Universidade do Minho
Escola de Ciências

Domingos Gonçalves Paulo

Modelo de Regressão de Poisson Generalizado: Análise de dados de contagem com sobredispersão e subdispersão

Modelo de Regressão de Poisson Generalizado: Análise de dados de contagem com sobredispersão e subdispersão
Domingos Gonçalves Paulo
UMinho | 2019

julho de 2019

Universidade do Minho
Escola de Ciências

Domingos Gonçalves Paulo

Modelo de Regressão de Poisson Generalizado: Análise de dados de contagem com sobredispersão e subdispersão

Dissertação de Mestrado
Mestrado em Estatística

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

julho de 2019

Direitos de autor e condições de utilização do trabalho por terceiros

Agradecimentos

A agradeço a Deus por ter me capacitado em todos os momentos deste curso, por ter me dado forças, entendimento e sabedoria, sem Deus nada seria e nada é. A ele a honra o poder e a glória para sempre.

A minha esposa, mesmo distante mas soube desempenhar o seu papel nos momentos de dificuldades e pelo apoio e confiança depositado.

Aos meus familiares, que compreenderam minha ausência dando força, amor carinho e incentivos para que alcançasse mais está vitória.

A orientadora Professora Doutora Susana Margarida Ferreira de Sá Faria por não medir esforço para conseguir material bibliográfico para a criação e desenvolvimento deste trabalho e total atenção, à você professora o meu muito obrigado.

Aos demais professores e funcionários do curso que, de uma forma ou outra, contribuíram para esta conquista, a todos o meu profundo agradecimento.

Declaração de integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Modelo de regressão de Poisson Generalizado: Análise de dados de contagem com sobredispersão e subdispersão

Resumo

A distribuição de Poisson é uma referência para os modelos de regressão para dados de contagem. No entanto, a restrição de equidispersão, isto é, o valor esperado e a variância condicionais são iguais, não representa com precisão os dados reais.

De facto, existem muitas situações, onde os dados de contagem apresentam sobredispersão ou subdispersão e na modelação deste tipo de dados, a aplicação do modelo de regressão de Poisson revela-se inadequada, uma vez que pode originar a subestimação da variância dos parâmetros.

Alguns autores têm desenvolvido novos modelos de regressão para ultrapassar o problema de sobredispersão ou subdispersão. Um desses modelos é o modelo de regressão de Poisson Generalizado, proposto por Consul e Famoye (1992) para ajustar dados de contagem que apresentam sobredispersão ou subdispersão, e que será apresentado neste trabalho.

Neste trabalho, o modelo de regressão de Poisson, o modelo de regressão Binomial Negativa e o modelo de regressão de Poisson Generalizado são ajustados a dois conjuntos de dados.

O primeiro conjunto de dados, que apresenta subdispersão, pretende estudar número de ofertas públicas de aquisição de empresas, após a oferta inicial recebida, durante o período de 1978-1985. O segundo conjunto de dados, que apresenta sobredispersão, tem como o objetivo estudar o número de relatórios com avaliação negativa, de uma conta de crédito, de clientes que solicitaram um cartão de crédito.

Os resultados obtidos mostram que o modelo de regressão de Poisson Generalizado apresenta melhor ajustamento, quando comparado com o modelo de regressão de Poisson e o modelo de regressão Binomial Negativa.

Palavras-chave: Sobredispersão, Subdispersão, Modelo de regressão de Poisson, Modelo de regressão Binomial Negativa, Modelo de regressão de Poisson Generalizado.

Generalized Poisson regression model: Modeling count data with Overdispersion and underdispersion

Abstract

The Poisson distribution is a reference to the regression models for modeling counting data. However, the equidispersion restriction, that is, the mean and the conditional variance are equal, does not accurately represent the real data.

However, there are many situations where the modeling count data have overdispersion or underdispersion and in the modelling of this type of data, the application of the Poisson regression model is inadequate, since it can underestimate the variances of the parameters.

Some authors have developed new regression models to overcome the problem of overdispersion or underdispersion. One of these models is the generalized Poisson regression model, proposed by Consul and Famoye (1992) to adjust modeling count data that presents Overdispersion or underdispersion, which will be presented in this work.

In this work, the Poisson regression model, the Negative Binomial regression model and the generalized Poisson regression model are adjusted to two data sets.

The first data set, which has underdispersion, intends to study the number of takeover bids of firms that were targets during the period 1978–1985. The second data set, which presents Overdispersion, aim to study the number of major derogatory reports for a sample of applicants for a type of credit card.

The results obtained show that the generalized Poisson regression model has a better adjustment when compared to the Poisson regression model and the Negative Binomial regression model.

Keywords: Overdispersion, Underdispersion, Poissom regression, Negative Binomial regression, Generalized Poisson regression

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

DBN	Distribuição Binomial Negativo
DPG	Distribuição de Poisson Generalizada
DPGI	Distribuição de Poisson Generalizada I
DPGII	Distribuição de Poisson Generalizada II
VIF	*Variance Inflation Factor*
MLG	Modelo Linear Generalizado
MLGs	Modelos Lineares Generalizados
MMV	Método da Máxima Verosimilhança

Capítulo 1

Introdução

Em diversas áreas de estudo os modelos de regressão de dados de contagem tornaram-se uma das ferramentas da estatística mais utilizados para análise de dados, onde a variável dependente assume apenas valores inteiros não negativos observadas num determinado período de interesse. Esses dados são obtidos a partir da observação do número de ocorrências de um determinado evento por unidade de tempo ou espaço. Por exemplo, o número de pessoas atendidas numa loja num dia de trabalho, ou o número de internamentos por doenças cardíacas em diversos hospitais num determinado dia.

Diante dos problemas relatados na aplicação de modelos de dados de contagem, diferentes abordagens foram propostas. Destaca-se o trabalho apresentado por Nelder e Wedderburn (1972) que introduz a teoria dos Modelos Lineares Generalizados. Essa nova classe de modelos flexibilizou a distribuição condicional permitindo que outras distribuições pertencentes à família exponencial fossem consideradas para a distribuição da variável resposta. Tal família contempla as distribuições de Poisson, Binomial, Gama e entre outras. A distribuição de Poisson é a mais usada na análise de dados de contagem.

Na teoria da probabilidade, a distribuição de Poisson é conhecida por assumir que o valor esperado e a variância são iguais (equidispersão). Todavia, em casos reais a variância dos dados pode ser maior ou menor que o valor esperado, causando no modelo problemas de sobredispersão ou subdispersão respetivamente.

Na prática, em muitas situações pode surgir o problema da sobredispersão, que podem ocorrer por diversas razões tais como a heterogeneidade das unidades experimentais, ausência de covariáveis, diferentes amplitudes de domínio não considerados, correlação entre as observações, excesso de zero, entre outras (Hinde e Demétrio, 1998). uma situação que menos comum, mas que tem ganho a atenção nos estudos estatísticos, é a subdispersão. Os processos que reduzem a variabilidade da contagem, abaixo do estabelecido pela distribuição de Poisson, não são tão conhecidos.

Nesta dissertação pretende-se estudar o modelo de regressão de Poisson Generalizado, para modelar dados de contagens que apresentam sobredispersão ou subdispersão. Os dois conjuntos de dados utilizados neste trabalho foram extraídos na biblioteca *Countreg* e *AER* do *software* R.

