} + \sum_{i=1}^{n}\frac{(y_i-\mu_i)z_{il}}{\mathrm{Var}\,(Y_i)}\frac{\partial\mu_i}{\partial\eta_i} \tag{2.23}$$

obtém-se na forma matricial

$$I(\widehat{\beta}^{(k)})\widehat{\beta}^{(k+1)} = Z^T W^{(k)} T^{(k)}, \tag{2.24}$$

onde Z é a matriz do modelo em questão, W é a matriz diagonal dos pesos com entradas dadas por

$$\varpi_i^{(k)} = \frac{w_i^{(k)}}{\phi b''(\theta_i)\left(\frac{\partial\eta_i^{(k)}}{\partial\mu_i^{(k)}}\right)^2}. \tag{2.25}$$

e T é o vetor com as entradas

$$\begin{aligned} t_i^{(k)} &= \widehat{\eta}_i^{(k)} + (y_i - \widehat{\mu}_i^{(k)})\frac{\partial\widehat{\eta}_i^{(k)}}{\partial\mu_i^{(k)}} \\ &= g(\widehat{\mu}_i^{(k)}) + (y_i - \widehat{\mu}_i^{(k)})g'(\widehat{\mu}_i^{(k)}). \end{aligned} \tag{2.26}$$

sendo a expressão final para a estimativa de β na $(k+1)$-ésima iteracção dada por [Turkman and Silva, 2000],

$$\widehat{\beta}^{(k+1)} = (Z^T W^{(k)} Z)^{-1} Z^T W^{(k)} T^{(k)}, \tag{2.27}$$

Este método é conhecido como o "algoritmo iterativo de mínimos quadrados ponderados"uma vez que a equação (2.27) é idêntica à que se obteria para os estimadores de mínimos quadrados ponderados se se aplicasse em cada passo, a regressão linear de respostas $T^{(k)}$ em Z onde $W^{(k)}$ é uma matriz de pesos.

O procedimento repete-se até que as estimativas se alterem a menos de uma constante pré-especificada.

Um critério de paragem, por exemplo, é dado por [Fahrmeir and Tutz, 2001],

$$\frac{\| \widehat{\beta}^{(k+1)} - \widehat{\beta}^{(k)} \|}{\| \widehat{\beta}^{(k)} \|} \leq \varepsilon, \tag{2.28}$$

para um valor de $\varepsilon > 0$ previamente definido.

Resumidamente, o cálculo das estimativas de máxima verosimilhança de β processa-se, iterativamente, da seguinte forma:

1. Dado $\widehat{\beta}^{(k)}$, $k = 0, 1, 2, ...$, calcula-se $t_i^{(k)}$ usando a expressão (2.26) e $W^{(k)}$ através da expressão (2.25).

2. A nova iteração $\widehat{\beta}^{(k+1)}$ é calculada através da expressão (2.27).

O parâmetro de dispersão, ϕ, pode ser estimado por [McCullagh and Nelder, 1989]

$$\begin{aligned} \widehat{\phi} &= \frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{w_i (y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \\ &= \frac{X^2}{n-(p+1)}. \end{aligned} \tag{2.29}$$

sendo $\widehat{\phi}$ um estimador consistente de ϕ e X^2 é a estatística de Pearson generalizada, que será definida na Secção (2.4) .

2.3.3 Testes de Hipóteses

Os testes de hipóteses são uma ferramenta útil para fazer inferências sobre uma população com base numa amostra. Neste caso, pretende-se avaliar se cada um dos

parâmetros é significativamente diferente de zero, e comparar a qualidade do ajustamento de vários modelos.

2.3.3.1 Teste de Wald

Se pretendermos testar a hipótese

$$H_0 : \beta_j = 0, \quad versus \quad H_1 : \beta_j \neq 0, \quad j = 0, ..., p$$

que indica que o coeficiente independente $\beta_0 = 0$ é irrelevante para o modelo $(j = 0)$ ou que se a variável explicativa X_j não é estatísticamente significativa para o modelo de regressão $(j \neq 0)$.

A estatística de teste, para grandes amostras, é dada por,

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

O teste de Wald encontra-se implementado no package *lmtest* do *software* R através da função *waldtest( )*.

2.3.3.2 Teste de razão de verosimilhanças

O teste de razão de verosimilhanças é utilizado para comparar a qualidade do ajustamento de dois modelos encaixados, isto é, modelos em que um é submodelo do outro.

Como a função de verosimilhança, $L(\beta)$ é inferior a 1, e geralmente muito pequena (uma vez que é o produto de várias probabilidades do intervalo $[0;1]$), é usual usar o $ln(L(\beta))$, que é um número negativo, pelo que se multiplica por -2 para torná-lo positivo, maior e com distribuição conhecida, a distribuição Qui-quadrado [Marôco, 2010].

Considere-se dois modelos encaixados, M_1 e M_2, com um número de parâmetros p_1 e p_2 respectivamente, tal que $p_1 < p_2$.

Para comparar a qualidade do ajustamento de dois modelos aplica-se o teste de razão de verosimilhanças, sob a hipótese nula,

H_0: Os dois modelos têm a mesma qualidade de ajustamento.

A estatística de teste é dada por

$$\begin{aligned} G &= -2\,\ell_{M_1}(\beta) - (-2\ell_{M_2}(\beta)) \\ &= -2\left(\frac{\ell_{M_1}(\beta)}{\ell_{M_2}(\beta)}\right) \end{aligned} \tag{2.30}$$

em que $\ell_{M_1}(\beta)$ é a função log-verosimilhança do modelo M_1 e $\ell_{M_2}(\beta)$ a função log-verosimilhança do modelo M_2. Repare-se que a estatística de teste se obtém a partir da razão de verosimilhanças dos dois modelos, daí a designação de "Teste de razão de Verosimilhanças".

A estatística de teste segue uma distribuição Qui-quadrado com $(p_2 - p_1)$ graus de liberdade.

$$G \sim \chi^2_{p_2-p_1} \tag{2.31}$$

O teste de razão de verosimilhanças encontra-se implementado no package *lmtest* do *software* R através da função *lrtest( )*.

2.4 Seleção e Validação de Modelos

2.4.1 Qualidade do Ajustamento

Nesta Secção serão apresentadas duas medidas para avaliar a qualidade do ajustamento de um determinado modelo em estudo. Essas medidas são a *deviance* e Estatística de Pearson generalizada.

2.4.1.1 *Deviance*

Comecemos por definir o modelo nulo que é o modelo mais simples onde somente o parâmetro constante é estimado, apresentando por isso um menor valor da função de verosimilhança. Por sua vez o modelo completo ou saturado é o modelo que estima um parâmetro para cada observação, isto é, as estimativas de máxima verosimilhança são as próprias observações, $\widehat{\mu}_i = y_i$.

O modelo saturado ou completo é útil para avaliar a qualidade do ajustamento de um determinado modelo ajustado aos dados, através da introdução de medida de distância dos valores ajustados $\widehat{\mu}$ com esse modelo e dos correspondentes valores observados y.

Se assumirmos que não existem pesos associados às observações, a razão de verosimilhanças entre o modelo ajustado (M_A) e o modelo completo (M_C) é dada por

$$\begin{aligned} -2(\ell_A(\beta)-\ell_C(\beta)) &= -2\left[\sum_{i=1}^{n}\frac{y_i\widehat{\theta}_i-b(\widehat{\theta}_i)}{\phi}+c(y_i,\phi)-\left(\sum_{i=1}^{n}\frac{y_i\tilde{\theta}_i-b(\tilde{\theta}_i)}{\phi}+c(y_i,\phi)\right)\right] \\ &= 2\sum_{i=1}^{n}\frac{y_i(\tilde{\theta}_i-\widehat{\theta}_i)-b(\tilde{\theta}_i)+b(\widehat{\theta}_i)}{\phi} \\ &= \frac{D(y,\widehat{\mu})}{\phi} \end{aligned}$$

onde, ℓ_A corresponde ao logaritmo da função de verosimilhança do modelo ajustado (M_A), ℓ_C ao logaritmo da função de verosimilhança do modelo completo (M_C), y_i é o valor ajustado da observação *i* dada pelo modelo (M_C) , $\widehat{\theta}_i$ os parâmetros estimados pelo modelo (M_A), $\tilde{\theta}_i$ os parâmetros do modelo (M_C).

A quantidade $D(y,\widehat{\mu})$, também conhecida por *deviance* é dada por

$$\begin{aligned} D(y,\widehat{\mu}) &= -2\phi(\ell_A(\beta)-\ell_C(\beta)) \\ &= 2\sum_{i=1}^{n} y_i(\tilde{\theta}_i-\widehat{\theta}_i)-b(\tilde{\theta}_i)+b(\widehat{\theta}_i). \end{aligned} \tag{2.32}$$

A *deviance* de um modelo avalia, portanto, a discrepância entre os valores ajustados pelo modelo completo e os valores ajustados pelo modelo em estudo. O valor de D é sempre maior ou igual a zero e será tanto maior, quanto maior for a discrepância entre o modelo ajustado e os valores observados.

Para se avaliar se um determinado modelo M_A, se ajusta bem aos dados, considera-se o teste de hipóteses com a seguinte hipótese nula:

H_0: O ajustamento do modelo M_A é igual ao ajustamento do modelo M_C.

Sob H_0, e para amostras grandes, D apresenta distribuição assimptótica Qui-quadrado com $J-p-1$ graus de liberdade,

$$D \sim \chi^2_{(J-(p+1))}$$

onde J é o número total de covariáveis existentes nos dados, p é o número de parâmetros do modelo M_A a menos da constante.

2.4.1.2 Percentagem de *Deviance* Explicada

Para os modelos GLM, nomeadamente para os modelos de regressão de Poisson e Binomial Negativo é possível calcular a percentagem de *deviance* explicada (%DevExp)

pelo modelo através da seguinte expressão [Zuur et al, 2009],

$$100 \times \frac{D_{M_0} - D_{M_A}}{D_{M_0}} \tag{2.33}$$

onde D_{M_0} é a *deviance* do modelo nulo e D_{M_A} é a *deviance* do modelo ajustado.

Dobson [Dobson, 2002] chamou-lhe aumento proporcional na *deviance* explicada, considerada similar ao coeficiente de determinação R^2 dos modelos lineares.

2.4.1.3 Estatística de Pearson generalizada

A Estatística de Pearson generalizada é outra medida de adequabilidade de modelos,

$$X^2 = \sum_{i=1}^{n} \frac{w_i(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \tag{2.34}$$

onde $V(\widehat{\mu}_i)$ é a função de variância estimada para a distribuição do modelo em estudo. No caso da distribuição Normal, a estatística X^2 coincide com a soma dos quadrados dos resíduos, enquanto que para os modelos de Poisson e Binomial coincide com a estatística χ^2 original de Pearson.

Uma vez que não é conhecida a distribuição para a diferença entre estatísticas de Pearson, a comparação entre modelos encaixados não pode ser feita usando a diferença entre estatísticas de Pearson, contrariamente ao que sucede com a função desvio.

2.4.2 Análise de Resíduos

A análise de resíduos é útil, para avaliar a qualidade do ajustamento de um modelo no que diz respeito à escolha da distribuição, da função de ligação, do preditor linear e também para identificar observações mal ajustados, e que por isso são mal explicadas pelo modelo [McCullagh and Nelder, 1989].

Os resíduos medem a diferença entre os valores observados y_i e os valores ajustados $\widehat{\mu}_i$. No caso dos modelos lineares generalizados é necessário alargar esta definição para que possam ser aplicados, não só à distribuição normal, mas também às outras distribuições.

2.4.2.1 Resíduos de Pearson

O resíduo de Pearson para uma dada observação é dado por

$$\begin{aligned} R_i^P &= \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\text{Var}\,(Y_i)}}} \\ &= \frac{(y_i - \widehat{\mu}_i) w_i}{\sqrt{\widehat{\phi}\, V(\widehat{\mu}_i)}} \end{aligned} \tag{2.35}$$

O resíduo R_I^P corresponde à contribuição de cada observação para o cálculo da estatística de Pearson generalizada.

O resíduo de Pearson padronizado é dado por

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

uma vez que $Var(Y_i - \widehat{\mu}_i) \approx Var(Y_i)(1 - h_{ii})$, [Turkman and Silva, 2000] onde h_{ii} são os valores da diagonal da matriz de projecção $H = Z_0\,(Z_0^T\, Z_0)^{-1} Z_0^T$, em que $Z_0^T = W^{\frac{1}{2}} Z$.

2.4.2.2 Resíduo *Deviance*

O resíduo *deviance* é dado por

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

onde d_i é a contribuição de cada observação i para a medida de discrepância *deviance*.

O resíduo *deviance* estandardizado é dado por

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

2.4.3 Seleção de modelos

Quando dois modelos não são encaixados não é possível utilizar o teste de razão de verosimilhanças pelo que se torna aconselhável outro critério. Akaike introduziu o *Akaike Information Criterion (AIC)* para a seleção de modelos [Akaike, 1974]. A formulação do *AIC* para selecionar um modelo entre M modelos pode ser expressa por

$$AIC = -2\ell(\beta) + 2p \tag{2.36}$$

onde ℓ é o logaritmo da função de máxima verosimilhança do modelo e p é o número de parâmetros a estimar do modelo. Um valor baixo do *AIC* é considerado como representativo de um melhor ajustamento, por isso na seleção de modelos devemos ter como objetivo a minimização de *AIC*.

Um outro critério foi proposto por Schwarz, o *Bayesian Information Criterion (BIC)* [Schwarz, 1978], sendo expresso por

$$BIC = -2\ell(\beta) + p\, ln(n) \tag{2.37}$$

onde ℓ é o logaritmo da função de máxima verosimilhança do modelo escolhido, p é o número de parâmetros a ser estimado do modelo e n é o número de observações.

De igual forma, um valor baixo do *BIC* é considerado como representativo de um melhor ajustamento.

Bozdogan, propôs a seguinte correção para o AIC, [Bozdogan, 1987]

$$AIC_c = -2\ell(\beta) + 2p + 2\frac{p(p+1)}{n-p-1} \tag{2.38}$$

Alguns autores recomendam o uso do AIC_c quando o tamanho da amostra, n, é relativamente pequeno e o número de parâmetros, p, é muito elevado.

2.4.4 Sobredispersão

Um fenómeno que ocorre com frequência nas aplicações é o fenómeno de sobredispersão. Sobredispersão surge quando a variância da variável resposta é superior ao valor da média. Designando por ϕ o parâmetro de sobredispersão, tal que $Var(Y) = \phi \mathbb{E}(Y) = \phi\mu$, quando ocorre sobredispersão na estimação dos parâmetros do modelo as estimativas pontuais, são iguais, caso não exista sobredispersão, mas a variância dos estimadores é inflacionada pelo parâmetro de sobredispersão ϕ.

Para identificar sobredispersão nos dados, podemos utilizar a *deviance*, estatística também utilizada para testar a qualidade do ajustamento do modelo. O calculo é baseado na aproximação χ^2 do desvio residual. Se existir sobredispersão, então $\frac{D}{\phi}$ segue uma distribuição qui-quadrado com $n-p$ graus de liberdade, e isso leva ao seguinte estimador para ϕ [Zuur et al, 2009]

$$\widehat{\phi} = \frac{D}{n-p} \tag{2.39}$$

Quando este rácio é próximo de um, pode-se assumir a não existência de sobredispersão, prosseguindo-se com o processo de validação do modelo. Caso seja maior que

um, pode haver alguma indicação da presença de sobredispersão nos dados. Alguns autores no entanto apenas recomendam verificar a presença de sobredispersão caso o rácio seja superior a dois [Lindsey, 1999].

Uma ferramenta gráfica adicional para determinar se o modelo é adequado ou se existe sobredispersão nos dados é o *envelope plot*. Este gráfico é parte do gráfico normal quantil-quantil (ou seja, o Q-Q plot), para o qual os resíduos obtidos do modelo ajustado, contra os resíduos teóricos obtidos da distribuição normal, são projetados. Se o gráfico for significativamente diferente de uma linha reta, há indícios claros de que os resíduos não seguem a distribuição normal, o que implica que o modelo ajustado não é adequado para os dados.