No primeiro conjunto de dados, que apresenta subdispersão, será estudado uma

amostra de dados de empresas que foram alvo de ofertas públicas de aquisição durante o período 1978-1985. O segundo conjunto de dados, que apresenta sobredispersão será estudado uma amostra de dados de um histórico de clientes que solicitaram um cartão de crédito, tendo como o principal objetivo, estudar o número de relatórios com avaliação negativa.

1.1 Estrutura do Trabalho

O conteúdo deste trabalho esta organizado em 5 capítulos.

No capítulo 2, apresenta-se a base teórica dos Modelos Lineares Generalizados, os modelos de regressão para dados de contagem nomeadamente, os modelo de regressão de Poisson e o modelo de regressão Binomial Negativa.

O capítulo 3 apresenta-se o modelo de regressão de Poisson Generalizado, as principais parametrizações para obtenção do modelo de regressão de Poisson Generalizado I e o modelo de regressão de Poisson Generalizados II.

O capítulo 4 descreve-se as bases de dados, e apresentam-se os modelos de regressão ajustados.

No capítulo 5 será apresentada as principais conclusões deste trabalho e as indicações para o trabalho futuro.

Capítulo 2

Modelos Lineares Generalizados

Em muitos estudos estatísticos, somos confrontados com problemas em que o objectivo principal é o de estudar a relação entre variáveis, ou mais particularmente, analisar a influência que uma ou mais variáveis (explicativas), medidas em indivíduos ou objectos, têm sobre uma variável de interesse, a que damos o nome de variável resposta. O modo como, em geral, o estatístico aborda tal problema é através do estudo de um modelo de regressão que relacione essa variável de interesse com as variáveis explicativas (Turkman e Silva, 2000).

O modelo estatístico clássico para a análise de regressão, o modelo linear, surgiu no inicio do século XIX, com Legendre e Gauss (Stigler, 1977). Muito resumidamente, este modelo exprime o valor esperado da variável resposta como uma combinação linear das variáveis explicativas, e é aplicado na situação em que a variável resposta segue uma distribuição normal.

Neste capítulo abordam-se os Modelos Lineares Generalizados, em particular, os modelos de regressão para dados de contagens.

2.1 Família Exponencial

Os Modelos Lineares Generalizados sugerido por McCullagh e Nelder (1989), pressupõem que a variável aleatória resposta (Y), segue uma distribuição pertencente a uma família particular, a família exponencial, com função densidade de probabilidade dada por:

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

onde se assume que ϕ é um parâmetro de dispersão geralmente conhecido, θ é a forma canónica do parâmetro de localização e $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas. Para esta família de distribuições temos,

$$E[Y] = \mu = b'(\theta) \text{ e } Var[Y] = a(\phi)V(\mu) = a(\phi)b''(\theta),$$

onde $V(\mu) = b''(\theta)$ é designada por função de variância.

A função $a(\phi)$ depende apenas do parâmetro de dispersão ϕ e em muitas situações, observa-se que $a(\phi) = \frac{\phi}{w}$, onde w é uma constante conhecida. Admite-se

ainda que a função $b(\cdot)$ é diferenciável e que o suporte da distribuição não depende dos parâmetros.

Os MLG são uma extensão do modelo linear clássico,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \varepsilon, \tag{2.2}$$

onde X é uma matriz de dimensão $n \times (p+1)$ de especificação do modelo, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ é o vector dos parâmetros do modelo e ε é um vector de erros aleatórios com distribuição que se supõe $N_n(0, \sigma^2 I_n)$, onde I_n é a matriz identidade de ordem n, ou seja, o valor esperado da variavel resposta é uma função linear das covariáveis. Tem-se assim,

$$E(\boldsymbol{Y}|X) = \boldsymbol{\mu} \text{ com } \boldsymbol{\mu} = X\boldsymbol{\beta}.$$

A extensão mencionada é feita em duas direcções. Por um lado, a distribuição da variável resposta não tem de ser normal, podendo ser qualquer distribuição da família exponencial, por outro lado, embora se mantenha a estrutura de linearidade, a função que relaciona o valor esperado e o vector de covariáveis pode ser qualquer função diferenciável (Turkman e Silva, 2000).

Os MLG são caracterizados pela seguinte estrutura:

- Componente aleatória
- Componente sistemática ou estrutural
- Função de ligação

Componente aleatória

Dado o vector de covariáveis $\mathbf{x_i} = (1, x_{i1}, ..., x_{ip})^T$, as variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à família exponencial da forma (2.1), com o seu valor médio dado por:

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

e, possivelmente, um parâmetro de dispersão ϕ, que não depende de i.

Componente sistemática ou estrutural

Se $Y_1, ..., Y_n$ são variáveis aleatórias independentes com distribuição pertencente à família exponencial de distribuições, a componente sistemática do modelo pode ser escrito na forma:

$$\eta_i = \mathbf{x_i}^T\boldsymbol{\beta}, \tag{2.3}$$

onde η_i é o preditor linear para a i-ésima observação, $\mathbf{x_i} = (1, x_{i1}, ..., x_{ip})^T$ é o vetor $(p+1) \times 1$ da variável explicativa para a i-ésima observação e $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$, é o vector $((p+1) \times 1)$ de parâmetros de regressão desconhecidos.

Função de ligação

A função de ligação de um Modelo Linear Generalizado (MLG) é uma função g, monótona e diferenciável, que relaciona o valor esperado da variável resposta, μ, com o preditor linear η,

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

A função de ligação g(.) estabelece assim a ligação entre a componente aleatória e a componente sistemática do modelo. Quando a função de ligação torna o preditor linear η igual ao parâmetro canónico θ da família exponencial, diz-se que a função de ligação é a função de ligação canónica.

2.2 Método de estimação dos parâmetros

Num MLG, o parâmetro $\boldsymbol{\beta}$ é geralmente estimado pelo método da máxima verosimilhança. O parâmetro de dispersão ϕ, quando existe, é considerado um parâmetro perturbador, sendo a sua estimação realizada, geralmente aplicando o método dos momentos.

A função de verosimilhança para estimar o parâmetro $\boldsymbol{\beta}$, dada em Turkman e Silva (2000), é:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} exp\left\{\frac{w_i}{\phi}(y_i\theta_i - b(\theta_i)) + c(y_i, \phi, w_i)\right\} = \\ &= exp\left\{\frac{1}{\phi}\sum_{i=1}^{n} w_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi, w_i)\right\}, \end{aligned} \tag{2.4}$$

e portanto a função de log-verosimilhança (logaritmo da função de verosimilhança) é dado por:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi, w_i) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}), \tag{2.5}$$

onde $\ell_i(\beta) = \dfrac{w_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi, w_i)$

Os estimadores de máxima verosimilhança para $\boldsymbol{\beta}$ são obtidos como solução do sistema de equações de verosimilhança

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i \boldsymbol{\beta}}{\partial \beta_j} = 0, j = 1, ..., p. \tag{2.6}$$

Para obter estas equações escrevemos:

$$\frac{\partial \ell_i(\beta)}{\partial \beta_j} = \frac{\partial \ell_i(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j}.$$

onde

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

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{w_i var(Y_i)}{\phi},$$

$$\frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j} = x_{ij}.$$

Assim

$$\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{w_i(y_i - \mu_i)}{\phi} \frac{\phi}{w_i var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}, \tag{2.7}$$

e as equações de verosimilhança para $\boldsymbol{\beta}$ são

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, j = 1, 2..., p. \tag{2.8}$$

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

$$s(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} s_i(\boldsymbol{\beta}), \tag{2.9}$$

onde $s_i(\boldsymbol{\beta})$ é o vector de componentes $\dfrac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j}$ obtidas em (2.7).

A matriz de covariância da função score $I(\beta)$, é designada por matriz de informação de *Fisher* e é dada por:

$$I(\beta) = E\left[-\frac{\partial s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]. \tag{2.10}$$

Prara obter a matriz de informação de *Fisher* temos,

$$\begin{aligned} I(\beta) &= -E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = E\left(\frac{\partial \ell_i}{\partial \beta_j}\frac{\partial \ell_i}{\partial \beta_k}\right) \\ &= E\left[\left(\frac{(Y_i - \mu_i)x_{ij}}{var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\right)\left(\frac{(Y_i - \mu_i)x_{ik}}{var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\right)\right] \\ &= E\left[\frac{(Y_i - \mu_i)^2 x_{ij}x_{ik}}{(var(Y_i))^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right] \\ &= \frac{x_{ij}x_{ik}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \end{aligned}$$

e, portanto o elemento genérico de ordem *(j,k)* da matriz de informação de Fisher é

$$I(\beta)_{j,k} = -\sum_{i=1}^{n} E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Na forma matricial temos

$$I(\boldsymbol{\beta}) = \mathbf{X}^T W \mathbf{X}$$

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

$$\boldsymbol{w_i} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \times \frac{1}{var(Y_i)}. \tag{2.11}$$

2.3 Testes de hipóteses

A maior parte dos problemas de inferência relacionados com testes de hipóteses sobre o vector de parâmetro $\boldsymbol{\beta}$, podem ser formulados em termos de hipóteses lineares (hipótese nula H_0 ou hipótese alternativa H_1). Os testes de hipóteses são utilizados para determinar se os resultados de um estudo científico podem levar à rejeição da hipótese nula H_0, a um nível de significância pré-estabelecido. O estudo da teoria das probabilidades e a determinação da estatística de teste correta são fundamentais para a coerência de um teste de hipótese (Williams e Iyer, 1985).

No contexto dos MLG, pretende-se, por exemplo, averiguar se um determinado conjunto de covariáveis é estatisticamente significativo para o modelo, ou, analisar a significância estatística de cada um dos parâmetros, individualmente, ou comparar a qualidade do ajustamento de dois modelos, entre outras questões.

Os testes de hipóteses que apresenta-se de seguida dizem respeito, à significância estatística dos coeficientes de regressão.

Teste de *Wald*

O teste de *Wald* baseia-se na distribuição assintótica normal dos estimadores de máxima verosimilhança dos parâmetros do modelo.

Seja $\hat{\beta}_j$, o estimador de máxima verosimilhança de β_j, um particular parâmetro de um MLG. Então pretende-se testar,

$$H_0 : \beta_j = 0 \; versus \; H_1 : \beta_j \neq 0, j = 0, 1, 2....., p,$$

e a estatística de teste é dado por:

$$z_j = \frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}},$$

e, sob H_0, a estatística de teste tem uma distribuição assintótica *N(0,1)* quando ϕ é conhecido.

Teste de Razão da Verosimilhança

O teste de razão de verosimilhança 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).

Considera-se dois modelos encaixados, N_1 e N_2, com um número de paramêtros p_1 e p_2 respectivamente, tal que $p_1 < p_2$.

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

H_0 : Os modelos têm a mesma qualidade de ajustamento.

A estatística do teste é definida por:

$$-2ln\Lambda = -2ln\left(\frac{L_{N_1}(\boldsymbol{\beta})}{L_{N_2}(\boldsymbol{\beta})}\right) = -2\left(\ell_{N_1}(\hat{\boldsymbol{\beta}}) - \ell_{N_2}(\hat{\boldsymbol{\beta}})\right) \tag{2.12}$$

sendo $\ell_{N_1}(\beta)$, é a função do logaritmo de verosimilhança do modelo N_1 e $\ell_{N_2}(\beta)$ é a função do logaritmo da verosimilhança do modelo N_2. Repara-se que a estatística de teste se obtém a partir da razão de verosimilhança dos dois modelos, dai a designação de Teste da razão de verosimilhança.

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

$$-2ln\Lambda \sim \chi^2_{(p_2-p_1)}$$

2.4 Seleção de Modelos

Em modelos de regressão é necessário determinar um subconjunto de variáveis explicativas que melhor explique a variável resposta, isto é, de entre todas as variáveis explicativas disponíveis, devemos encontrar um subconjunto de variáveis importantes para o modelo. Qualquer procedimento para seleção ou exclusão de variáveis de um modelo é baseado num algoritmo que verifica a importância das variáveis, incluindo ou excluindo-as do modelo, baseando em uma regra de decisão. A importância da variável é definida em termos de uma medida de significância estatística do coeficiente associado à variável para o modelo. Existem três procedimentos automáticos para selecionar modelos.

- Método de *Forward*
- Método de *Backward*
- Método de *Stepwise*

Método *Forward*

Esse procedimento considera o modelo inicial, apenas a constante. A ideia do método é adicionar uma variável de cada vez. A primeira variável selecionada é aquela com maior correlação com a variável resposta.

Método de *Backward*

Enquanto que no método Forward o modelo inicial não tem nenhuma variável e adiciona variáveis a cada passo, o método *Backward* faz o caminho oposto, inicialmente incorpora todas as variáveis e depois, por etapas, retira sucessivamente as variáveis não significativas do modelo.

Método de *Stepwise*
Este método é uma combinação dos dois métodos: *Forward* e *Backward*.

2.5 Qualidade de Ajustamento

Nesta secção apresentam-se as medidas como o desvio e a estatística de Pearson generalizada que ajudam a avaliar a qualidade do modelo.

Análise de Desvio
Segundo Cordeiro e Demétrio (1986), um dos critérios de avaliação do modelo mais utilizado e que pode ser usada em modelos encaixados, assenta no valor de uma medida usualmente denominada de desvio, baseada na função de verosimilhança do modelo em estudo.

Consideremos a equação (2.5), o logaritmo da função de verosimilhança de um modelo linear generalizado, em que se substitui θ_i por $q(\mu_i)$, para fazer salientar, na função log-verosimilhança, a relação funcional existente entre θ_i e μ_i, é:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i \left[y_i q(\mu_i) - b(q(\mu_i)) \right]}{\phi} + c(y_i, \phi, w_i) \tag{2.13}$$

O modelo saturado ou completo (modelo com n parâmetros $\mu_1, ..., \mu_n$) 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 $\hat{\mu}$, com esse modelo e dos correspondentes valores observados y, o modelo saturado, que tem-se $\hat{\mu}_i = y_i$, e o máximo da função log-verosimilhança para este modelo é:

• Modelo Saturado com n parâmetros

$$\ell_S(\hat{\boldsymbol{\beta}_S}) = \sum_{i=1}^{n} \frac{w_i \left[y_i q(y_i) - b(q(y_i)) \right]}{\phi} + c(y_i, \phi, w_i). \tag{2.14}$$

Por outro lado, se designarmos por $\hat{\mu}_i$ a estimativa de máxima verosimilhança de μ_i, para *i=1,2,....n*, o máximo da função log-verosimilhança para o modelo em estudo com m parâmetros, é

• Modelo com n parâmetros

$$\ell_M(\hat{\boldsymbol{\beta}_M}) = \sum_{i=1}^{n} \frac{w_i \left[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i)) \right]}{\phi} + c(y_i, \phi, w_i). \tag{2.15}$$

Se compararmos o modelo em estudo M, com o modelo saturado S através da estatística de razão de verosimilhança, obtemos

• Desvio reduzido

$$D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2\left(\ell_M(\hat{\boldsymbol{\beta}_M}) - \ell_S(\hat{\boldsymbol{\beta}_S})\right) = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi}. \tag{2.16}$$

onde $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{w_i\left[y_i q(y_i) - b(q(y_i))\right] - w_i\left[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i))\right]}{\phi}$

A $D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ definido em (2.16) damos o nome de desvio reduzido; $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ damos o nome de desvio para o modelo corrente.