O *envelope plot* simula intervalos de confiança empíricos para determinar se os resíduos diferem significativamente da linha recta. O cálculo destes intervalos baseiam-se na simulação de várias amostras para a variável de resposta. Essas amostras são geradas a partir de estimativas obtidas no modelo que foi ajustado tendo em consideração a distribuição assumida para a variável de resposta. Se houver sobredispersão, a projeção dos resíduos cairá fora dos intervalos.

2.4.4.1 Quasi-verosimilhança

Em muitos casos devido à existência de sobredispersão nos dados, é necessário introduzir um parâmetro de sobredispersão, ϕ desconhecido, isto é, admitir que se tem $\mathrm{Var}(y_i) = \phi V(\mu_i)$. Com esta alteração o modelo deixa de estar especificado dentro da família exponencial, uma vez que deixa de existir uma distribuição com estes valores de média e variância, impossibilitando o uso da função de verosimilhança.

No entanto é possível fazer inferências sobre este modelo, considerando modelos de quasi-verosimilhança. Contudo estes modelos não serão abordados nesta dissertação.

Capítulo 3

Modelos de Regressão para Dados de Contagem

Os dados de contagem são um tipo de dados muito frequentes nas mais diversas áreas de estudo, como por exemplo, nas Ciências Sociais, Ciências da Saúde, Engenharia, Ciências Económicas, Ciências Políticas, etc.

A contagem de dados é definida como o número de eventos que ocorrem numa mesma unidade de observação durante um intervalo de tempo ou espaço.

Dados de contagem surgem de várias formas, podendo ser, por exemplo, o número de defeitos, o número de acidentes, o número de ligações perdidas ou o número de vezes que uma tarefa foi concluída.

Para modelar este tipo de dados, habitualmente é usado o modelo de regressão de Poisson, que é construído com base na distribuição de Poisson. Um possível problema no modelo de regressão de Poisson surge quando a variância das respostas é superior ao seu valor médio, designando-se a este fenómeno de sobredispersão. O excesso de zeros nos dados pode também levar a problemas de ajustamento no modelo de regressão de Poisson.

O modelo de regressão Binomial Negativa, que é uma generalização do modelo de regressão de Poisson, permite resolver o problema da sobredispersão, mas não resolve o problema do excesso de zeros.

Os modelos de regressão de zeros inflacionados, foram desenvolvidos para ter em conta o excesso de zeros nos dados. O modelo de regressão de Poisson de zeros inflacionados (ZIP) e o modelo de regressão Binomial Negativa de zeros inflacionados (ZINB) modelam as contagens como uma mistura de duas distribuições com dois processos subjacentes, um processo que trata do excesso de zeros, modelado por uma massa pontual em zero e assumindo que com probabilidade π a única observação possível é zero, e um outro que trata das contagens, modelado por uma distribuição de Poisson ou Binomial

Negativa, com probabilidade $1-\pi$.

Considere-se que as variáveis resposta, $Y = (Y_1, \ldots, Y_n)^T$ são independentes, onde n é o número de observações. Para cada variável Y_i, existem dois processos possíveis para cada modelo considerado. Resumindo,

$$Y_i \sim \begin{cases} 0 & \textit{com probabilidade } \pi_i \\ \textit{Poisson}(\mu_i) \textit{ ou Binomial Negativa}(\mu_i, \alpha) & \textit{com probabilidade } 1-\pi_i, \end{cases}$$

onde π_i corresponde à probabilidade de existir um zero que não deriva de uma contagem de Poisson ou binomial negativa [Zuur et al, 2009].

A função de probabilidade de uma variável de Poisson de zeros inflacionados pode ser visualizado para diferentes valores de π, na Figura 3.1.

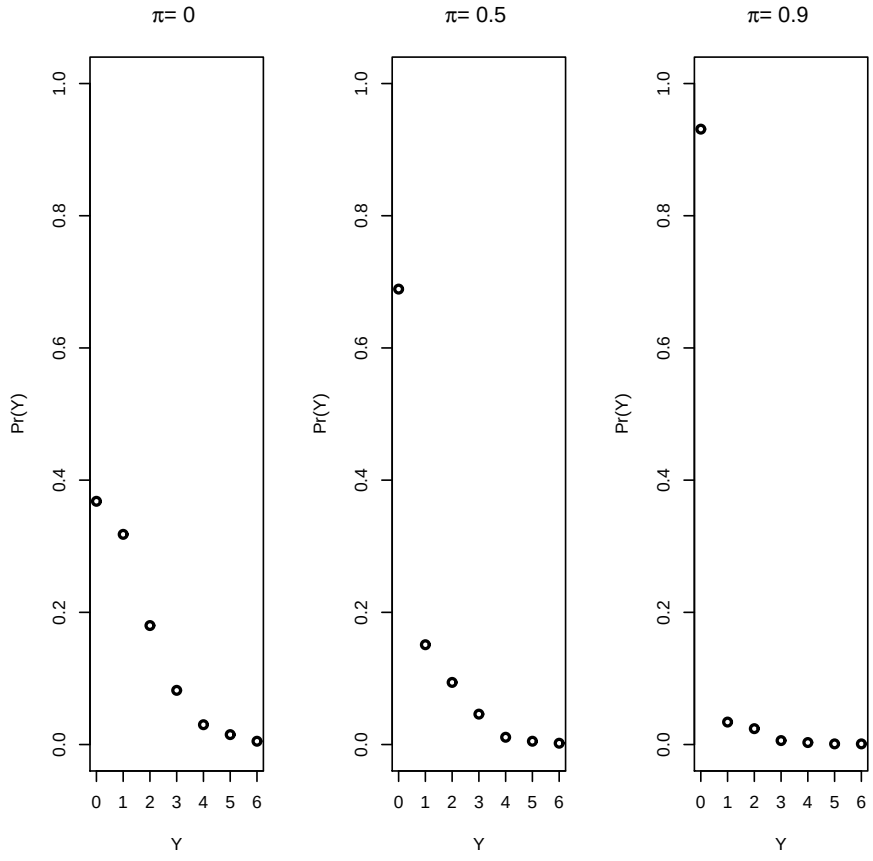


Figura 3.1: Função de probabilidade de uma variável de Poisson com zeros inflacionados para diferentes valores para π.

A função de probabilidade de uma variável Binomial Negativa com zeros inflacionados pode ser visualizado para diferentes valores de π, na Figura 3.2.

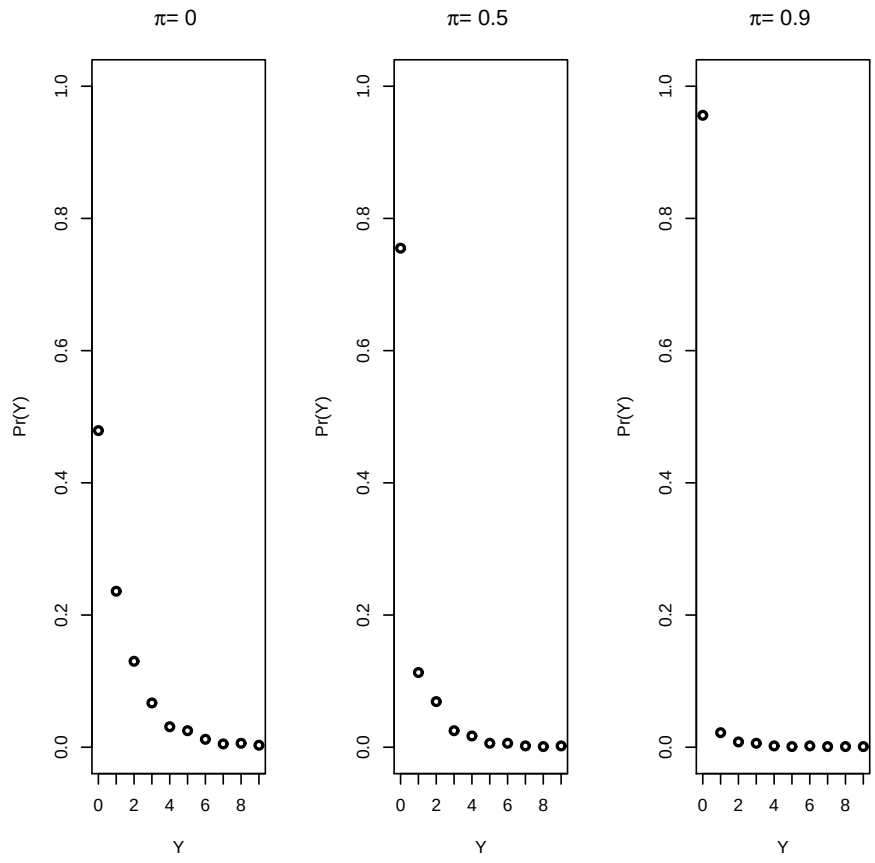


Figura 3.2: Função de probabilidade de uma variável Binomial Negativa com zeros inflacionados para diferentes valores para π.

3.1 Modelo de Regressão de Poisson

Suponhamos que $Y_1, \ldots, Y_n$ são variáveis aleatórias independentes tais que $Y_i \sim P(\mu_i)$, a função de probabilidade de Y_i é dada por

$$f(y_i|\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, y_i = 0, 1, \ldots \tag{3.1}$$

onde μ_i representa o número médio de ocorrências de um determinado acontecimento, $\mu_i > 0$.

A média e a variância são dadas por,

$$\mathbb{E}(Y) = \mathrm{Var}(Y) = \mu$$

Considere a variável aleatória Y que representa o número de ocorrências de um acontecimento num determinado espaço ou periodo de tempo. Dado $X = (X_1, \ldots, X_p)$ um

vetor de covariáveis e $\mathsf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ uma observação do indivíduo i, assume-se

$$Y|X = \mathsf{x}_i \sim P(\mu(\mathsf{x}_i))$$

onde $\mu_i = \mu(\mathsf{x}_i)$ é o número médio de ocorrências de um dado acontecimento dada a observação x_i.

Para se modelar $\mathbb{E}[Y|X = \mathsf{x}_i]$ poderia escrever-se um modelo linear da forma

$$\mu_i = z_i^T \beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

onde β é o vetor dos coeficientes de regressão e $z_i = (1, \mathsf{x}_i^T)^T$.

No entanto, este modelo não pode ser usado, uma vez que o preditor linear pode assumir qualquer valor real, enquanto que μ_i só assume valores não negativos.

Para ultrapassar-se este problema, pode usar-se a transformação logarítmica como função de ligação do modelo linear generalizado e tem-se

$$ln(\mu(\mathsf{x}_i)) = z_i^T \beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Assim, o modelo de regressão de Poisson é dado por

$$Y|X = \mathsf{x}_i \sim P(\mu(\mathsf{x}_i)) \tag{3.2}$$

e

$$ln(\mu(\mathsf{x}_i)) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{3.3}$$

Os coeficientes de regressão β_j, $j = 0, \ldots, p$ representam a variação esperada no logaritmo da média por unidade de variação na covariável X_j.

Conforme apresentado na Secção (2.3), a estimação dos parâmetros do modelo pode ser realizada usando o método da máxima verosimilhança. Para o modelo de regressão de Poisson o logaritmo da função de máxima verosimilhança é dada por

$$\ell(\beta) = \sum_{i=1}^{n} (y_i \, ln(\mu(\mathsf{x}_i)) - \mu(\mathsf{x}_i) - ln(y_i!)) \tag{3.4}$$

onde $ln(\mu(\mathsf{x}_i))$ é dado pela equação (3.3), e $\mu(\mathsf{x}_i)$ depende do vetor de covariáveis x_i e $ln(y_i!)$ é uma constante. Derivando em ordem a β e igualando a zero o lado direito da equação (3.4), prova-se que as estimativas de máxima verosimilhança $\widehat{\beta}$ de β satisfazem

$$X^T y = X^T \widehat{\mu} \tag{3.5}$$

onde $\widehat{\mu}$ é o vetor dos valores previstos pelo modelo.

Para calcular a estimativa de cada parâmetro de regressão é necessário recorrer ao método iterativo dos mínimos quadrados ponderados, conforme descrito na Secção (2.3.2).

Considerando que y_i são os valores observados, e que, $\widehat{\mu}_i$ são os valores previstos pelo modelo de regressão de Poisson, a *deviance* é dada por

$$\begin{aligned} D &= 2\sum_{i=1}^{n}(y_i\, ln(y_i) - y_i - ln(y_i!) - y_i\, ln(\widehat{\mu}_i) + \widehat{\mu}_i + ln(y_i!)) \\ &= 2\sum_{i=1}^{n}\left(y_i\, ln\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right). \end{aligned} \tag{3.6}$$

A *deviance* reduz-se a

$$D = 2\sum_{i=1}^{n}\left(y_i\, ln\left(\frac{y_i}{\widehat{\mu}_i}\right)\right) \tag{3.7}$$

para modelos com termo constante, β_0 porque neste caso

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0$$

A estatística de Pearson generalizada é dada por

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

O modelo de regressão de Poisson encontra-se implementado no package *stats* do *software* R recorrendo à função *glm( )*.

Quando nos modelos de regressão de Poisson existe sobredispersão pode-se recorrer ao modelo de regressão Binomial Negativa.

3.2 Modelo de Regressão Binomial Negativa

Suponhamos então que $Y_1, \ldots, Y_n$ são variáveis aleatórias independentes tais que $Y_i \sim BN(\mu_i, \alpha)$, a função de probabilidade de Y_i é dada por

$$
\begin{aligned}
f(y_i|\mu_i,\alpha) &= \frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \\
&= \binom{y_i+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}, y_i = 0,1,2,\ldots
\end{aligned} \tag{3.8}
$$

onde α é denominado por parâmetro de heterogenidade e fazendo $k = \frac{1}{\alpha}$ e $p = (\frac{1}{1+\alpha\mu_i})$ obtém-se a expressão (2.7) apresentada na Secção 2.1. A média e a variância são dadas por,

$$\mathbb{E}(Y) = \mu$$

$$\mathrm{Var}(Y) = \mu + \alpha\mu^2$$

Repare-se que a variância da distribuição Binomial Negativa tem um termo adicional positivo $\alpha\mu^2$, comparativamente com a variância da distribuição de Poisson, que, em muitos casos, ajuda a ajustar melhor um conjunto de dados onde existe sobredispersão. A distribuição Binomial Negativa aproxima-se à distribuição de Poisson quando α tende para 0 [Cameron and Trivedi, 1998].

Seja Y uma variável aleatória, representando o número de ocorrências de um determinado acontecimento com n observações, $X = (X_1,\ldots,X_p)$ um vetor de covariáveis e $\mathsf{x}_i^T = (x_{i1},\ldots,x_{ip})$ uma observação do indivíduo i, e assume-se

$$Y|X = \mathsf{x}_i \sim BN(\mu(\mathsf{x}_i),\alpha)$$

onde $\mu_i = \mu(\mathsf{x}_i)$ é igual ao número médio de ocorrências de um dado acontecimento dada a observação x_i.

O modelo de regressão Binomial Negativa, é então dado por [Hilbe, 2001]

$$Y|X = \mathsf{x}_i \sim BN(\mu(\mathsf{x}_i),\alpha) \tag{3.9}$$

e

$$ln(\mu(\mathsf{x}_i)) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Na estimação dos parâmetros aplica-se o método da máxima verosimilhança e para o modelo de regressão Binomial Negativa o logaritmo da função de máxima verosimilhança é dado por

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i \, ln \left(\frac{\alpha\mu_i}{1+\alpha\mu_i} \right) - \left(\frac{1}{\alpha} \right) ln(1+\alpha\mu_i) + ln \left(\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)} \right) \right) \quad (3.10)$$

As estimativas de máxima verosimilhança para β e α são obtidas através do algoritmo de mínimos quadrados ponderados conforme descrito na Secção 2.3.2.