O desvio de um modelo avalia, portanto, a discrepância entre os valores ajustados pelo modelo completo e os valores ajustado pelo modelo saturado. 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.

Estatística de Pearson generalizada

Outra medida para avaliar a qualidade de ajustamento de um modelo não utilizados para modelos encaixados é a estatística de Pearson generalizada definida por

$$\chi^2 = \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{2.17}$$

onde $V(\hat{\mu}_i)$ é a função variância estimada para a distribuição do modelo em causa.

No caso da distribuição Normal, a estatística χ^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 estatística de *Pearson*, a comparação entre modelos encaixados não podem ser feita usando a diferença entre estatísticas de *Pearson*, contrariamente ao que sucede com a função desvio.

2.6 Análise dos Resíduos

Os resíduos constituem uma ferramenta de extrema importância para verificar se os pressupostos considerados na formulação do modelo são ou não correctos.

Um resíduo pode ser definido como sendo a discrepância entre cada valor observado e o respectivo valor ajustado pelo modelo, sendo conveniente usar valores padronizados, isto é, que tenham variância constante e o valor médio zero.

Resíduo Ordinário

O resíduo ordinário é simplesmente a diferença do valor observado para o valor ajustado para uma observação particular, e é definida por,

$$R_i = y_i - \hat{\mu}_i.$$

Resíduos de Pearson

O resíduo de Pearson para uma dada observação é dado por

$$R_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(Y_i)}} = \frac{(y_i - \hat{\mu}_i)w_i}{\sqrt{\hat{\phi}V(\hat{\mu}_i)}}$$

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

Resíduos de Pearson Padronizado

O resíduo de Pearson padronizado é definido por

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

uma vez que $Var(Y_i - \hat{\mu}_i) \approx Var(Y_i)(1 - h_{ii})$, (Turkman e Silva, 2000) onde h_{ii} são os valores da diagonal da matriz de projecção $H = w^{\frac{1}{2}}X(X^T wX)^{-1}X^T w^{\frac{1}{2}}$.

Resíduo do Desvio

O resíduo do desvio para i-ésima observação correspondente à contribuição dessa observação para o desvio do modelo. É uma medida de distância y_i em relação a $\hat{\mu}_i$ na escala do logarítmico da verosimilhança. O resíduo do desvio fica definido por:

$$R_i^D = sinal(y_i - \hat{\mu}_i)\sqrt{d_i},$$

onde d_i é a contribuição de cada observação i para a função desvio e o $sinal(x) = -1$, se $x < 0$, e $sinal(x) = +1$, se $x > 0$.

Resíduo do Desvio Padronizado

O resíduo do desvio padronizado é dado por

$$R_i^{D^*} = \frac{R_i^D}{\sqrt{\hat{\phi}(1 - h_{ii})}}.$$

Resíduo Quantílico Aleatorizado

Nas situações em que se tem uma variável resposta que não segue a distribuição Normal, os resíduos, muitas vezes não têm boa aproximação à distribuição normal, ainda que o modelo se ajuste bem aos dados. Isto é, particularmente notável na modelação de dados discretos, sobretudo quando os dados assumem valores pequenos. Propostos por Dunn e Smyth (1996), os resíduos quantílicos aleatorizados apresentam distribuição Normal, independente da distribuição da variável resposta

Estes resíduos baseiam-se no teorema da inversa da função distribuição acumulada. No contexto de MLG, seja $F(y; \mu, \phi)$ a função distribuição acumulada de uma

variável aleatória Y. Se Y é uma variável aleatória contínua, o teorema da inversa da função distribuição acumulada garante que $U_i = F(y; \mu, \phi)$ tem distribuição uniforme no intervalo $[0; 1]$. Se os parâmetros do modelo são consistentemente estimados então, R_i^q converge para uma distribuição Normal padrão. O resíduo quantílico aleatorizado é dado por

$$R_i^q = \Phi^{-1}\left\{F(y_i; \hat{\mu}_i, \phi)\right\},$$

onde $\Phi(.)$ é a função de distribuição acumulada da distribuição normal padrão.

Tipos de gráficos

Além dos gráficos, dos resíduos *versus* valores ajustados e o *QQplot*, utilizados para identificar problemas relacionados com o ajustamento de um modelo, nesta dissertação apresenta-se também o *rootogram* uma ferramenta gráfica adicional com objectivo de diagnosticar ou identificar problemas relacionados com o ajustamento e tratar questões como sobredispersão (ou subdispersão);

1) O gráfico dos resíduos *versus* valores ajustados permite identificar observações consideradas *outlieres*;

2) O gráfico *QQplot* representa o ajuste dos resíduos do modelo em uma distribuição normal, ou seja, os quantis empíricos dos resíduos são comparados com os quantis teóricos de uma distribuição, quanto mais próximo for essa comparação teremos uma recta linear perfeita indicando que o quantil teórico e empírico são iguais;

3) O *rootogram* é uma ferramenta gráfica associada ao trabalho de Tukey e Cleveland (1984) que foi utilizado originalmente para avaliar a qualidade do ajuste de distribuições univariadas. Zeileis *et al.* (2008) estenderam a sua utilidade em modelos de regressão para dados de contagem e mostram que esta ferramenta gráfica é particularmente útil para diagnosticar ou identificar problemas relacionados ao ajuste e tratar questões como sobredispersão (ou subdispersão).

2.7 Critérios de Informação

Ao selecionarmos modelos é preciso ter em conta que não existem modelos exactos. Há apenas modelos aproximados da realidade que podem causar perda de informações. Deste modo, é necessário fazer a selecção do "melhor"modelo, dentre aqueles que foram ajustados, para explicar o fenómeno sob estudo.

Akaike (1974) definiu o critério de informação como:

$$AIC = -2\ell(\beta) + 2p,$$

onde $\ell(\beta)$ é a função do logaritmo da verosimilhança do modelo e p número de parâmetros. O modelo selecionado é o modelo com menor valor AIC.

Um outro critério de informação usado para selecionar modelo é o Critério de Informação Bayesiano (BIC), proposto por Schwarz *et al.* (1978) e é dado por:

$$BIC = -2\ell(\beta) + plog(n),$$

onde $\ell(\beta)$ é a função do logaritmo da verosimilhança do modelo e p o número de parâmetros e n é o número de observações do mesmo modelo. Do mesmo modo, o modelo com o menor BIC é o modelo selecionado.

2.8 Modelos de Regressão para Dados de Contagem

Dentre os MLG, os modelos de regressão para dados de contagem têm, por objectivo principal estudar o comportamento de uma variável dependente, que assume apenas valores inteiros não negativos, com base no comportamento de variáveis explicativas.

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

Segundo Cameron e Trivedi (2009), o ponto inicial para o estudo do modelo de regressão para dados de contagem, é a apresentação da distribuição de Poisson.