No modelo de regressão Binomial Negativa, a *deviance* é dada pela seguinte expressão,

$$D = 2\sum_{i=1}^{n} \left(y_i ln \left(\frac{y_i}{\widehat{\mu}_i} \right) - \left(\frac{1}{\alpha} + y_i \right) ln \left(\frac{1+\alpha y_i}{1+\alpha\widehat{\mu}_i} \right) \right) \quad (3.11)$$

A estatística de Pearson generalizada é dada por

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)}{\widehat{\mu}_i + \alpha\widehat{\mu}_i^2}$$

O modelo de regressão Binomial Negativa encontra-se implementado no package *MASS* do *software* R recorrendo à função *glm.nb( )*.

3.3 Modelo de Regressão para Dados Inflacionados

Os modelos de regressão de zeros inflacionados que serão apresentados terão em conta a distribuição de Poisson e a distribuição Binomial Negativa, uma vez que são estas as distribuições mais utilizadas em dados de contagem [Cameron and Trivedi, 1998]. Na literatura existem trabalhos recentes sobre o modelo ZIP e o modelo ZINB, nas mais diversas áreas.

Na produção [Lambert, 1992] faz análise do número de defeitos num processo de fabricação industrial, propondo uma mistura finita das distribuições de Bernoulli e Poisson para modelar o excesso de zeros em dados de contagem. Na medicina, [Böhning, 1999] apresenta um estudo sobre a análise do número de crianças com cáries dentárias recorrendo ao modelo de regressão ZIP; [Lewsey and Thomson, 2004] comparam os modelos ZIP e ZINB fazendo também um estudo relativo a dentes com cáries; [Yau et al, 2003] ajustam modelo ZINB para estudar a recuperação de doentes que efectuaram cirurgia no fígado; [Lee et al, 2001] utilizam o modelo ZIP para modelar dados provenientes ferimentos ocupacionais. Na epidemologia, [Cheung, 2002] faz um estudo sobre o crescimento e o desenvolvimento de crianças modelando os dados através da utilização de

modelos de regressão de zeros inflacionados. Em ecologia, [Potts and Elith, 2006] efectua estudo sobre a abundância de especies de plantas vulneráveis e [Martin et al, 1989] estuda a presença de determinada espécie num determinado habitat utilizando modelos de contagem com excesso de zeros. Em Ciências Sociais, [Famoye and Singh, 2006] estuda um conjunto de dados sobre violência doméstica. [Ridout et al, 2001] abordam teste *score* para testar se o modelo correcto é o modelo ZIP contra o modelo alternativo ZINB; [Yau et al, 2003] efectuam estudo de dados com sobredispersão e excesso de zeros, usando uma mistura do modelo de ZINB; [Broek, 1995] apresenta uma estatística escore para testar se o modelo ZIP se ajusta melhor que a usual distribuição de Poisson. [Hall, 2000] estuda os modelos ZIP e ZINB incluindo efeitos aleatórios, e outros autores utilizam estes modelos nos mais variados contextos.

Os modelos ZIP e ZINB encontram-se implementados no package *pscl* do *software* R recorrendo à função *zeroinfl( )*.

3.3.1 Modelo de Regressão de Poisson de zeros inflacionados (ZIP)

Suponhamos então que $y_i, \ldots, y_n$ são realizações da variável resposta Y_i, $i = 1, 2, \ldots, n$, o modelo de zeros inflacionado de Poisson é dado por,

$$P(Y_i = y_i|\mathrm{x}_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i}, & y_i = 0 \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, & y_i > 0. \end{cases}$$

em que $0 < \pi_i < 1$ e $\mu_i > 0$.

A média e a variância da distribuição são dadas por,

$$\mathbb{E}(Y_i) = (1-\pi_i)\mu_i$$

$$\mathrm{Var}\,(Y_i) = \mu_i(1-\pi_i)(1+\pi_i\mu_i)$$

O modelo de regressão de Poisson de zeros inflacionados, modela a média μ de uma variável de Poisson através de uma regressão de Poisson e a probabilidade de π através de uma regressão logística com função de ligação $\eta_i = logit(\pi_i)$, ou seja,

$$log(\mu_i) = X_i\beta \tag{3.12}$$

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = G_i\gamma \tag{3.13}$$

em que X_i e G_i são as matrizes de covariáveis. Nestas duas equações de regressão, as duas matrizes de covariáveis podem ou não coincidir [Lambert, 1992]; [Ridout et al, 2001].

Os parâmetros do modelo podem ser estimados aplicando o método de máxima verosimilhança.

Como o ajustamento de um modelo de regressão de Poisson de zeros inflacionados é feito à custa de duas regressões, a função de verosimilhança é dada por

$$L = \prod_{i:y_i=0} [\pi_i + (1-\pi_i)e^{-\mu_i}] \prod_{i:y_i>0} \left[(1-\pi_i) \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \right] \tag{3.14}$$

e logaritmo da função de verosimilhança

$$\ell = \sum_{i:y_i=0} ln(e^{\mathbf{G}_i^T\gamma} + e^{-e^{\mathbf{X}_i^T\beta}}) + \sum_{i:y_i>0} (y_i \mathbf{X}_i^T\beta - e^{\mathbf{X}_i^T\beta} - ln(y_i!)) - \sum_{i=1}^{n} ln(1+e^{\mathbf{G}_i^T\gamma}) \tag{3.15}$$

Os estimadores de máxima verosimilhança de γ e β podem ser obtidos aplicando o método de *scores* de Fisher ou o algoritmo *EM*, [Dempster et al, 1977][Lambert, 1992]. Neste trabalho vamos aplicar o algoritmo *EM*. Este algoritmo consiste basicamente, em um processo iterativo de dois passos, o passo *E* (*Expectation*)que calcula o valor esperado do logaritmo da função de verosimilhança, e o passo *M* (*Maximization*) que é a etapa da maximização na qual utiliza os dados observados e os estimados no passo E. Os passos são repetidos iterativamente até se atingir a convergência.

Considere-se a variável não observada $W = (w_1, \ldots, w_n)^T$ com

$$w_i = \begin{cases} 1 & se\ y_i = 0 \\ 0 & se\ y_i > 0 \end{cases}$$

A função de log-verosimilhança dos dados completos é

$$\begin{aligned}
\ell_c &= ln \prod_{i=1}^{n} \Pr(Y_i = y_i, W_i = w_i) \\
&= \sum_{i=1}^{n} w_i\ ln\ \pi_i + (1-w_i) ln(1-\pi_i)(1-w_i)\ ln\ \left(\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \right) \\
&= \sum_{i=1}^{n} w_i\ ln \left(\frac{e^{\mathbf{G}_i\gamma}}{1+e^{\mathbf{G}_i\gamma}} \right) + (1-w_i) ln \left(\frac{1}{1+e^{\mathbf{G}_i\gamma}} \right) (1-w_i)\ ln\ \left(\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \right) \\
&= \sum_{i=1}^{n} [w_i \mathbf{G}_i^T\gamma - ln(1+e^{\mathbf{G}_i^T\gamma})] + \sum_{i=1}^{n} (1-w_i)(y_i \mathbf{X}_i^T\beta - e^{\mathbf{X}_i^T\beta}) - \sum_{i=1}^{n} (1-w_i) ln(y_i!) \\
&= \ell_c(\gamma) + \ell_c(\beta)
\end{aligned} \tag{3.16}$$

em que

$$\ell_c(\gamma) = \sum_{i=1}^{n} w_i \mathbf{G}_i^T \gamma - ln(1 + e^{\mathbf{G}_i^T \gamma})$$

e

$$\ell_c(\beta) = \sum_{i=1}^{n} (1 - w_i)(y_i \mathbf{X}_i^T \beta - e^{\mathbf{X}_i^T \beta} - ln(y_i!))$$

O passo *E* do algoritmo *EM* consiste em substituir w_i pela sua esperança condicional dado $y, \gamma^{(k)}, \beta^{(k)}$

$$\widehat{w}_i = \mathbb{E}(w_i | y_i, \widehat{\gamma}^{(k)}, \widehat{\beta}^{(k)}) \begin{cases} (1 + e^{\mathbf{G}_i \widehat{\gamma}^{(k)} - e^{\mathbf{X}_i \widehat{\beta}^{(k)}}}) & se\ y_i = 0 \\ 0 & se\ y_i > 0 \end{cases}$$

e tem-se

$$\mathbb{E}(\ell_c) = \sum_{i=1}^{n} [\widehat{w}_i \mathbf{G}_i^T \gamma - ln(1 + e^{\mathbf{G}_i^T \gamma})] + \sum_{i=1}^{n} (1 - \widehat{w}_i)(y_i \mathbf{X}_i^T \beta - e^{\mathbf{X}_i^T \beta}) - \sum_{i=1}^{n} (1 - \widehat{w}_i) ln(y_i!) \tag{3.17}$$

No passo *M* a função log-verosimilhança completa pode ser facilmente maximizada porque a função $\ell_c(\gamma)$ e $\ell_c(\beta)$ podem ser maximizadas separadamente. O passo *M* para estimar β consiste em encontrar $\beta^{(k+1)}$ maximizando $\ell_c(\beta)$ o que é o mesmo que maximizar a função log-verosimilhança ponderada para o modelo Poisson utilizando os pesos $1 - \widehat{w}_i^{(k)}$, y_i como variável resposta e uma função de ligação. O passo *M* para estimar γ consiste em encontrar $\gamma^{(k+1)}$ maximizando $\ell_c(\gamma)$ o que é o mesmo que maximizar a função log-verosimilhança para a regressão logistica não ponderada de $\widehat{w}^{(k)}$ em G.

No modelo de regressão de Poisson de zeros inflacionados a *deviance*, é calculada como a diferença entre o logaritmo da função de máxima verosimilhança dos dois modelos multiplicada por -2. Para amostras grandes a distribuição da *deviance* segue aproximadamente uma distribuição qui-quadrado com $n - p$ graus de liberdade, em que n é o número de observações e p é o número de parâmetros estimados.

3.3.2 Modelo de Regressão Binomial Negativo de zeros inflacionados (ZINB)

O modelo de regressão Binomial Negativo com zeros inflacionados é dado por,

$$P(Y_i = y_i | \mathsf{x}_i) = \begin{cases} \pi_i + (1 - \pi_i) \left(\frac{1}{1 + \alpha\mu} \right)^{\alpha^{-1}} & y_i = 0 \\ (1 - \pi_i) \frac{\Gamma(y_i + \alpha^{-1})}{y_i! \Gamma(\alpha^{-1})} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i} \right)^{y_i} \left(\frac{1}{1 + \alpha\mu_i} \right)^{\alpha^{-1}} & y_i > 0 \end{cases}$$

em que $0 \leq \pi_i < 1, \mu_i > 0$. A média e a variância são dadas por,

$$\mathbb{E}(Y_i) = (1-\pi_i)\mu_i$$

$$\text{Var}(Y_i) = \mu_i(1-\pi_i)(1+\pi_i\mu_i+\alpha\mu_i)$$

A função de verosimilhança é dada por

$$\begin{aligned} L &= \prod_{i:y_i=0}\left[\pi_i+(1-\pi_i)\left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}}\right] \\ &\quad \prod_{i:y_i>0}\left[(1-\pi_i)\frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\right] \end{aligned} \tag{3.18}$$

e logaritmo da função de verosimilhança é dado por

$$\begin{aligned} \ell &= -ln\left[1+e^{\mathbf{G}_i^T\gamma}\right]+ln\left[e^{\mathbf{X}_i^T\beta}+\left(\frac{1}{1+\alpha e^{\mathbf{X}_i^T\beta}}\right)^{\alpha^{-1}}\right] \\ &\quad -ln\left[1+e^{\mathbf{G}_i^T\gamma}\right]+ln\left[\Gamma\left(\frac{1}{\alpha}+y_i\right)\right]-ln[\Gamma(y_i+1)] \\ &\quad -ln\left[\Gamma\left(\frac{1}{\alpha}\right)\right]+y_i\,ln\left[\frac{\alpha e^{\mathbf{X}_i^T\beta}}{1+\alpha e^{\mathbf{X}_i^T\beta}}\right] \end{aligned} \tag{3.19}$$

Considerando a variável não observada $W=(w_1,\ldots,w_n)^T$ com

$$w_i = \begin{cases} 1 & se\ y_i = 0 \\ 0 & se\ y_i > 0 \end{cases}$$

A função de log-verosimilhança dos dados completos é

$$\begin{aligned} \ell_c &= ln\prod_{i=1}^{n}\Pr(Y_i=y_i,W_i=w_i) \\ &= \sum_{i=1}^{n} w_i\ ln\ \pi_i+(1-w_i)ln(1-\pi_i)+(1-w_i)\ ln\ g(y_i;\beta,\alpha^{-1}) \\ &= \sum_{i=1}^{n} w_i\mathbf{G}_i\gamma-ln(1+e^{\mathbf{G}_i^T\gamma})+(1-w_i)ln\ g(y_i;\beta,\alpha^{-1}) \\ &= \ell_c(\gamma)+\ell_c(\beta) \end{aligned} \tag{3.20}$$

onde

$$g(y_i;\beta,\alpha^{-1}) = \frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}$$

O passo E do algoritmo EM consiste em substituir w pela sua esperança condicional dado $y, \gamma^{(k)}, \beta^{(k)}$

$$\widehat{w}_i = \mathbb{E}(w_i|y_i,\widehat{\gamma}_i^{(k)},\widehat{\beta}_i^{(k)}) \begin{cases} (1+e^{-\mathbf{G}_i\widehat{\gamma}^{(k)}}\left[\frac{1}{\widehat{\alpha}e^{\mathbf{x}_i^{\widehat{\beta}}}+1}\right]^{\frac{1}{\widehat{\alpha}}})^{-1} & se\ y_i = 0 \\ 0 & se\ y_i > 0 \end{cases}$$

O passo M a função log-verosimilhança pode ser facilmente maximizada porque a função $\ell_c(\gamma)$ e $\ell_c(\beta)$ podem ser maximizadas separadamente. O passo M para calcular β consiste em encontrar $\beta^{(k+1)}$ maximizando $\ell_c(\beta)$ o que é o mesmo que maximizar a função log-verosimilhança ponderada para o modelo Binomial Negativa utilizando os pesos $1-\widehat{w}_i^{(k)}$, y_i como variável resposta e uma função de ligação. O passo M para calcular γ consiste em encontrar $\gamma^{(k+1)}$ maximizando $\ell_c(\gamma)$ o que é o mesmo que maximizar a função log-verosimilhança para a regressão logistica não ponderada de $\widehat{w}^{(k)}$ em G.

3.3.3 Resíduos

Os resíduos de Pearson para os modelos de regressão de zeros inflacionados, ZIP e ZINB, podem ser obtidos pela seguinte expressão [Zuur et al, 2009],

$$X_i^2 = \frac{Y_i - (1-\pi_i)\mu_i}{\sqrt{\text{Var}\,(Y_i)}} \tag{3.21}$$

3.3.4 Teste Vuong

Vuong [Vuong, 1989] introduziu um teste que é um método adequado para comparar modelos aninhados. Em particular utiliza-se este teste nos modelos ZIP e Poisson bem como nos modelos Binomial Negativa.

Seja $P_N(y_i|x_i)$ a probabilidade prevista de uma contagem observada para o caso i de um dado modelo N e m_i é definido da seguinte forma:

$$m_i = ln\left(\frac{P_1(y_i|x_i)}{P_2(y_i|x_i)}\right)$$

Para testar a $H_0 : \mathbb{E}(m_i) = 0$ a estatística de teste é dada por

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}} \tag{3.22}$$

em que n é a dimensão da amostra.

Sob a hipótese nula, a estatística de teste é assintoticamente normalmente distribuída. Para um nível de significância de 5%, o primeiro modelo é preferível se $V > 1.96$, no entanto se $V < -1.96$ então o segundo modelo é o melhor modelo, caso $|V| < 1.96$ os dois modelos são equivalentes.

O teste de Vuong encontra-se implementado no package *pscl* do *software* R.

Capítulo 4

Aplicação a dados reais

Neste capítulo, a metodologia dos modelos lineares generalizados para dados de contagem e dos modelos de regressão com excessos de zeros serão aplicados a dados reais. As análises estatísticas destes modelos foram efectuadas no *software* livre R versão (2.13.2).