2.8.1 Modelo de regressão de Poisson

Definição: Uma variável aleatória Y tem uma distribuição de Poisson com parâmetro μ se a sua função de probabilidade é dada por:

$$f(y_i, \mu_i) = P(Y_i = y_i) = \frac{exp(-\mu)\mu_i^{y_i}}{y_i!}, y_i = 0, 1, 2, ... \tag{2.18}$$

em que $\mu_i > 0$ representa o número médio de ocorrências de um dado acontecimento.

O valor esperado e a variância da distribuição de Poisson são dadas por:

$$E(Y) = \sum_{y=0}^{\infty} yP(Y = y)$$

Neste caso,

$$E(Y) = \sum_{y=0}^{\infty} y\frac{exp(-\mu)\mu^y}{y!} = exp(-\mu)\sum_{y=1}^{\infty}\frac{\mu^y}{(y-1)!} \tag{2.19}$$

fazendo a mudança de variáveis, j=y-1, substituindo na equação (2.19) tem-se

$$E(Y) = exp(-\mu) \sum_{y=j+1}^{\infty} \frac{\mu^{j+1}}{j!} = \mu exp(-\mu) \sum_{j=0}^{\infty} \frac{\mu^{j}}{j!} = \mu exp(-\mu) exp(\mu) = \mu. \quad (2.20)$$

e, de maneira análoga, tem-se que

$$Var(Y) = E\left(Y^2\right) - [E(Y)]^2 = E(Y) = \mu \quad (2.21)$$

Sejam $Y_1, ..., Y_n$ uma amostra aleatória de uma variável aleatória Y que representa o número de ocorrências de um acontecimento raro num determinado período de tempo ou espaço. Dado um vector de variáveis explicativas $X = (X_1, ..., X_P)$ e uma observação $\mathbf{x_i}^T = (1, x_{i1}, ..., x_{ip})$ do indivíduo i, assume-se que

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

onde

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

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

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

Pretende-se modelar o valor esperado de $Y|X = \mathbf{x_i}$ como combinação linear das variáveis explicativas. Poderia escrever-se um modelo linear simples na forma

$$\mu(\mathbf{x_i}) = \mathbf{x_i}^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_P x_{iP}$$

.

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

$$log(\mu(\mathbf{x_i})) = \beta_0 + \beta_1 x_{i1} + ... + \beta_P x_{iP}.$$

Assim, o modelo de regressão de Poisson é definido por

$$Y|X = \mathbf{x_i} \sim P(\mu(\mathbf{x_i})) \quad (2.22)$$

e

$$log(\mu(\mathbf{x_i})) = \beta_0 + \beta_1 x_{i1} + ... + \beta_P x_{iP} \quad (2.23)$$

2.8.2 Estimação dos parâmetros

Para estimar os parâmetros, utiliza-se o método de estimação de máxima verosimilhança. A função de verosimilhança para o modelo de regressão de Poisson é dado por,

$$L(\beta) = \prod_{i=1}^{n} \frac{exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \tag{2.24}$$

e a função de log-verosimilhança é dado por:

$$\ell(\beta) = \sum_{i=1}^{n} [-\mu_i + y_i ln(\mu_i) - ln(y_i!)] \tag{2.25}$$

Como $\mu_i = exp(\mathbf{x_i}^T \boldsymbol{\beta})$, vem

$$\ell(\beta) = \sum_{i=1}^{n} \left[-exp(\mathbf{x_i}^T \boldsymbol{\beta}) + y_i(\mathbf{x_i}^T \boldsymbol{\beta}) - ln(y_i!)\right].$$

Portanto, maximizando a função de log-verosimilhança $\ell(\beta)$ com respeito a $\boldsymbol{\beta}$ temos:

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left(y_i - exp(\mathbf{x_i}^T \boldsymbol{\beta})\right) \mathbf{x_i}, j = 0, 1, 2....p$$

e as equações de verosimilhança para $\boldsymbol{\beta}$ são:

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left(y_i - exp(\mathbf{x_i}^T \boldsymbol{\beta})\right) \mathbf{x_i} = 0, j = 0, 1, 2...p. \tag{2.26}$$

A função desvio no modelo de regressão de Poisson é definida pela seguinte expressão,

$$D(Y; \hat{\mu}) = 2 \sum_{i=1}^{n} \left\{ y_i ln(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i) \right\} \tag{2.27}$$

A função desvio reduz-se a

$$D(Y; \hat{\mu}) = 2 \sum_{i=1}^{n} \left(y_i ln(\frac{y_i}{\hat{\mu}_i}) \right) \tag{2.28}$$

para modelos com termo constante, β_0, porque neste caso

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

A estatística de Pearson generalizada é definida por

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

2.8.3 Sobredispersão e Subdispersão

A sobredispersão é um problema que ocorre frequentemente na prática quando se aplica o modelo de regressão de Poisson, surgindo quando a variância da variável resposta é superior ao valor esperado. Outro problema menos comum, mas que tem ganhado a atenção nos estudos de regressão para dados de contagem é a subdispersão, surge quando a variância da variável resposta é menor ao valor esperado.

Os critérios que indicam a sobredispersão aparente do modelo de regressão de Poisson são (Hilbe, 2011).

- O modelo omite importantes variáveis explicativas;
- Os dados incluem outliers;
- O modelo não inclui um número suficiente de termos de interação;
- Uma variável explicativa precisa ser transformado em outra escala;
- Há situações em que os dados são muito escassos e mais dados precisam ser recolhidos e incluídos no modelo;
- Valores ausentes existem nos dados, mas não são distribuídos aleatoriamente nos dados.

Para identificar a sobredisperão nos dados, podemos utilizar o desvio, $D(y, \hat{\mu})$ também utilizado para testar a qualidade do ajustamento do modelo. O cálculo é baseado na aproximação χ^2 do desvio reduzido. Se existir sobredispersão, então $\frac{D}{\alpha}$ segue uma distribuição qui-quadrado com *n-p* graus de liberdade, e isso leva ao seguinte estimador para α (Zuur *et al.*, 2009),

$$\hat{\alpha} = \frac{D}{n-p}$$

Se a estimativa deste parâmetro for maior que um, é uma indicação da existência de sobredispersão. Caso seja menor que um (1), prossegue-se com o teste da dispersão definido em Cameron e Trivedi (2005).

Teste da Dispersão

Um teste estatístico de sobredispersão ou subdispersão é, altamente desejável após a estimação de um modelo de regressão de Poisson. A maioria dos modelos de contagem que apresentam sobredispersão ou subdispersão especificam a variância da seguinte forma:

$$Var\left[y_i|x_i\right] = \mu_i + \alpha * g(\mu_i), \tag{2.29}$$

onde α é um parâmetro desconhecido e g (.) é uma função conhecida. As especificações mais comuns da função g (μ_i) são $g(\mu) = \mu^2$ ou $g(\mu) = \mu$ (Cameron e Trivedi, 2005).

Se $\alpha = 0$, há equidispersão; se $\alpha < 0$ ocorre subdispersão e se $\alpha > 0$ indica sobredispersão.

Para testar a sobredispersão, define-se

$$H_0 : \alpha = 0 \; versus \; H_1 : \alpha > 0$$

e depois de estimar o modelo de regressão de Poisson, calculam-se $\hat{\mu}_i = exp(\mathbf{x_i}^T\hat{\beta})$ e estima-se o modelo de regressão linear

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha\frac{g(\hat{\mu}_i)}{\hat{\mu}_i} + e_i$$

onde e_i é um erro aleatório. Sob $H_{0,}$ a estatística de teste e a sua distribuição são

$$T = \frac{\hat{\alpha} - \alpha}{s(\hat{\alpha})} \sim N(0;1)$$

Este teste pode ser utilizado para testar a subdispersão ($\alpha < 0$).

O teste de dispersão encontra-se implementado no *Package Countreg* do R *package* recorrendo à função *disptest()* ou *dispersiontest()*.

2.9 Modelo de Regressão Binomial Negativa

O modelo de regressão Binomial Negativa é utilizado, quando ocorre sobredispersão no modelo de regressão de Poisson.

Considere então que $Y_i \sim BN(\mu_i, \alpha)$, com os parâmetros $\mu_i \geqslant 0$ e $\alpha \geqslant 0$ cuja função de probabilidade é dada por

$$f(y;\mu,\alpha) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} com, y_i = 0,1,2,... \quad (2.30)$$

onde α é denominado parâmetro de dispersão.

Na distribuição Binomial Negativa a variância é superior que a média, e elas são representadas por:

$$E(Y) = \mu \; e \; Var(Y) = \mu + \alpha\mu^2$$

.

Repara-se que, quando o valor do parâmetro de heterogeneidade (α) tende para zero, a distribuição Binomial Negativa tende para a distribuição de Poisson.

Considere-se a variável aleatória Y, com n observações que representa o número de ocorrência de um determinado acontecimento num certo período de tempo ou espaço. Dado um vector de variáveis explicativas $X = (X_1, ..., X_P)$ e uma observação $\mathbf{x_i}^T = (1, x_{i1}, ..., x_{ip})$ do individuo i, assume-se que

$$Y|X = \mathbf{x_i} \sim NB(\mu(\mathbf{x_i}), \alpha)$$

onde

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

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

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

O modelo de regressão Binomial Negativa é expresso por:

$$log(\mu(\mathbf{x_i})) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{iP}. \tag{2.31}$$

onde o log representa a função de ligação do modelo em questão.

2.9.1 Estimação dos Parâmetros

O método utilizado para estimar os parâmetros de regressão Binomial Negativa é a máxima verosimilhança.

A função log-verosimilhança para n observações da distribuição Binomial Negativa é dada por Hilbe (2011),

$$\ell(\beta) = \sum_{i=1}^{n} \left\{ y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \left(\frac{1}{\alpha}\right) ln(1+\alpha\mu_i) + ln\left(\frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})}\right) \right\}. \tag{2.32}$$

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

O desvio do modelo de regressão Binomial Negativa, é dado por

$$D = 2\sum_{i=1}^{n} \left\{ y_i ln\left(\frac{y_i}{\mu_i}\right) - \left(\frac{1}{\alpha} + y_i\right) ln\left(\frac{1+\alpha y_i}{1+\alpha\mu_i}\right) \right\}. \tag{2.33}$$

Nesta situação, a estatística de Pearson generalizada é,

$$\chi^2_P = \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i + \alpha\hat{\mu_i^2}}. \tag{2.34}$$

2.9.2 Teste de Vuong

Vuong (1989) introduziu um teste que é um método adequado para comparar modelos aninhados. Em particular utiliza-se este teste nos modelos de regressão de Poisson bem como nos modelos de regressão 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 $H_0 : E(m_i) = 0$ a estatística de teste é dada por

$$V = \frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} m_i)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} \tag{2.35}$$

onde 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, caso $|V| < 1,96$ os dois modelos são equivalentes. Ou seja, $M_1 \subset M_2$

Se $-1.96 < V < 1.96 \Rightarrow M_1 \approx M_2$.

Se $V > 1.96 \Rightarrow M_1$ eleito.

se $V < 1.96 \Rightarrow M_2$ eleito.

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

Capítulo 3

Modelo de Regressão de Poisson Generalizado

Neste capítulo pretende-se descrever a Distribuição de Poisson Generalizada (DPG) e apresentar os modelos de regressão de Poisson Generalizada para dados de contagem.

3.1 Distribuição de Poisson Generalizado

Em dados de contagem, nem sempre é observada a equidispersão. Muitos investigadores têm propostos distribuições baseadas na distribuição de Poisson. Uma dessas distribuições é a Distribuição de Poisson Generalizada (DPG).

Definição I: Uma variável aleatória Y que assume valores inteiros não negativos que segue uma Distribuição de Poisson Generalizada com parâmetros $\theta_i > 0$ e $0 \leq \gamma < 1$, DPG (θ, γ) se a função de probabilidade é dada por:

$$y \sim DPG(\theta, \gamma)$$

$$P(Y_i = y_i|\theta_i, \gamma) = \begin{cases} \theta_i(\theta_i + \gamma y_i)^{y_i - 1} \dfrac{exp(-(\theta_i + \gamma y_i))}{y_i!}, y_i = 0, 1, 2, 3, ... \\ 0, \text{caso contrário.} \end{cases} \tag{3.1}$$

3.1.1 Restrições no Espaço Paramétrico

Consul e Shenton (1973) definiram que γ poderia assumir valores negativos desde que $\theta + m\gamma > 0$, onde m é o maior valor que a variável aleatória Y pode assumir para que expressão $\theta + m\gamma > 0$ seja válida. A restrição do domínio da variável aleatória Y quando γ assume valores negativos não foi suficiente para evitar que a soma de todas as probabilidades seja diferente de 1, isto é, $\sum_{y=0}^{+\infty} P(Y = y) \neq 1$.

Diante disso, Consul e Shoukri (1984) corrigiram os valores possíveis de γ para $0 \leq \gamma < 1$. Contudo, isto implicava que a DPG deixasse de ser indicada para modelar dados caracterizados por subdispersão ($\gamma < 0$), quando esta distribuição já tinha demonstrado ter um bom ajustamento a alguns conjuntos de dados deste tipo. Para corrigir o problema da soma das probabilidades ser diferente de 1, Consul e Shoukri (1985) fizeram uma outra modificação propondo uma correção feita habitualmente em modelos truncados,

$$P\left(Y_i = y_i|\theta_i, \gamma\right) = \frac{P(Y_i = y_i|\theta_i, \gamma)}{\sum_{y=0}^{m} f(Y_i = y_i|\theta_i, \gamma)} = \frac{P(Y_i = y_i|\theta_i, \gamma)}{F_m} \tag{3.2}$$

sendo $\theta_i < 0$ e $F_m = \sum_{y=0}^{m} f(Y_i = y_i|\theta_i, \gamma)$. A soma das probabilidades passou a ser igual a 1, mas o modelo revelou-se inadequado por omitir valores negativos para θ_i.

Os mesmos autores realizaram um estudo de simulação para verificar os efeitos do fator de correção F_m, chegando à conclusão de que o erro resultante, inferior a 0.05% era desprezável, desde que o número de acontecimentos distintos com probabilidades não nulas fosse pelo menos de cinco (daí impuseram a restrição $m \geqslant 4$) ou que θ estivesse fora do intervalo $[0.7, 4.5]$. Assim, sugeriram a utilização do modelo sem factor corretivo quando $m \geq 4$, devendo-se, nos restantes casos, aplicar o modelo truncado, apesar da dificuldade de determinar por este meio, estimativas razoáveis para γ e θ.