4.1 Base de Dados

Neste trabalho, os dados utilizados dizem respeito a uma amostra aleatória de clientes a quem foi garantido crédito de consumo por um banco. O crédito de consumo, também conhecido como crédito pessoal, é um tipo de crédito atribuido a pessoas individuais para propósitos pessoais, familiares ou relacionados com as despesas da casa. Este tipo de empréstimo normalmente tem um curto periodo de saldação da dívida, não têm um propósito específico e não são necessariamente cobertos por garantias.

A amostra foi tirada a 31 de Dezembro de 2011 e contém 5366 observações.

Todos os dados correspondem a clientes que lhes foi garantido crédito e em que o contrato ainda não foi finalizado. Para cada cliente foi recolhida informação sobre as várias caracteristicas no início do contrato, assim como, o número total consecutivo de meses sem pagamento da prestação.

Empréstimos com um número superior a 12 meses de incumprimento, foram excluídos porque, nestes casos, acção judicial e custos adicionais ocorrem.

Na Tabela 4.1 são apresentadas as variáveis que serão alvo de estudo.

Algumas das variáveis categóricas da amostra apresentaram valores em falta, como uma das categorias já existente era uma categoria descrita como desconhecida, estes valores em falta foram incluídos neste categoria.

O principal objetivo da análise é aplicar a metodologia dos modelos de regressão para dados de contagem, para explicar o número de não pagamentos (incumprimento) da

prestação do crédito pelo cliente em função das caracteristicas do cliente e do contrato.

Tabela 4.1: Informação sobre os dados.

Variável	Descrição
IdadeContrato	Número de meses desde a data do contrato à data da amostra.
MontanteContratado	Montante contratado pelo cliente (euros).
CapitalVincendo	Capital liquidado até 31 Dezembro de 2011 (euros).
NMesesLC	Número de meses até liquidação do montante contratado.
PrestacaoMensal	Valor da prestação mensal do empréstimo (euros).
Idade	Idade do cliente (anos).
SldMdSem.cat	Saldo médio semestral do cliente (euros).
NAnosCliente	Número de anos que é cliente do banco.
Sexo	Sexo do cliente.
EstadoCivil	Estado civil do cliente.
Habilitacoes	Habilitações literárias do cliente.
Profissao	Profissão do cliente.
Ordenado	Indicador se o cliente recebe o ordenado através do banco.
Regiao	Região de residência do cliente.
MesessemPagar	Número consecutivo de meses sem pagamento da prestação . (variável dependente)

A variável número consecutivo de meses sem pagamento (*MesessemPagar*) é a variável dependente, e as suas frequências podem ser observadas na Tabela 4.2. A média e a variância desta variável é 0.853 e 3.378, respetivamente, sugerindo sobredispersão dos dados. O gráfico de barras pode ser visualizado na Figura 4.1.

Tabela 4.2: Tabela de frequências do número de meses sem pagamento.

MesessemPagar	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
0	3779	70.42	3779	70.42
1	669	12.47	4448	82.89
2	248	4.62	4696	87.51
3	199	3.71	4895	91.22
4	120	2.24	5015	93.46
5	130	2.42	5145	95.88
6	59	1.10	5204	96.98
7	70	1.30	5274	98.28
8	42	0.78	5316	99.06
9	21	0.39	5337	99.45
10	11	0.20	5348	99.65
11	10	0.20	5358	99.84
12	8	0.15	5366	100.00

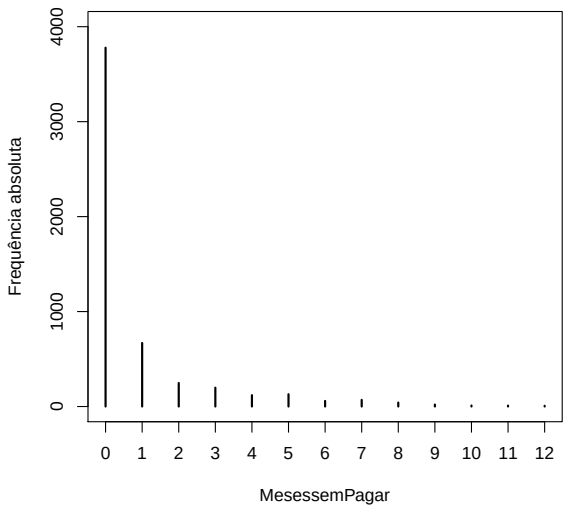


Figura 4.1: Gráfico de Barras da variável *MesessemPagar*

4.2 Análise descritiva

O principal objetivo desta secção é sumariar e descrever as variáveis explicativas usadas neste trabalho. Para as variáveis quantitativas são apresentadas as medidas de tendência central, medidas de dispersão, de assimetria e achatamento, assim como caixas com bigodes, histogramas e gráficos de barras. Para as variáveis qualitativas, são apresentados tabelas de frequências.

4.2.1 Variáveis Explicativas Quantitativas

Com o objetivo da visualização de caracteristicas das variáveis em estudo de forma simples e de fácil interpretação, a caixa com bigodes é uma das representações mais utilizadas. Da Figura 4.2 até à Figura 4.8, as caixas com bigodes, os histogramas e gráficos de barras apresentados, sugerem que as variáveis *MontanteContratado*, *CapitalVincendo*, *PrestacaoMensal* têm distribuições enviesadas à direita, ou enviesamento positivo, uma vez que se apresentam concentradas no lado esquerdo com uma longa cauda para a direita. A variável *NAnosCliente* tem distribuição moderadamente enviesada à direita. As variáveis *Idade*, *IdadeContrato* e *NMesesLC* apresentam uma distribuição aproximadamente simétrica.

De entre as várias formas de caraterizar amostras tomam especial importância as medidas de tendência central, as medidas de dispersão e as medidas de assimetria e acha-

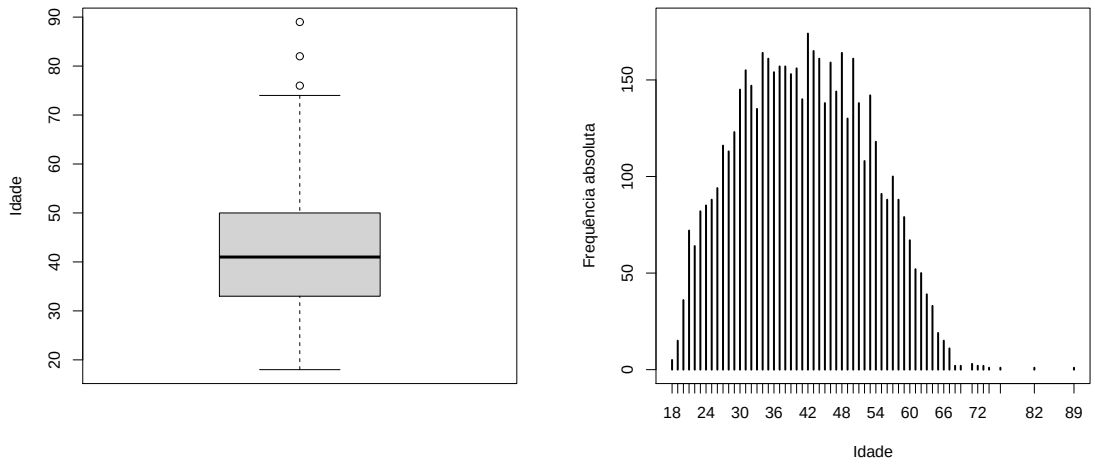


Figura 4.2: Caixa com bigodes e gráfico de barras para a variável *Idade*

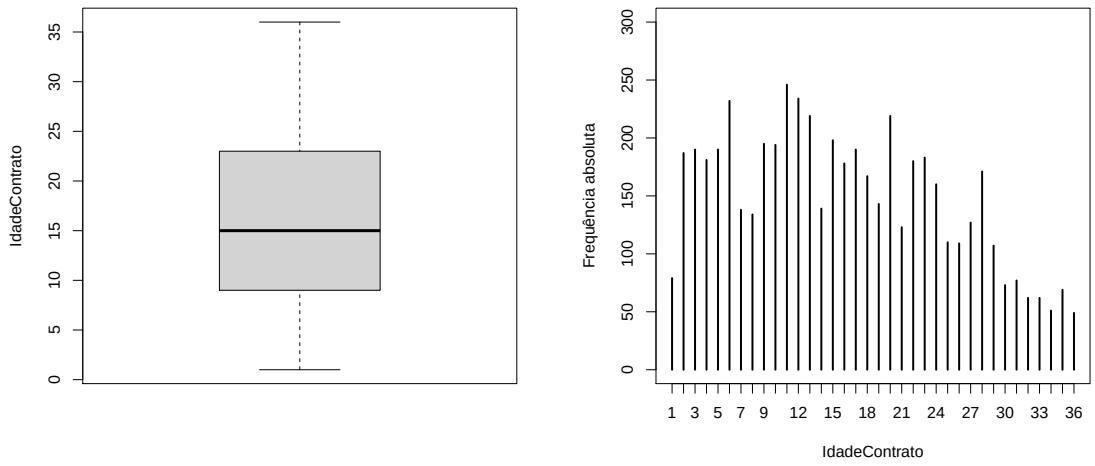


Figura 4.3: Caixa com bigodes e gráfico de barras para a variável *IdadeContrato*

tamento. Das medidas de tendência central as estatísticas mais utilizadas são a média, a mediana e a moda. As medidas de dispersão mais utilizadas são a variância, o desvio padrão e o coeficiente de variação que caracterizam a dispersão das observações em torno das estatísticas de tendência central. As medidas de assimetria e achatamento, que caracterizam a forma da distribuição dos elementos da amostra em torno da média, mais utilizadas são o coeficiente de assimetria e o coeficiente de achatamento ou kurtose.

Na Tabela 4.3, o mínimo, a mediana, a média, o máximo e o desvio padrão de cada variável são apresentadas, assim como o coeficiente de variação.

A idade dos clientes a quem foi concedido crédito varia entre os 18 e os 89 anos, respectivamente, sendo a média das idades dos clientes aproximadamente igual a 41 anos.

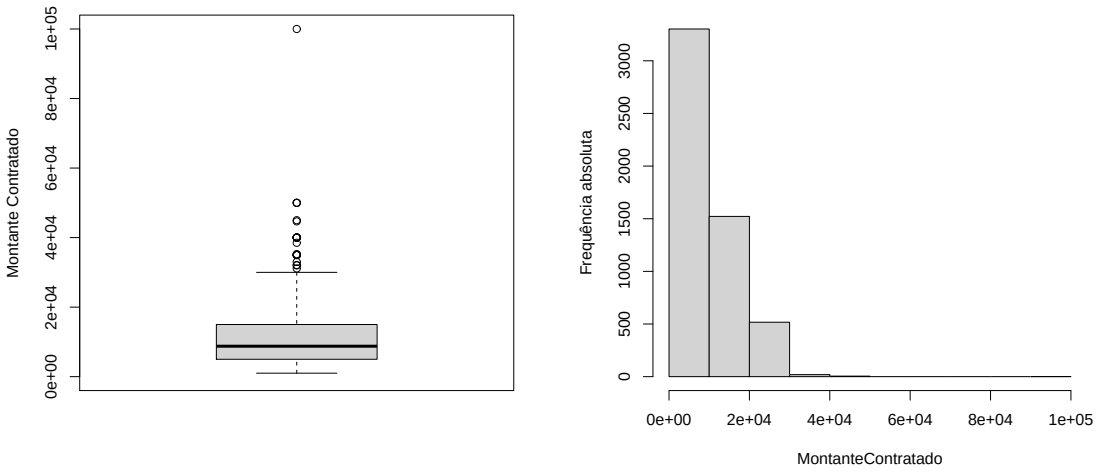


Figura 4.4: Caixa com bigodes e histograma para a variável *MontanteContratado*

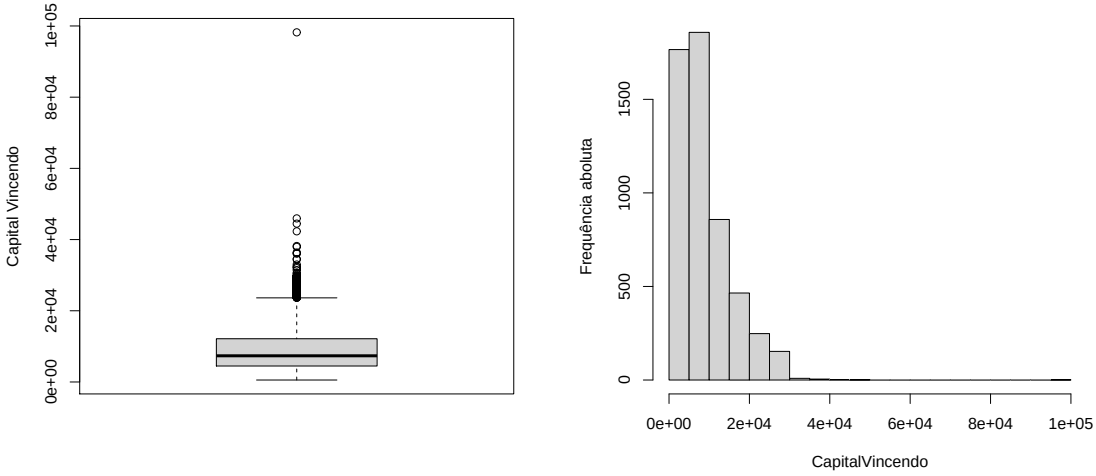


Figura 4.5: Caixa com bigodes e histograma para a variável *CapitalVincendo*

A variável *MontanteContratado* apresenta um mínimo de 1000 e um máximo de 100000 euros, sendo a média do valor do montante contratado pelo cliente de 10800 euros. O valor médio da mensalidade foi de 192 euros.

As variáveis *MontanteContratado* e *PrestacaoMensal* apresentam uma dispersão semelhante, enquanto que a variável *NMesesLC* e a variável *CapitalVincendo* apresentam a menor e a maior dispersão, respetivamente.

Os valores apresentados na Tabela 4.4, confirmam a descrição já efectuada no início da Secção 4.2.1 relativamente à assimetria das variáveis quantitativas.