Por último, Consul e Famoye (1989) propuseram a restrição final:

tem-se $max\left(-1, -\frac{\theta}{m}\right) < \gamma < 1$ acrescido da restrição de que $m \geq 4$ quando $\gamma < 0$.

Definição II: Uma variável aleatória Y, que assume valores inteiros não negativos segue uma Distribuição de Poisson Generalizada com parâmetros θ e γ, se a função de probabilidade é dada por (Zamani e Ismail, 2012),

$$P(Y_i = y_i|\theta_i, \gamma) = \begin{cases} \theta_i(\theta_i + \gamma y_i)^{y_i - 1} \dfrac{exp(-(\theta_i + \gamma y_i))}{y_i!}, y_i = 0, 1, 2, 3, ... \\ 0, y_i > m, quando\ \gamma < 0 \end{cases} \tag{3.3}$$

onde $\theta_i > 0$ e $max\left(-1, -\frac{\theta_i}{m}\right) < \gamma < 1$ e $m \geqslant 4$ é o maior número positivo para o qual $\theta_i + m\gamma > 0$, quando $\gamma < 0$.

O valor médio e a variância da DPG são dadas por:

$$E\left[Y_i\right] = \frac{\theta_i}{1 - \gamma}, \tag{3.4}$$

$$Var\left[Y_i\right] = \frac{\theta_i}{(1-\gamma)^3} = \frac{E\left[Y_i\right]}{(1-\gamma)^2} = \phi E\left[Y_i\right] \tag{3.5}$$

e quando $\gamma = 0$, $E\left[Y_i\right] = Var\left[Y_i\right] = \theta_i$.

O termo $\phi = \dfrac{1}{(1-\gamma)^2}$ é o factor de dispersão. Esta relação entre o valor médio e a variância permitem concluir que:

- Se $\gamma = 0$, ocorre a equidispersão e a DPG será a distribuição de Poisson.
- Se $\gamma > 0$, ocorre a sobredispersão e se $\gamma < 0$ ocorre a subdispersão.

A título de exemplo, usando a biblioteca *RNGforGPD* no R *package*, foram simuladas três amostras aleatórias com dimensão *n=100* com a DGP. Onde se verifica situações em que ocorre equidispersão, sobredispersão ou subdispersão.

Para cada amostra simulada, calcula-se as estimativas dos parâmetros θ e γ pelo método *Inversion* (Demirtas, 2017). A Tabela 3.1 mostra os valores dos parâmetros reais e os valores estimados.

Tabela 3.1: Valores dos parâmetros e os valores estimados na DGP

Amostra	Valores reais	Média e variância populacional	Parâmetros estimados	Média e variância amostral	Situação
1	$\theta = 3$	$E\left[Y\right] = 3$	$\hat{\theta} = 2,891$	$\overline{x} = 2,810$	Equidispersão
	$\gamma = 0$	$Var\left[Y\right] = 3$	$\hat{\gamma} = -0,029$	s^2=2,650	
2	$\theta = 5$	$E\left[Y\right] = 7,143$	$\hat{\theta} = 5,047$	$\overline{x} = 6,749$	Sobredispersão
	$\gamma = 0,3$	$Var\left[Y\right] = 14,577$	$\hat{\gamma} = 0,252$	s^2= 12,072	
3	$\theta = 5$	$E\left[Y\right] = 3,333$	$\hat{\theta} = 5,360$	$\overline{x} = 3,230$	
	$\gamma = -0,5$	$Var\left[Y\right] = 1,481$	$\hat{\gamma} = -0,659$	s^2=1,173	Sudispersão

As Figuras 3.1, 3.2 e 3.3 apresentam os gráficos das amostras simuladas para diferentes valores dos parâmetros θ e γ. A primeira amostra foi gerada com $\theta = 3$ e $\gamma = 0$ (Figura 3.1), a segunda amostra com $\theta = 5$ e $\gamma = 0.3$ (Figura 3.2) e a terceira amostra com $\theta = 5$ e $\gamma = -0.5$ (Figura 3.3).

3.2 Distribuição de Poisson Generalizada I

Alguns autores propuseram diferentes parametrizações para a DGP. Wang e Famoye (1997) propuseram a parametrização,

$$\theta_i = \frac{\mu_i}{(1+\alpha\mu_i)} \; e \; \gamma = \alpha\frac{\mu_i}{(1+\alpha\mu_i)}, \tag{3.6}$$

Figura 3.1: Equidispersão na DGP

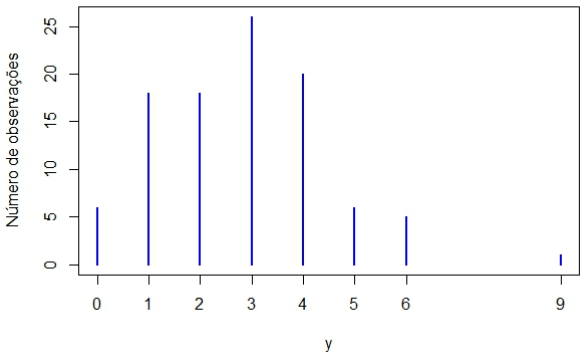


Figura 3.2: Sobredispersão na DGP

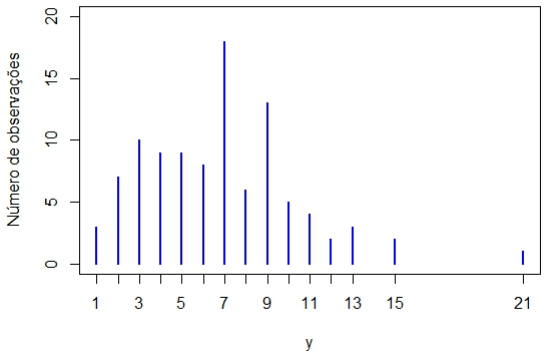
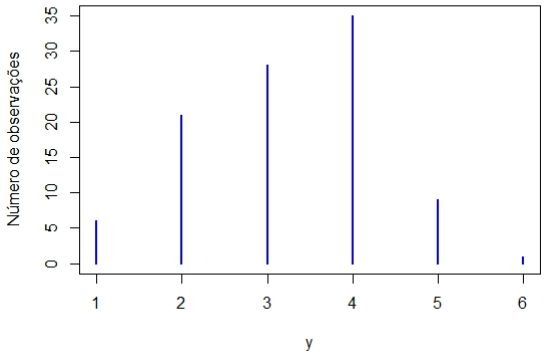


Figura 3.3: Subdispersão na DGP



para se ter $E[Y_i] = \mu_i$. Assim, substituindo as equações (3.6) na função (3.3), obtém-se a distribuição de Poisson Generalizada I (DPGI) com parâmetros (μ e α), dada por:

$$P(Y_i = y_i|\mu_i, \alpha) = \begin{cases} \dfrac{(\mu_i)^{y_i}(1+\alpha y_i)^{y_i-1} exp\left(-\dfrac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right)}{(1+\alpha\mu_i)^{y_i} y_i!}, y_i = 0, 1, ... \\ 0, y_i > m, \alpha < 0 \end{cases} \quad (3.7)$$

onde valor esperado $E[Y_i] = \mu_i$ e a variância $Var[Y_i] = \mu_i(1+\alpha\mu_i)^2$.