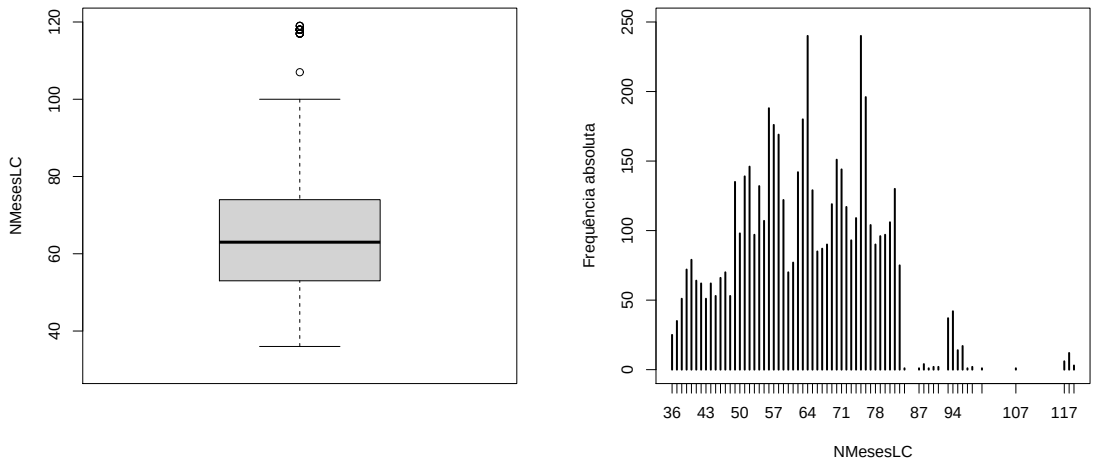


Figura 4.6: Caixa com bigodes e gráfico de barras para a variável *NMesesLC*

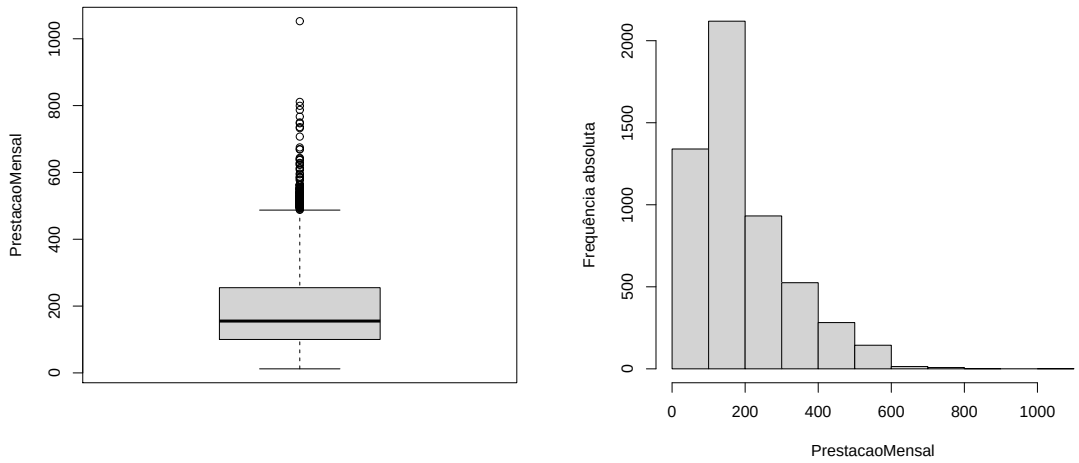


Figura 4.7: Caixa com bigodes e histograma para a variável *PrestacaoMensal*

Tabela 4.3: Tabela das medidas de tendência central e dispersão das variáveis quantitativas

Variável	Mínimo	Mediana	Média	Máximo	Desvio Padrão	Coeficiente Variação
IdadeContrato	1	15	15.89	36	9.122	0.574
MontanteContratado	1000	8756	10863.50	100000	7458.938	0.687
CapitalVincendo	500.4	7351.3	9189.60	98205.5	6467.464	0.704
NMesesLC	36	63	63.19	119	13.438	0.213
PrestacaoMensal	12.15	155.15	192.09	1052.64	124.027	0.646
Idade	18	41	41.41	89	11.248	0.272
NAnosCliente	3	9	10.19	30	5.181	0.509

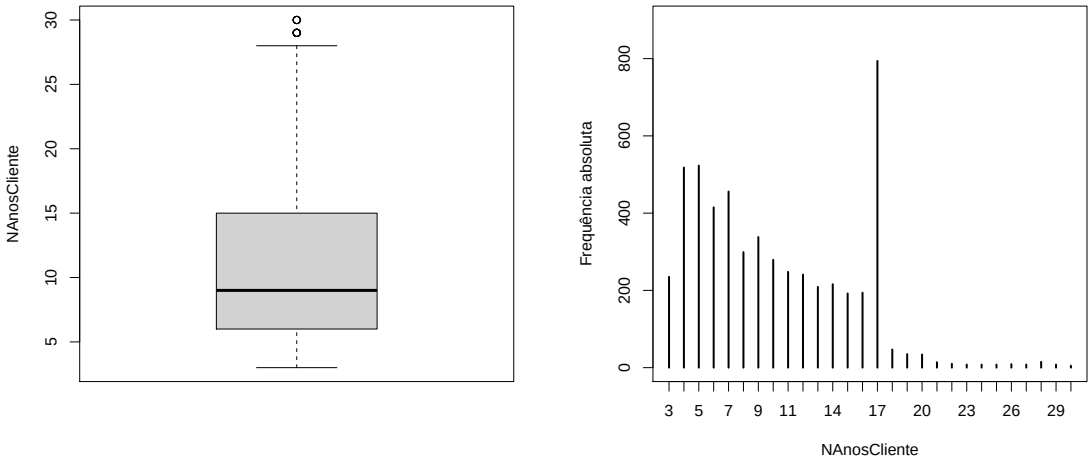


Figura 4.8: Caixa com bigodes e gráfico de barras para a variável *NAnosCliente*

Tabela 4.4: Tabela das medidas de assimetria e achatamento das variáveis quantitativas

Variável	Coeficiente Assimetria	Coeficiente Achatamento
IdadeContrato	0.271	-0.895
MontanteContratado	1.510	4.772
CapitalVincendo	1.764	8.026
NMesesLC	0.241	0.184
PrestacaoMensal	1.320	1.676
Idade	0.123	-0.723
NAnosCliente	0.567	-0.284

4.2.2 Variáveis Explicativas Qualitativas

Da Tabela 4.5 à Tabela 4.9 são apresentadas as frequências das variáveis categóricas.

Tabela 4.5: Tabela de frequências da variável *Sexo*.

Código	Descrição	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	Feminino	1807	33.7	1807	33.7
2	Masculino	3559	66.3	5366	100.0

Das Tabelas 4.5 e 4.6 podemos verificar que aproximadamente dois terços dos indivíduos na amostra são do sexo masculino, e que aproximadamente dois terços dos clientes não recebem o ordenado através do banco.

Das Tabelas 4.7 e 4.8 podemos verificar que a categoria 'Casado' para a variável *EstadoCivil* é a mais frequente, com mais de metade dos clientes da amostra categorizados

Tabela 4.6: Tabela de frequências da variável *Ordenado*.

Código	Descrição	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	Não	3474	64.7	3474	64.7
2	Sim	1892	35.3	5366	100.0

Tabela 4.7: Tabela de frequências da variável *EstadoCivil*.

Código	Descrição	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	Desconhecido	665	12.4	665	12.4
2	Casado	2780	51.8	3445	64.2
3	Divorciado	513	9.6	3958	73.8
4	Solteiro	1311	24.4	5269	98.2
5	Viúvo	97	1.8	5366	100.0

Tabela 4.8: Tabela de frequências da variável *Habilitacoes*.

Código	Descrição	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	Ensino Secundário	1632	30.4	1632	30.4
2	Ensino Superior	319	5.9	1951	36.4
3	Escolaridade Obrigatória	2027	37.8	3978	74.1
4	Formação Profissional	182	3.4	4160	77.5
5	Desconhecidas	1206	22.5	5366	100.0

desta forma, enquanto que as categorias ’Viúvo’ e ’Divorciado’ apresentam as menores frequências. Relativamente à variável *Habilitacoes*, é possível verificar-se que a categoria ’Escolaridade Obrigatória’ é a mais frequente com aproximadamente 38% dos clientes classificados desta forma.

Quanto à variável *Regiao* pela Tabela 4.9 podemos observar que a categoria mais frequente é os ’Açores’ enquanto que a categoria menos frequente é ’Alentejo e Algarve’. Repare-se que aproximadamente 50% dos clientes residem nas ilhas. Pela Tabela 4.10 para a variável *Profissao* verifica-se que as categorias mais frequentes são ’Outros’ e ’Emp. Escrit./Comer./Serv.’ e que aproximadamente 28% dos clientes não se conhece a profissão.

Por último, para a variável *SldMdSem.cat*, através da Tabela 4.11 verifica-se que clientes com saldo médio inferior a 11.5 e clientes com saldo médio superior a 325 são os mais frequentes, com 35% e 40% classificados desta forma, enquanto que os outros clientes

Tabela 4.9: Tabela de frequências da variável *Regiao*.

Código	Descrição	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	Açores	1952	36.4	1952	36.4
2	Alentejo e Algarve	329	6.1	2281	42.5
3	Centro	469	8.7	2750	51.2
4	Lisboa e Vale do Tejo	933	17.4	3683	44.7
5	Madeira	689	12.8	4372	81.5
6	Norte	994	18.5	5366	100.0

Tabela 4.10: Tabela de frequências da variável *Profissao*.

Código	Descrição	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	Desconhecida	1476	27.5	1476	27.5
2	Emp. Escrit./Comer./Serv.	1534	28.6	3010	56.1
3	Estudante ou Doméstica	268	5.0	3278	61.1
4	Pequeno / Médio Empresário	347	6.5	3625	67.6
5	Quadro Médio	80	1.5	3705	69.1
6	Outros	1661	31.0	5336	100.0

Tabela 4.11: Tabela de frequências da variável *SldMdSem.cat*.

Código	Descrição	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	$-\infty$ a 11.5	1882	35.1	1882	35.1
2	11.5 a 136.5	705	13.1	2587	48.2
3	136.5 a 325	605	11.3	3192	59.5
4	325 a $+\infty$	2174	40.5	5366	100.0

correspondem a menos de 25% dos observados.

4.2.3 Associação entre a variável dependente e as variáveis explicativas

As medidas de associação, quantificam a intensidade e a direcção da associação entre variáveis. Nesta secção pretende-se estudar a presença ou ausência de correlação entre a variável dependente *MesessemPagar* e cada uma das variáveis explicativas. O coeficiente utilizado foi o coeficiente de correlação de Spearman.

Na Tabela 4.12 é apresentado o coeficiente de correlação de Spearman entre a variável

dependente *MesessemPagar* e cada uma das variáveis explicativas quantitativas, assim como o p-valor associado ao teste de hipótese do coeficiente de correlação ser zero entre as variáveis, com o objetivo de medir a associação entre as variáveis.

Para um nível de significância de 5%, as variáveis *IdadeContrato* e *MontanteContratado*, apresentam uma correlação significativa positiva, isto é, existe uma tendência para o número de meses sem pagar crescer à medida que os valores destas variáveis aumentam. Enquanto as variáveis *NAnosCliente*, *NMesesLC* e *CapitalVincendo* apresentam uma correlação significativa negativa. Relativamente às variáveis *PrestacaoMensal* e *Idade*, estas apresentam uma correlação não significativa, indicando a não associação destas variáveis com a variável dependente.

Na Tabela 4.13 é apresentado o coeficiente de correlação de Spearman entre as variáveis quantitativas. Coeficientes superiores a 0.75 estão destacados em negrito. Morrison [Morrison, 2004] sugere apenas incluir covariáveis no modelo de regressão que têm uma correlação entre si que é inferior a 0.75 em valor absoluto. Na presença de covariáveis que são altamente correlacionadas uma com a outra, o conselho é manter a covariável que é a mais altamente correlacionada com a variável dependente e deixar cair a outra.

Neste caso as variáveis *MontanteContratado*, *CapitalVincendo* e *PrestacaoMensal* apresentam correlações bastante altas. Tendo em conta os coeficientes nas Tabelas 4.12 e 4.13, optou-se por manter a variável *PrestacaoMensal* no estudo uma vez que parece fazer mais sentido do ponto de vista empresarial, retirando *MontanteContratado* e *CapitalVincendo*.

4.2.4 Teste Kruskal-Wallis

Nesta secção pretende-se comparar a distribuição da variável dependente *MesessemPagar* nas várias categorias das variáveis qualitativas.

Na Tabela 4.14, é apresentada a estatística de teste e o p-valor do teste de Kruskal-

Tabela 4.12: Coeficiente de Correlação de *Spearman* das covariáveis quantitativas e a variável *MesessemPagar*.

Variável	Coeficiente	p-valor
MontanteContratado	0.03199	0.0191
CapitalVincendo	-0.04599	0.0008
PrestacaoMensal	0.01998	0.1433
Idade	0.00682	0.6173
IdadeContrato	0.47462	$<$ 2e-16
NAnosCliente	-0.04474	0.0010
NMesesLC	-0.31298	$<$ 2e-16

Tabela 4.13: Coeficiente de Correlação de *Spearman* das covariáveis quantitativas.

Variável							
MontanteContratado	1						
CapitalVincendo	**0.968**	1					
PrestacaoMensal	**0.956**	**0.962**	1				
Idade	0.139	0.131	0.136	1			
IdadeContrato	0.147	-0.011	0.096	0.066	1		
NAnosCliente	0.099	0.078	0.093	0.248	0.116	1	
NMesesLC	0.277	0.392	0.177	-0.018	-0.539	-0.026	1

Wallis entre a variável dependente *MesessemPagar* e cada uma das variáveis qualitativas. Como se pode verificar, ao rejeitar-se a hipótese nula da igualdade das distribuições, pode-se afirmar que existe evidência estatística de que entre as categorias das variáveis *EstadoCivil*, *Habilitacoes*, *Profissao*, *Ordenado*, *Região*, *SldMdSem.cat* e a variável *MesessemPagar* ocorrem diferenças significativas em termos do número de meses consecutivos sem pagamento da prestação mensal. O que não acontece com a variável *Sexo*, o que significa que não existe diferença entre os clientes do sexo masculino e do sexo feminimo em termos do número de meses consecutivos sem pagamento da prestação mensal ao banco.

Tabela 4.14: Teste Kruskal-Wallis para as covariáveis qualitativas e a variável *MesessemPagar*.

Variável	Estatística	p-valor
Sexo	2.4100	0.1206
EstadoCivil	22.9060	0.0001
Habilitacoes	33.0434	1.17e-06
Profissao	35.7661	1.058e-06
Ordenado	111.016	$<$ 2e-16
Regiao	185.6490	$<$ 2e-16
SldMdSem.cat	322.3587	$<$ 2e-16

4.3 Seleção de Modelos

Uma vez que o principal objetivo deste trabalho é analisar modelos de regressão para dados de contagem, para explicar o número de não pagamentos em função de determinadas caracteristicas dos clientes e do contrato, iniciou-se este estudo estimando o modelo de regressão de Poisson com apenas uma variável explicativa de cada vez. Os resultados podem ser visualizados na Tabela 6.1 que se encontra em anexo.

Da Tabela 6.1 verifica-se que, quando modeladas individualmente, com exceção das

variáveis *Idade* e *Sexo*, todas as outras são estatisticamente significativas.

O método de seleção *stepwise* é um método de seleção em que a escolha das variáveis preditivas é realizada por um procedimento automático, no entanto decidiu-se não utilizar este método devido ao facto de que alguns autores referirem que nem sempre esse método deve ser o método preferencial [Kleinbaum and Klein, 2002].

Para avaliar a multicolineariedade das variáveis explicativas utilizou-se o *Fator de Inflação de Variação* (*VIF*), que mede o quanto a variância de um coeficiente de regressão estimado é maior devido à colinearidade entre as variáveis explicativas.

Os valores de *VIF* obtidos, foram valores elevados ($VIF > 10$) para as variáveis *MontanteContratado*, *CapitalVincendo*, *PrestacaoMensal*, significando que estas variáveis estão linearmente dependentes, decidiu-se retirar do estudo as variáveis *MontanteContratado*, *CapitalVincendo*. Este facto já tinha sido observado na Tabela 4.13.

- Modelo de regressão Poisson

Resolvido o problema da multicolinearidade, começou-se por ajustar o modelo seguinte,

Modelo Inicial:

log(MesessemPagar)= β_0 + β_1 * Sexo + β_2 * EstadoCivil + β_3 * PrestacaoMensal + β_4 * Profissao + β_5 * Idade + β_6 * NMesesLC + β_7 * Habilitacoes + β_8 * NAnosCliente + β_9 * Regiao + β_{10} * Ordenado + β_{11} * IdadeContrato + β_{12} * SldMdSem.cat

De seguida foram retiradas do modelo as variáveis estatísticamente não significativas, sucessivamente, até se obter o modelo final em que todas as variáveis restantes são significativas. O modelo selecionado foi o seguinte modelo,

Modelo Final:

log(MesessemPagar)= β_0 + β_1 * Profissao + β_2 * Idade + β_3 * NMesesLC + β_4 * Habilitacoes + β_5 * NAnosCliente + β_6 * Regiao + β_7 * Ordenado + β_8 * IdadeContrato + β_9* SldMdSem.cat

Na Tabela 4.15, representa-se os valores das estatísticas de ajustamento desses modelos.

Tabela 4.15: Estatísticas de ajustamento dos modelos de regressão de Poisson

Estatísticas	Modelo Inicial	Modelo Final
% Dev.Exp.	41.95	41.86
AIC	12345.90	12346.33
BIC	12536.95	12497.85
ℓ	-6143.95	-6150.17
X^2	11694.40	11662.49
Parâmetros Estimados	29	23

Aplicou-se o teste de razão de verosimilhanças e concluiu-se que não existem diferenças entre os modelos na qualidade do ajustamento ($p\text{-}valor = 0.0530$). Assim o modelo escolhido, apesar do valor do AIC ser ligeiramente superior, foi o modelo Final com um menor número de variáveis explicativas pois as restantes estatísticas de ajustamento sugerem que o modelo melhorou, podendo os seus coeficientes ser visualizados na Tabela 4.16.

Tabela 4.16: Modelo de regressão de Poisson

Variável	Coeficiente	Erro Padrão	p-valor
constante	-0.3024	0.1405	0.0314
ProfissaoEscritComercServiços	0.0256	0.0455	0.5741
ProfissaoEstudanteDomestica	-0.0653	0.0803	0.4160
ProfissaoOutros	0.0924	0.0428	0.0307
ProfissaoPeqMedEmpresario	0.1028	0.0600	0.0869
ProfissaoQuadroMedio	-0.2812	0.1389	0.0430
Idade	-0.0045	0.0014	0.0018
NMesesLC	-0.0054	0.0016	0.0009
HabilitacoesEnsinoSuperior	-0.2058	0.0844	0.0148
HabilitacoesEscolaridadeObrig.	0.0967	0.0392	0.0137
HabilitacoesFormacaoProfiss.	-0.0262	0.0884	0.7672
HabilitacoesDesconhecidas	-0.0316	0.0431	0.4634
NAnosCliente	-0.0267	0.0040	1.45e-11
RegiaoAlentejoAlgarve	0.7209	0.0673	$<$ 2e-16
RegiaoCentro	0.5419	0.0607	$<$ 2e-16
RegiaoLisboaValeTejo	0.5496	0.0548	$<$ 2e-16
RegiaoMadeira	0.6088	0.0525	$<$ 2e-16
RegiaoNorte	0.5666	0.0507	$<$ 2e-16
OrdenadoSim	-1.2036	0.0421	$<$ 2e-16
IdadeContrato	0.0770	0.0021	$<$ 2e-16
SldMdSem.cat2	-0.7937	0.0411	$<$ 2e-16
SldMdSem.cat3	-1.2469	0.0534	$<$ 2e-16
SldMdSem.cat4	-1.8408	0.0449	$<$ 2e-16

Analisando os valores apresentados na Tabela 4.16, o teste de Wald, para um nível de significância de 5%, indica-nos que todas as variáveis são estatísticamente significativas.

Este modelo contudo evidencia problemas de sobredispersão, uma vez que apresenta um valor de $\widehat{\phi} = 1.5008$ [Zuur et al, 2009]. A sobredispersão no modelo, é possível também ser visualizada no *envelope plot* da Figura 4.9, em que uma grande parte dos resíduos não pertence ao intervalo de confiança.

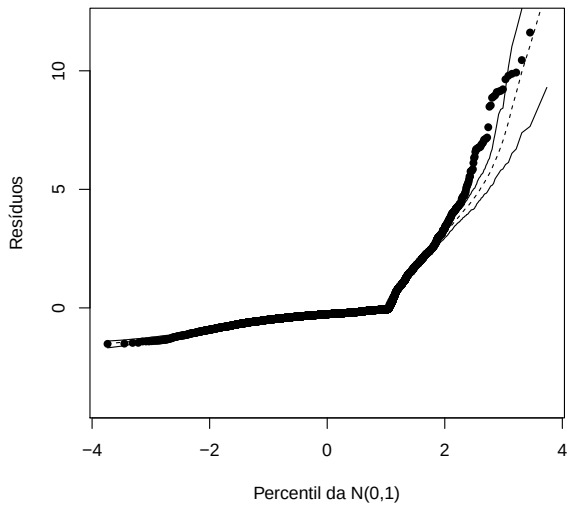


Figura 4.9: *Envelope plot* dos resíduos de Pearson do modelo Poisson

Na Tabela 4.17, são apresentados os valores estimados por este modelo, assim como a diferença entre estes valores e os valores observados. Verifica-se que cerca 30% dos valores estimados pelo modelo são diferentes dos observados.

Tabela 4.17: Valores estimados pelo modelo de regressão de Poisson.

MesessemPagar	0	1	2	3	4	5	6	7	8	9	10	11	12
Observado	3779	669	248	199	120	130	59	70	42	21	11	10	8
Estimado	3141	1197	494	233	121	68	41	25	16	11	7	5	3
Diferença	-638	528	246	34	1	-62	-18	-45	-26	-10	-4	-5	-5

Comparando as frequências estimadas pelo modelo de Poisson com as frequências observadas, constata-se que o modelo de Poisson é claramente desajustado, já que se observa subestimação do número de incumprimentos igual a zero meses e maior que cinco meses. Enquanto se verifica sobrestimação do número de incumprimentos entre 1 e 4 meses. Repare-se ainda, que o valor ajustado pelo modelo para o número de incumprimentos igual a zero difere 17% dos casos.

Na Figura 4.10, são apresentados os residuos de Pearson, assim como os resíduos *Deviance* do modelo de regressão de Poisson.

Relativamente à percentagem de resíduos de Pearson entre -2 e 2, o modelo de regressão de Poisson apresenta cerca de 90% de resíduos dentro do intervalo, valor muito semelhante ao apresentado pelos resíduos *deviance* com cerca de 91% dos resíduos no intervalo.

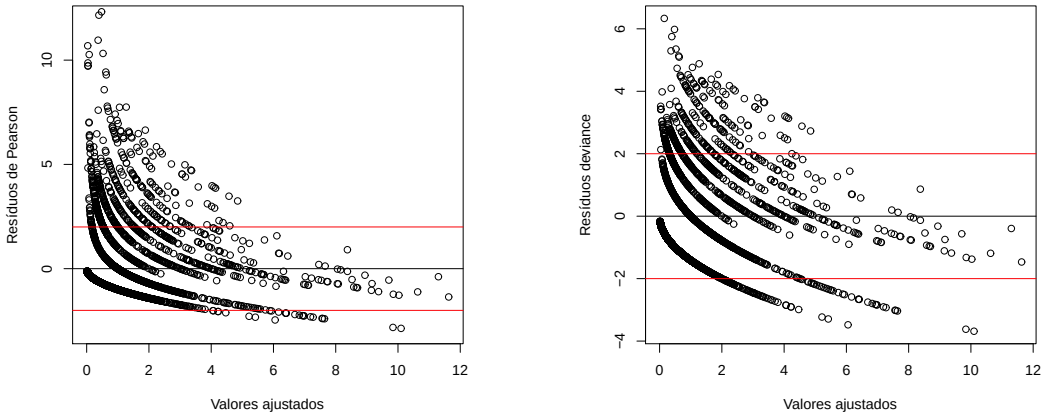


Figura 4.10: Gráficos dos resíduos do modelo de regressão de Poisson

Visto que o modelo de regressão de Poisson apresenta sobredispersão, decidiu-se ajustar aos dados um modelo de regressão Binomial Negativa com o objetivo de modelar a sobredispersão presente.

- Modelo de regressão Binomial Negativa

Começou-se por ajustar o modelo inicial, com as mesmas variáveis explicativas do modelo de Poisson, retirando-se do modelo todas variáveis não significativas, utilizando o método passo a passo. O modelo finalmente selecionado foi o modelo,

Modelo Final:

log(MesessemPagar)= β_0 + β_1* Idade + β_2* SldMdSem.cat + β_3* Ordenado + β_4* Regiao + β_5 * IdadeContrato + β_6* NAnosCliente + β_7 * NMesesLC

Na Tabela 4.18, apresentam-se os valores das estatísticas de ajustamento desses modelos.

Com aplicação do teste de razão de verosimilhanças, concluiu-se que não existem diferenças entre os modelos na qualidade do ajustamento ($p\text{-}valor = 0.4237$). As estatísticas de ajustamento indicam que o modelo que melhor se ajusta ao dados é o modelo

Tabela 4.18: Estatísticas de ajustamento dos modelos de regressão Binomial Negativos

Estatísticas	Modelo Inicial	Modelo Final
% Dev.Exp.	41.51	41.25
AIC	10625.44	10610.83
BIC	10823.08	10709.65
ℓ	-5282.72	-5290.42
X^2	6775.34	6851.39
Parâmetros Estimados	30	15

Final com o menor número de variáveis explicativas, podendo os seus coeficientes ser visualizados na Tabela 4.19.

Tabela 4.19: Modelo de regressão Binomial Negativa

Variável	Coeficiente	Erro Padrão	p-valor
constante	-0.0820	0.2216	0.7113
Idade	-0.0065	0.0025	0.0096
SldMdSem.cat2	-0.8600	0.0783	$<$ 2e-16
SldMdSem.cat3	-1.3262	0.0897	$<$ 2e-16
SldMdSem.cat4	-1.9983	0.0718	$<$ 2e-16
OrdenadoSim	-1.2255	0.0677	$<$ 2e-16
RegiaoAlentejoAlgarve	0.8271	0.1135	3.13e-13
RegiaoCentro	0.5485	0.1008	5.20e-08
RegiaoLisboaValeTejo	0.5780	0.0854	1.29e-11
RegiaoMadeira	0.7900	0.0882	$<$ 2e-16
RegiaoNorte	0.6327	0.0820	1.19e-14
IdadeContrato	0.0884	0.0035	$<$ 2e-16
NAnosCliente	-0.0345	0.0064	8.19e-08
NMesesLC	-0.0094	0.0026	0.0002

Os valores apresentados na Tabela 4.19, para um nível de significância de 5%, indicam-nos que todas as variáveis são estatísticamente significativas no modelo. Comparando este modelo com o modelo de Poisson, deixaram de ser significativas as variáveis *Prestacao-Mensal*, *Profissao*, *EstadoCivil* e *Habilitacoes*, sugerindo que estas variáveis eram as que provocavam sobredispersão no modelo.

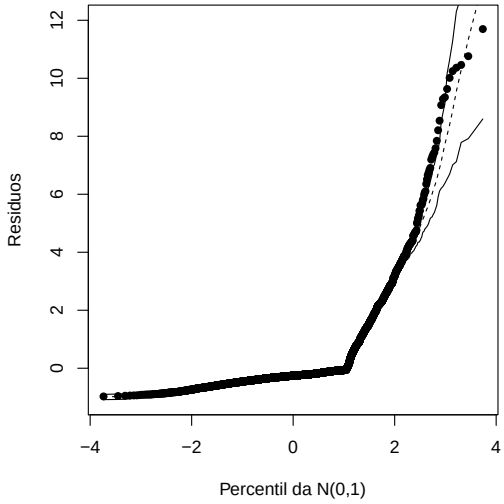


Figura 4.11: *Envelope plot* dos resíduos de Pearson do modelo Binomial Negativa

Analisando a Figura 4.11, no *envelope plot* dos resíduos de Pearson é evidente que o modelo Binomial Negativa efectua um melhor ajuste, já que uma grande parte dos resíduos se encontra no intervalo.

Na Tabela 4.20, são apresentados os valores estimados por este modelo, assim como a diferença entre estes valores e os valores observados. Verifica-se que apenas 9% dos valores estimados pelo modelo são diferentes dos observados, e que a diferença do valor observado com o valor previsto pelo modelo para o número de incumprimentos igual a zero é de apenas 1.80%.

Tabela 4.20: Valores estimados pelo modelo de regressão Binomial Negativa.

MesessemPagar	0	1	2	3	4	5	6	7	8	9	10	11	12
Observado	3779	669	248	199	120	130	59	70	42	21	11	10	8
Estimado	3711	806	322	166	98	64	44	31	23	18	14	11	9
Diferença	-68	137	74	-33	-22	-66	-15	-39	-19	-3	3	1	1

Na Figura 4.12, são apresentados os residuos de Pearson e os resíduos *Deviance* do modelo de Binomial Negativa.

Relativamente à percentagem de resíduos de Pearson, o modelo de regressão Binomial Negativa apresenta cerca de 94% de resíduos dentro do intervalo -2 e 2, enquanto que os resíduos *deviance* aumentam para cerca de 98% dentro do intervalo. Repare-se ainda que, quer no modelo Binomial Negativa quer no modelo de Poisson existe uma certa tendência para os valores dos resíduos fora deste intervalo serem positivos.

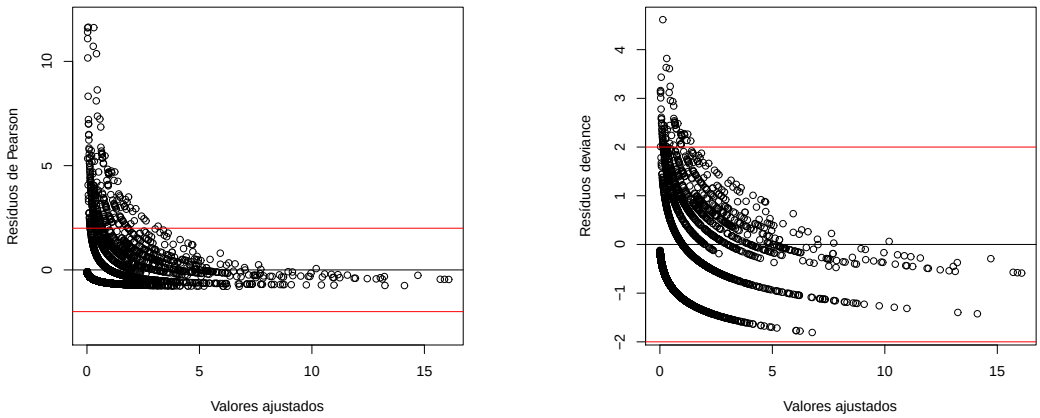


Figura 4.12: Gráficos dos resíduos do modelo de regressão Binomial Negativa.

Para comparar a qualidade do ajustamento dos dois modelos, do modelo de Poisson e do modelo Binomial Negativa, efectuou-se o teste de razão de verosimilhanças, como sugerido em [Zuur et al, 2009] na página 238. Os resultados podem ser visualizados na Tabela 4.21, indicando que o modelo Binomial Negativa é preferível ao modelo de Poisson.

Tabela 4.21: Teste de razão de verosimilhanças entre o modelo de regressão de Poisson e o modelo de regressão Binomial Negativa.

Modelo	Gl	ℓ	Estatística	p-valor
Binomial Negativa	15	-5290.4		
Poisson	23	-6150.2	1719.5	$<$2.2e-16

De seguida, consideramos os modelos com excesso de zeros, uma vez que os dados apresentam um grande número de zeros.

- Modelo de regressão de Poisson de zeros inflacionados

Começou-se por ajustar o modelo inicial, com as mesmas variáveis explicativas do modelo de Poisson e do modelo binomial negativa, na parte do modelo de contagem e na parte do modelo de zeros.

Modelo Inicial

Modelo de Contagem: log(MesessemPagar)= β_0 + β_1 * Sexo + β_2* EstadoCivil + β_3* PrestacaoMensal + β_4 * Profissao + β_5* Idade + β_6* NMesesLC + β_7* Habilitacoes + β_8* NAnosCliente + β_9* Regiao + β_{10} * Ordenado + β_{11}* IdadeContrato +

β_{12} * SldMdSem.cat

Modelo de zeros: logit(MesessemPagar)= γ_0 + γ_1 * Sexo + γ_2 * EstadoCivil + γ_3* PrestacaoMensal + γ_4 * Profissao + γ_5* Idade + γ_6 * NMesesLC + γ_7* Habilitacoes + γ_8* NAnosCliente + γ_9 * Regiao + γ_{10} * Ordenado + γ_{11} * IdadeContrato + γ_{12} * SldMdSem.cat

de seguida retirou-se do modelo as variáveis estatísticamente não significativas, sucessivamente, até se obter o modelo final,

Modelo Final

Modelo de Contagem: log(MesessemPagar)= β_0 + β_1 * SldMdSem.cat + β_2* EstadoCivil + β_3* Ordenado + β_4 * IdadeContrato + β_5 * NAnosCliente + β_6 * NMesesLC

Modelo de zeros: logit(MesessemPagar)= γ_0 + γ_1 * SldMdSem.cat + γ_2 * EstadoCivil + γ_3* Ordenado + γ_4* Regiao + γ_5* IdadeContrato + γ_6 * NMesesLC

Na Tabela 4.22, apresentam-se os valores das estatísticas de ajustamento desses modelos.

Tabela 4.22: Estatísticas de ajustamento dos modelos ZIP

Estatísticas	Modelo Inicial	Modelo Final
AIC	10106.12	10100.02
BIC	10488.22	10284.48
ℓ	-4995.06	-5022.01
X^2	7440.87	7542.09
Parâmetros Estimados	58	28

Aplicou-se o teste de razão de verosimilhanças e concluiu-se que existem diferenças entre os modelos na qualidade do ajustamento ($p\text{-}valor = 0.0047$). Assim, pelo resultado do teste de razão de verosimilhanças e também pelas estatísticas de ajustamento, o modelo escolhido foi o modelo Final, podendo os seus coeficientes ser visualizados na Tabela 4.23.

Os valores apresentados na Tabela 4.23, para um nível de significância de 5%, confirmam que todas as variáveis são estatísticamente significativas.

Na Tabela 4.24, são apresentados os valores estimados por este modelo, assim como a diferença entre estes valores e os valores observados. Verifica-se que 13% dos valo-

Tabela 4.23: Modelo de regressão de Poisson de zeros inflacionados

Modelo de Contagem	Coeficiente	Erro Padrão	p-valor
constante	1.0123	0.1519	2.65e-11
SldMdSem.cat2	-0.4127	0.0437	$<$ 2e-16
SldMdSem.cat3	-0.7315	0.0604	$<$ 2e-16
SldMdSem.cat4	-1.0958	0.0535	$<$ 2e-16
EstadoCivilDesconhecido	0.0070	0.0483	0.8849
EstadoCivilDivorciado	-0.1810	0.0613	0.0032
EstadoCivilSolteiro	-0.0387	0.0401	0.3337
EstadoCivilViuvo	-0.0806	0.1358	0.5526
OrdenadoSim	-0.8223	0.0485	$<$ 2e-16
IdadeContrato	0.0157	0.0023	1.51e-11
NAnosCliente	-0.0194	0.0038	2.70e-07
NMesesLC	0.0060	0.0019	0.0019
Modelo de Zeros	Coeficiente	Erro Padrão	p-valor
constante	1.9805	0.3400	5.68e-09
SldMdSem.cat2	0.9423	0.1341	2.12e-12
SldMdSem.cat3	1.2404	0.1518	3.11e-16
SldMdSem.cat4	1.6507	0.1223	$<$ 2e-16
EstadoCivilDesconhecido	0.0371	0.1418	0.7934
EstadoCivilDivorciado	-0.2126	0.1601	0.1842
EstadoCivilSolteiro	-0.2663	0.1079	0.0135
EstadoCivilViuvo	-0.1231	0.3552	0.7290
OrdenadoSim	0.8397	0.1115	5.06e-14
RegiaoAlentejoAlgarve	-1.1390	0.1857	8.49e-10
RegiaoCentro	-1.2645	0.1624	6.91e-15
RegiaoLisboaValeTejo	-1.1257	0.1343	$<$ 2e-16
RegiaoMadeira	-1.2705	0.1510	$<$ 2e-16
RegiaoNorte	-1.3371	0.1311	$<$ 2e-16
IdadeContrato	-0.1602	0.0062	$<$ 2e-16
NMesesLC	0.0217	0.0043	4.44e-07

res estimados pelo modelo são diferentes dos observados, valor bastante inferior quando comparado com o modelo de regressão de Poisson. Verifica-se ainda que o valor estimado do número de incumprimentos igual a zero apenas difere em 1.2% dos casos.

Tabela 4.24: Valores estimados pelo modelo ZIP.

MesessemPagar	0	1	2	3	4	5	6	7	8	9	10	11	12
Observado	3779	669	248	199	120	130	59	70	42	21	11	10	8
Estimado	3824	398	382	288	193	122	74	42	22	11	5	2	1
Diferença	45	-271	134	89	73	-8	15	-28	-20	-10	-6	-8	-7

No *envelope plot*, apresentado na Figura 4.13, é possível verificar-se que todos os resíduos pertencem ao intervalo de confiança revelando um bom ajustamento deste mo-

delo.

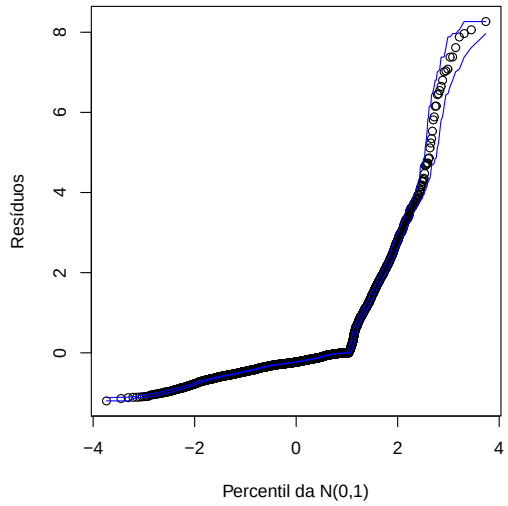


Figura 4.13: *Envelope plot* dos resíduos de Pearson do modelo ZIP

Na Figura 4.14 são representados os resíduos de Pearson do modelo ZIP. A percentagem de resíduos de Pearson dentro do intervalo -2 e 2, é de 94%.

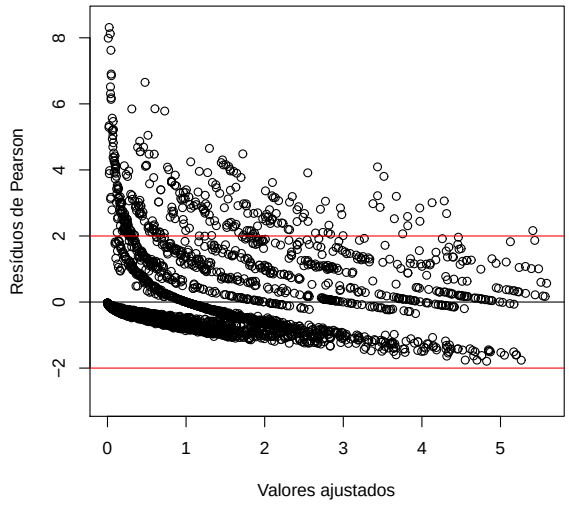


Figura 4.14: Gráfico dos resíduos de Pearson do modelo ZIP

Realizou-se também o teste de Vuong, cujos resultados se apresentam na Tabela 4.25, entre este modelo e o modelo de regressão de Poisson, concluíndo-se que o modelo de Poisson de zeros inflacionados é preferível ao modelo de regressão de Poisson, como seria

de esperar.

Tabela 4.25: Teste Vuong entre modelo de regressão de Poisson e o modelo ZIP.

Modelo	ZIP
Poisson	Estatística = 18.1836
	p-valor=< 2e-16
	modelo ZIP preferível ao modelo Poisson

- Modelo de regressão Binomial Negativa de zeros inflacionados

À semelhança dos modelos anteriores começou-se por ajustar o modelo inicial, com as mesmas variáveis explicativas, retirando-se sucessivamente do modelo as variáveis estatísticamente não significativas, até se obter o seguinte modelo,

Modelo Final

Modelo de Contagem: log(MesessemPagar)= β_0 + β_1 * SldMdSem.cat + β_2 * Ordenado + β_3 * Regiao + β_4 * IdadeContrato + β_5 * NAnosCliente + $\beta 6$ * NMesesLC

Modelo de zeros: logit(MesessemPagar)= γ_0 + γ_1 * SldMdSem.cat + γ_2 * Ordenado + γ_3 * Regiao + γ_4 * IdadeContrato + γ_5 * NMesesLC

Na Tabela 4.26 são apresentados os valores das estatísticas de ajustamento dos modelos.

Tabela 4.26: Estatísticas de ajustamento dos modelos ZINB

Estatísticas	Modelo Inicial	Modelo Final
AIC	9919.43	9900.34
BIC	10308.11	10097.98
ℓ	4900.72	-4920.17
X^2	6530.14	6498.42
Parâmetros Estimados	59	26

Pelo teste de razão de verosimilhanças concluiu-se que não existem diferenças entre os modelos relativamente à qualidade do ajustamento ($p\text{-}valor = 0.1034$). Analisando as estatísticas da Tabela 4.26 escolheu-se o modelo Final com o menor número de variáveis explicativas, como o modelo que melhor se ajusta aos dados. Na Tabela 4.27 apresentam-se os coeficientes do modelo selecionado. Repare-se que o modelo de zeros utiliza menos variáveis explicativas do que o modelo de contagem.

Tabela 4.27: Modelo de regressão Binomial Negativa de zeros inflacionados

Modelo de Contagem	Coeficiente	Erro Padrão	p-valor
constante	0.8132	0.2126	0.0001
SldMdSem.cat2	-0.4347	0.0610	1.01e-12
SldMdSem.cat3	-0.7852	0.0790	$<$ 2e-16
SldMdSem.cat4	-1.1505	0.0670	$<$ 2e-16
OrdenadoSim	-0.8530	0.0608	$<$ 2e-16
RegiaoAlentejoAlgarve	0.2125	0.0936	0.0232
RegiaoCentro	-0.0642	0.0820	0.4335
RegiaoLisboaValeTejo	-0.0008	0.0704	0.9913
RegiaoMadeira	0.0846	0.0767	0.2700
RegiaoNorte	0.0105	0.0671	0.8756
IdadeContrato	0.0203	0.0032	3.56e-10
NAnosCliente	-0.0215	0.0052	3.09e-05
NMesesLC	0.0063	0.0027	0.0191
Log (theta)	1.3276	0.1131	$<$ 2e-16
Modelo de Zeros	Coeficiente	Erro Padrão	p-valor
constante	1.5690	0.3954	7.25e-05
SldMdSem.cat2	0.9463	0.1492	2.27e-10
SldMdSem.cat3	1.2123	0.1721	1.85e-12
SldMdSem.cat4	1.6595	0.1361	$<$ 2e-16
OrdenadoSim	0.7819	0.1248	3.78e-10
RegiaoAlentejoAlgarve	-1.1091	0.2093	1.16e-06
RegiaoCentro	-1.4172	0.1996	1.24e-11
RegiaoLisboaValeTejo	-1.2179	0.1689	5.49e-11
RegiaoMadeira	-1.3094	0.1710	1.92e-14
RegiaoNorte	-1.4124	0.1574	$<$ 2e-16
IdadeContrato	-0.1595	0.0068	$<$ 2e-16
NMesesLC	0.0242	0.0050	1.07e-06

Os valores estimados pelo modelo ZINB, são apresentados na Tabela 4.28, verificando-se que aproximadamente 10% dos valores estimados diferem dos valores observados, e que o valor estimado do número de incumprimentos igual a zero difere apenas 1.80% dos valores observados.

Tabela 4.28: Valores estimados pelo modelo ZINB.

MesessemPagar	0	1	2	3	4	5	6	7	8	9	10	11	12
Observado	3779	669	248	199	120	130	59	70	42	21	11	10	8
Estimado	3848	465	360	246	160	102	65	42	27	18	11	7	5
Diferença	69	-204	112	47	40	-28	6	-28	-15	-3	0	-3	-3

O *envelope plot* da Figura 4.15 sugere-nos um bom ajustamento do modelo, já que os resíduos de Pearson, se encontram dentro de intervalo.

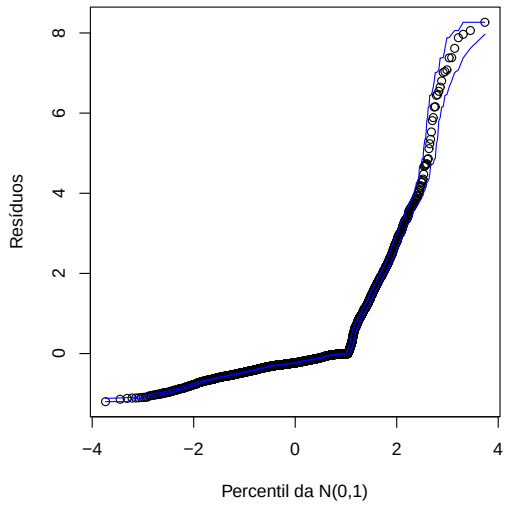


Figura 4.15: *Envelope plot* dos resíduos de Pearson do modelo ZINB

Na Figura 4.16 são representados os resíduos de Pearson do modelo ZINB. Podendo observar-se que a maior parte dos resíduos de Pearson, aproximadamente de 94%, se encontram entre os valores -2 e 2.

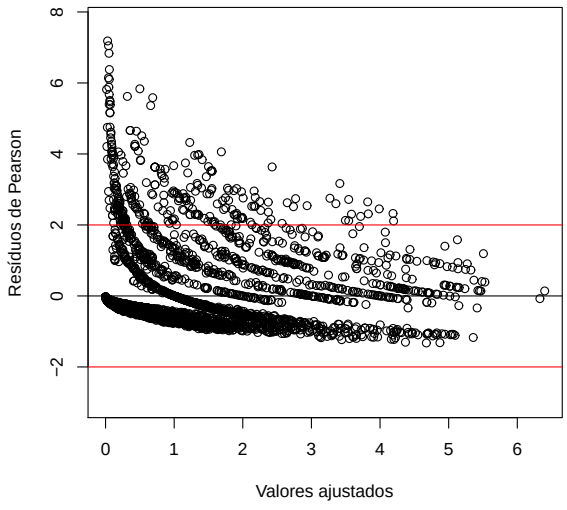


Figura 4.16: Gráfico dos resíduos de Pearson do modelo ZINB

O teste de Vuong foi realizado para se comparar os modelo ZINB e Binomial Negativo, o resultado é apresentado na Tabela 4.29, concluindo-se que o modelo de Binomial Negativo de zeros inflacionados é preferível ao modelo Binomial Negativo.

Tabela 4.29: Teste Vuong entre modelo Binomial Negativa e o modelo ZINB.

Modelo	ZINB
Binomial Negativa	Estatística = 16.3695
	p-valor=$<$ 2e-16
	modelo ZINB preferível ao modelo Binomial Negativa

Na Tabela 4.30 encontram-se resumidas as estatísticas de ajustamento dos quatro diferentes modelos anteriormente selecionados.

A análise dos indicadores de qualidade de ajustamento indicam como sendo os modelos que melhor se ajustam aos dados, os modelos de regressão de zeros inflacionados, os menores valores de AIC (ZIP:10100.02 e ZINB:9900.34) e BIC (ZIP:10284.48 e ZINB:10097.98) e os maiores valores da função log-verosimilhança (ZIP:-5022.01 e ZINB:-4920.17) correspondem aos modelos ZIP e ZINB, respectivamente.

Tabela 4.30: Estatísticas de ajustamento dos modelos escolhidos

Estatísticas	Modelo Poisson	Modelo Binomial Negativa	Modelo ZIP	Modelo ZINB
AIC	12346.33	10610.83	10100.02	9900.34
BIC	12497.85	10709.65	10284.48	10097.98
ℓ	-6150.17	-5290.42	-5022.01	-4920.17
X^2	11662.49	6851.39	7542.09	6498.42
Parâmetros Estimados	23	15	28	26

A Tabela 4.31, com os resultados do teste Vuong entre os modelos, para se perceber se existe diferenças significativas relativamente ao ajustamento, permite-nos reforçar a conclusão anterior, de que os modelos de regressão de zeros inflacionados efectuam um melhor ajustamento. Em ambos os testes a hipótese nula foi rejeitada, optando-se pela hipótese alternativa de que um dos modelos tem um melhor ajustamento aos dados. O modelo selecionado como modelo que melhor se ajusta, foi o modelo ZINB.

Tabela 4.31: Teste Vuong entre os modelos de regressão.

Teste Vuong	B.Negativa vs Poisson	ZINB vs ZIP
Estatística	14.8557	6.5998
p-valor	$<$ 2e-16	2.06e-011
Ajustamento preferível	Binomial Negativa	ZINB

A variável *Regiao* no modelo de contagem, apenas a categoria 'Alentejo e Algarve' é significativa.

O modelo fica então definido por,

Modelo de Contagem: log(MesessemPagar)= 0.8132 - 0.4347 * *SldMdSem.cat2* - 0.7852 * *SldMdSem.cat3* - 1.1505 * *SldMdSem.cat4* - 0.8530 * *OrdenadoSim* + 0.2125 * *RegiaoAlentejoAlgarve* - 0.0642 * *RegiaoCentro* - 0.0008 * *RegiaoLisboaValeTejo* + 0.0846 * *RegiaoMadeira* + 0.0105 * *RegiaoNorte* + 0.0203 * *IdadeContrato* - 0.0215 * *NAnosCliente* + 0.0063 * *NMesesLC*

Modelo de zeros: logit(MesessemPagar)= 1.5690 + 0.9463 * *SldMdSem.cat.2* + 1.2123 * *SldMdSem.cat.3* + 1.6595 * *SldMdSem.cat.4* + 0.7819 * *OrdenadoSim* -1.1091* *RegiaoAlentejoAlgarve* - 1.4172 * *RegiaoCentro* - 1.2179 * *RegiaoLisboaValeTejo* - 1.3094 * *RegiaoMadeira* - 1.4124 * *RegiaoNorte* - 0.1595 * *IdadeContrato* + 0.0242 * *NMesesLC*

Para se perceber qual o efeito das variáveis no número esperado de *MesessemPagar*, começamos por analisar as variáveis que apresentam coeficientes com sinal positivo no modelo de contagem. Se um cliente for da região 'Alentejo e Algarve' o número esperado de meses de incumprimento aumenta cerca de 24% em relação aos clientes dos 'Açores', sendo esta variável a que apresenta um maior efeito de crescimento no número consecutivo de meses sem pagamento da prestação. Por sua vez, quanto mais elevada for a idade do contrato maior será também o número de incumprimentos, já que esta variável sofre um aumento de 2% por cada mês que se aumente na idade do contrato. De igual modo o número esperado de incumprimentos aumenta cerca de 1% por cada mês que se aumente o número de meses até liquidação do montante contratado.

Quanto às variáveis que apresentam coeficientes com sinal negativo, concluí-se que um cliente que receba o ordenado através do banco diminui cerca de 57% o número esperado de incumprimentos em relação aos que não recebem. Um cliente que pertença à categoria *SldMdSem.cat4*, isto é, um cliente com um saldo médio semestral nunca inferior a 325 euros, diminui o número esperado de não pagamentos cerca de 68% em relação aos que pertencem à categoria *SldMdSem.cat1* que têm um saldo médo semestral que nunca ultrapassa os 11.5 euros. Repare-se que à medida que aumenta a categoria do saldo médio semestral, aumenta também a probabilidade de o número de incumprimentos ser zero, o que faz todo o sentido, porque quanto maior for o saldo médio semestral do cliente menor é o risco de este falhar no pagamento da mensalidade ao banco. O número esperado de incumprimentos decresce também cerca de 2% por cada ano que se aumente ao número de anos que é cliente do banco.

No modelo de zeros, à medida que aumenta o saldo médio semestral, aumenta também

a probabilidade de ocorrer um zero no valor esperado no número de meses de incumprimento. Já se o cliente receber o ordenado pelo banco, a probabilidade de pagar a mensalidade também aumenta, o que significa que o valor esperado de incumprimentos diminui. A medida que aumenta também o número de meses desde a data do contrato diminui a probabilidade de ocorrer um zero no valor esperado do número de meses de incumprimento.

O parâmetro de dispersão do modelo $Log(theta) = 1.3276$ é significativamente diferente de zero, sugerindo que existe sobredispersão nas contagens, confirmando mais uma vez que o modelo binomial negativa é mais apropriado do que o modelo de Poisson para modelar os dados.

Estes modelos permitem ainda, determinar a probabilidade do cliente do banco cumprir com o pagamento da mensalidade do crédito. Por exemplo, um cliente que resida na região do Alentejo ou do Algarve, que não receba o ordenado através do banco, que o seu saldo médio semestral não ultrapasse os 11.5 euros, com contrato celebrado há 24 meses, faltando 64 meses para liquidar o empréstimo e sendo cliente da instituição bancária à 6 anos, a probabilidade de este cliente cumprir com pagamento de todas as mensalidades é de 0.164, o que significa que o mais provável é este cliente não cumprir com os compromissos assumidos com o banco.

Por outro lado, um cliente da região da Madeira, que também não receba o ordenado pelo banco, mas que tenha um saldo médio semestral superior a 325 euros, que seja cliente do banco à 16 anos, com contrato celebrado há 7 meses, em que falte 54 meses para liquidação do contrato, a probabilidade deste cliente já é de 0.940, o que significa que este cliente tem uma grande probabilidade de cumprir com o que foi acordado com o banco.

Capítulo 5

Conclusões e trabalho futuro

Neste trabalho foram estudados modelos para dados de contagem, o modelo de regressão de Poisson e o modelo de regressão binomial negativa, o modelo de regressão de Poisson de zeros inflacionados e o modelo de regressão binomial negativa de zeros inflacionados. Foi ainda aplicada a metodologia desses modelos a dados sobre empréstimos bancários, com o objetivo de se explicar o número de não pagamentos da prestação do crédito de um cliente em função das caracteristicas do cliente e do contrato.

O modelo inicialmente ajustado, o modelo de regressão de Poisson revelou-se inapropriado, uma vez que apresentava sobredispersão. Neste modelo cerca de 30% dos valores ajustados são diferentes dos observados, e o valor ajustado pelo modelo para o número de não incumprimentos difere em 17% dos casos.

Ajustou-se então o modelo de regressão binomial negativa, verificando-se que este modelo é preferível ao modelo de regressão de Poisson, não só porque se corrigiu a sobredispersão do modelo anterior, mas também porque se conseguiu um modelo mais parcimonioso. Para este modelo, a diferença entre os valores ajustados e os valores observados é bastante menor, com apenas 9% de diferenças. Também o valor ajustado pelo modelo para o número de incumprimentos igual a zero descresce para 2%, relativamente ao modelo de regressão de Poisson.

Como os dados apresentam um grande número de zeros na variável dependente, ou seja 70% dos clientes não têm incumprimentos da sua prestação mensal do crédito pessoal, ajustou-se modelos de regressão de zeros inflacionados. Tendo-se concluído que os modelos de regressão de zeros inflacionados apresentam um melhor ajustamento, quando comparados com os modelos que não têm em consideração o excesso de zeros.

Para o modelo de regressão de Poisson de zeros inflacionados, os valores ajustados diferem 13% dos observados e o valor ajustado para o número de incumprimentos igual a zero é aproximadamente igual a 1%, revelando uma melhoria significativa em comparação

com o modelo de regressão de Poisson.

Por sua vez, o modelo de regressão binomial negativa de zeros inflacionados é o que melhor se ajusta aos dados, apresentando apenas 10% de diferenças entre os valores ajustados e os valores observados. O valor estimado para o número de não incumprimentos é de aproximadamente 2% dos casos.

Os modelos baseados na distribuição Binomial Negativa, revelaram-se mais adequados à modelação dos dados, comparando com os modelos baseados na distribuição de Poisson.

Segundo este modelo, o número esperado de incumprimentos aumenta com o aumento da idade do contrato e com o aumento do número de meses que ainda faltam para liquidar o empréstimo. Se o cliente for da região 'Alentejo e Algarve' o número esperado de meses de incumprimento aumenta em relação aos clientes dos 'Açores'. Enquanto que, se o cliente for um cliente antigo, se receber o ordenado pelo banco e se o seu saldo médio semestral for positivo, o número esperado de incumprimentos ao banco, diminui.

Para trabalho futuro, pretendemos estudar e aplicar os modelos de barreira a esta base de dados. Gostaríamos ainda alargar este estudo a modelos de regressão inflacionados de zeros com efeitos aleatórios e aplicá-los a dados sobre empréstimos bancários.

Capítulo 6

Anexos

Tabela 6.1: Modelos de Regressão de Poisson com apenas uma variável explicativa.

	Variável	Coeficiente	Erro Padrão	*p-valor*
Modelo 1	Constante	-2.15e-01	2.61e-06	$<$ 2e-16
	MontanteContratado	5.08e-06	1.92e-06	0.0083
Modelo 2	Constante	-5.45e-02	2.59e-02	0.0351
	CapitalVincendo	-1.17e-05	2.43e-06	1.55e-06
Modelo 3	Constante	-0.2177	0.0273	1.34e-15
	PrestacaoMensal	0.0003	0.0001	0.0094
Modelo 4	Constante	-0.0715	0.0561	0.2030
	Idade	-0.0021	0.0013	0.1070
Modelo 5	Constante	0.3362	0.0195	$<$ 2e-16
	SldMdSem.cat2	-0.1503	0.0395	0.0001
	SldMdSem.cat3	-0.6917	0.0523	$<$ 2e-16
	SldMdSem.cat4	-1.5132	0.0433	$<$ 2e-16
Modelo 6	Constante	-0.1969	0.0260	3.29e-14
	SexoM	0.0566	0.0316	0.0731
Modelo 7	Constante	-0.2009	0.0210	$<$ 2e-16
	EstadoCivilDesconhecido	0.2393	0.0434	3.63e-08
	EstadoCivilDivorciado	-0.0681	0.0547	0.2130
	EstadoCivilSolteiro	0.0716	0.0362	0.0477
	EstadoCivilViuvo	-0.1543	0.1231	0.2099
Modelo 8	Constante	-0.1266	0.0264	1.59e-06
	HabilitacoesEnsinoSup.	-0.5572	0.0831	2.01e-11
	HabilitacoesEscolaridadeObrig.	-0.0811	0.0361	0.0247
	HabilitacoesFormacaoProf.	-0.1077	0.0874	0.2181
	HabilitacoesDesconhecidas	0.1107	0.0392	0.0048
Modelo 9	Constante	-0.0027	0.0261	0.9171
	ProfissaoEmp.EscritComerc.Serv.	-0.3226	0.0398	5.01e-16
	ProfissaoEstudanteDomestica	-0.3256	0.0766	2.11e-05
	ProfissaoOutros	-0.1276	0.0376	2.35e-07
	ProfissaoPeqMedEmpresario	0.1040	0.0573	0.0695
	ProfissaoQuadroMedio	-0.353961	0.1361	0.0093
Modelo 10	Constante	0.0834	0.0163	3.02e-07
	OrdenadoSim	-0.9429	0.0389	$<$ 2e-16
Modelo 11	Constante	-0.6144	0.0308	$<$ 2e-16
	RegiaoAlentejoAlgarve	0.7291	0.0605	$<$ 2e-16
	RegiaoCentro	0.6883	0.0541	$<$ 2e-16
	RegiaoLisboaValeTejo	0.5872	0.0453	$<$ 2e-16
	RegiaoMadeira	0.5468	0.0500	$<$ 2e-16
	RegiaoNorte	0.7111	0.0431	$<$ 2e-16
Modelo 12	Constante	-1.7569	0.0412	$<$ 2e-16
	IdadeContrato	0.0824	0.0017	$<$ 2e-16
Modelo 13	Constante	0.2045	0.0317	1.08e-10
	NAnosCliente	-0.0375	0.0030	$<$ 2e-16
Modelo 14	Constante	1.8174	0.0688	$<$ 2e-16
	NMesesLC	-0.0328	0.0012	$<$ 2e-16

Bibliografia

[Agresti, 2007] Agresti A., An Introduction to Categorical Data Analysis, second edition. John Wiley & Sons, Inc, New Jersey.

[Akaike, 1974] Akaike H., A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716-723.

[Böhning, 1999] Böhning D., Dietz E., Schlattmann P., Mendonça L., Kirchner U., The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of Royal Statistical Society. Series A, 162(2), 195-209.

[Bozdogan, 1987] Bozdogan H., Model selection and Akaike's information criterion, the general theory and its analytical extensions. Psychometrica 52, 345-370.

[Broek, 1995] Broek J.V., A score test for zero inflation in a Poisson distribution. Biometrics, 51, 738-743.

[Cameron and Trivedi, 1998] Cameron A.C., Trivedi P.K., Regression Analysis of Count Data. Cambridge university Press, Cambridge, UK.

[Cheung, 2002] Cheung,Y. B., Zero-inflated models for regression analysis of count data: a study of growth and development. Statistics in Medicine 21, 1461-1469.

[Dempster et al, 1977] Dempster A.P., Laird N.M., Rubin D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, series B, 39, 1-38.

[Dobson, 2002] Dobson A.J., Introduction to Generalized Linear Models, Second Edition. Chapman & Hall /CRC Press.

[Fahrmeir and Tutz, 2001] Fahrmeir L., Tutz G., Multivariate Statistical Moddelling Based on Generalized Linear Models, second edition. Springer, New York.

[Famoye and Singh, 2006] Famoye F., Singh K.P., Zero-inflated generalized Poisson regression model with and application to domestic violence data. Journal of Data Science 4, 117-130.

[Garay et al, 2011] Garay A.M., Hashimoto E.M.M.,Ortega E.M., Lachos V.H., On estimation and influence diagnostics for zero-inflated negative binomial regression models. Computacional Statistics and Data Analysis, 55(3), 1304-1318.

[Hair et al, 1998] Hair J.F., Anderson R.E., Black W., Multivariate Data Analysis, 5th Edition. Prentice-Hall Inc., 167-168.

[Hall, 2000] Hall D., Zero-inflated Poisson and Binomial Regression with Random Effects: A case of Study. Biometrics, 56(4) 1030-1039.

[Hilbe, 2001] Hilbe J.M., Negative Binomial Regression, Second Edition, Cambridge University Press, New York.

[Kleinbaum and Klein, 2002] Kleinbaum D.G., Klein M., Logistic Regression: A Self-Learning Text, second edition. Springer-Verlag, New York.

[Lambert, 1992] Lambert D., Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.

[Lee et al, 2001] Lee A., Wang K., Yau K., Analysis of zero-inflated Poisson data incorporating extent of exposure. Biometrical Journal, 43, 963-975.

[Lewsey and Thomson, 2004] Lewsey J.D., Thomson, W.M., 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, 183-189.

[Lindsey, 1999] Lindsey J.K., On the use of corrections for overdispersion. Appl. Stat. 48(4), 553-561.

[Marôco, 2010] Marôco, João, Análise Estatística com o PASW Statistics. ReportNumber, Lda. Pêro Pinheiro.

[Martin et al, 1989] Martin T., Wintle B., Rhodes J., Kuhnert P., Field S., Low-Choy S., Tyre A., Possingham H., Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters 8, 1235-1246.

[McCullagh and Nelder, 1989] McCullagh P., Nelder J. A., Generalized Linear Models. Chapman & Hall.

[Morrison, 2004] Morrison J.S., Preparing for basel common problems, practical solutions: part 2: modelling strategies (capital requirements; accounting). The RMA Journal, 98-102.

[Nelder and Wedderburn, 1972] Nelder J. A., Wedderburn R. W. M., Generalized Linear Models. Journal of the Royal Statistical Society, series A, 135, 370-384.

[Potts and Elith, 2006] Potts J. M., Elith J., Comparing species abundance models. Ecological Modelling, 199, 153-163.

[Ridout et al, 2001] Ridout M., Hinde J., Demétrio C., A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives, Biometrics 57, 219-223.

[Rodrígues, 2007] Rodrígues G., Lecture notes on generalized linear models. Retrieved October 11, 2011 from http://data.princeton.edu/wws509/notes/.

[Schwarz, 1978] Schwarz G., Estimating the dimension of a model. The Annals of Statistics 6, 461-464.

[Turkman and Silva, 2000] Turkman M. A., Silva G. L., Modelos lineares Generalizados da Teoria à Prática. Edições SPE, Lisboa.

[Vuong, 1989] Vuong, Q.H., Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 307-334.

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

[Zuur et al, 2009] Zuur A.F., Ieno N.E., Walker N.J., Saveliev A.A., Smith G.M., Mixed effects models and extensions in ecology with R. Springer.