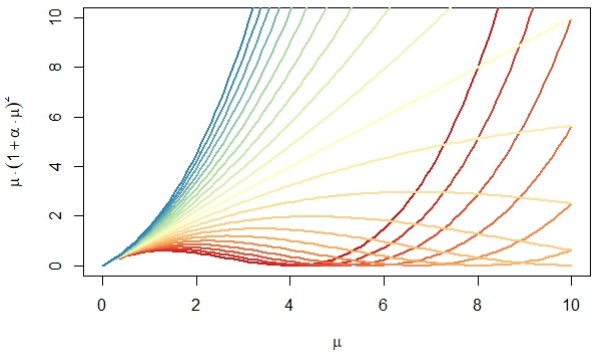
• Se $\alpha = 0$, a DGPI, reduz à distribuição de Poisson, resultando $E[Y_i] = Var[Y_i] = \mu_i$.

• Se $\alpha > 0$, tem-se $E[Y_i] < Var[Y_i]$, e a distribuição representa dados de contagem com sobredispersão.

• Se $\frac{-2}{\mu_i} < \alpha < 0$, tem-se $E[Y_i] > Var[Y_i]$ e a distribuição representa dados de contagem com subdispersão (Sellers e Morris, 2017).

A relação entre o valor esperado e a variância na DPGI é cúbica, como se pode observar na Figura (3.4).

Figura 3.4: Relação entre o valor esperado e variância para diferentes valores de α na DPGI



3.2.1 Distribuição de Poisson Generalizada II

Para estabelecer a linearidade entre a variância e o valor esperado da variável aleatória Y, Consul e Famoye (1992) propuseram a parametrização,

$$\theta_i = \frac{\mu_i}{\alpha} \; e \; \gamma = \frac{\alpha - 1}{\alpha}, \tag{3.8}$$

substituindo as equações (3.8) na função (3.3), obtém-se a distribuição de Poisson Generalizada II (DPGII), dada por:

$$P(Y_i = y_i|\mu_i, \alpha) = \begin{cases} \mu_i(\mu_i + (\alpha - 1)y_i)^{y_i - 1}\alpha^{-y_i}\dfrac{exp(-\alpha^{-1}(\mu_i + (\alpha - 1)y_i))}{y_i!} \\ 0, y_i > m, \alpha < 1 \end{cases} \tag{3.9}$$

onde $\alpha \geq max\left(\frac{1}{2}, 1 - \frac{\mu_i}{4}\right)$, e m é o maior número positivo para o qual $\mu_i + m(\alpha - 1) > 0$ quando $\alpha < 1$. Para essa distribuição, tem-se que o valor esperado $E[Y_i] = \mu_i$ e a variância $Var[Y_i] = \alpha^2\mu_i$.

• Se $\alpha = 1$, a DGPII reduz à distribuição de Poisson, resultando $E[Y_i] = Var[Y_i] = \mu_i$.

• Se $\alpha > 1$, tem-se $E[Y_i] < Var[Y_i]$, e a distribuição representam dados de contagem com sobredispersão.

• Se $\frac{1}{2} \leqslant \alpha < 1$ e $\mu > 2$, tem-se $E[Y_i] > Var[Y_i]$, e a distribuição representa dados de contagem com subdispersão.

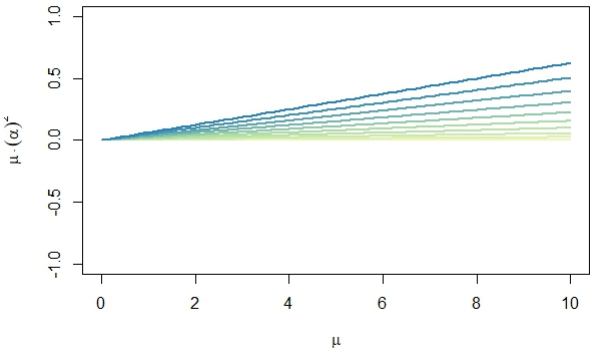
A Figura 3.5 , mostra a relação linear entre o valor esperado e a variância na DPGII.

Muitas aplicações da DGP podem ser encontradas na literatura.

Janardan *et al.* (1979) utilizaram a DPG para o estudo do número de aberrações cromossómicas nas células do linfócito humano. A aberração cromossómica é uma mudança no número ou na estrutura dos cromossomas, é produzida por agentes que podem elevar a frequência das mutações e é causada principalmente por radiações ou substâncias cancerígenas.

Segundo Consul e Shenton (1973), o número de clientes servidos, numa fila de espera em qualquer período, será uma variável aleatória X com distribuição de probabilidade *Lagrangiana*. Em um processo de fila de espera no qual se aplica a DPG, X é uma variável aleatória que designa o número de clientes que esperam em uma fila antes do início do serviço.

Figura 3.5: Relação entre o valor esperado e variância para diferentes valores de α na DPGII



Na Biologia, há exemplos de conjunto de dados de contagem que sugerem distribuição diferente da distribuição de Poisson. Por exemplo, a concentração ou a dispersão de ovos de pragas postos em folhas de plantas indicam a variabilidade diferente do que se espera para uma distribuição de Poisson. Diante disso, Janardan *et al.* (1979) demonstraram que alguns padrões de comportamento podem ser explicados e descritos por uma DPG.

Os modelos de dados do total dos sinistros é um dos mais importantes para a teoria de risco. Usualmente, quando a quantidade de indemnizações apresenta valor esperado igual à variância, utiliza-se a distribuição de Poisson e, quando a variância é maior que a média, utiliza-se a distribuição Binominal Negativa. Consul (1993) comparou a DPG sugerida por Consul e Jain (1973) com outras distribuições e concluiu que é plausível a utilização dessa distribuição.

3.3 Modelo de Regressão de Poisson Generalizado

Nesta secção serão apresentados os dois modelos de regressão de Poisson Generalizado : o modelo de regressão Poisson Generalizado I e o modelo de regressão Poisson Generalizado II.

3.3.1 Modelo de Regressão Poisson Generalizado I

Consul e Famoye (1992) demonstram como as covariáveis podem ser introduzidas num modelo de regressão cuja variável resposta segue uma DPG.

A relação entre o valor médio da variável Y e o preditor linear do modelo é representada pela função de ligação,

$$log\left(\mu_i\right) = log\left(\frac{\theta_i}{1-\gamma}\right) = \mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta} = \sum_{j=1}^{p}\beta_j x_{ij} \tag{3.10}$$

onde $\mathbf{x_i} = (1, x_{i1}, x_{i2}, ...x_{ip})$ é o vector das variáveis explicativas, $\boldsymbol{\beta}$ é o vector de regressão dos parâmetros e p é o número de variáveis explicativas.

Substituindo a equação (3.10) na função (3.7) obtém-se o modelo de regressão de Poisson Generalizado I, dada por:

$$P(Y_i = y_i|\boldsymbol{\beta},\alpha) = \begin{cases} \left(\frac{exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}{1+\alpha exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}\right)^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\times \\ \times exp\left(-\frac{exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})(1+\alpha y_i)}{1+\alpha exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}\right), y_i = 0,1,.... \\ 0, y_i > m, \alpha < 0 \end{cases} \tag{3.11}$$

e portanto a função de verosimilhança do modelo de regressão de Poisson Generalizado I é dada por:

$$L(\boldsymbol{\beta},\alpha) = \prod_{i=1}^{n}\left(\frac{exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}{1+\alpha exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}\right)^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\times \\ \times exp\left(-\frac{exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})(1+\alpha y_i)}{1+\alpha exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}\right), \tag{3.12}$$

logo a função de log-verosimilhança é dada por:

$$\ell(\boldsymbol{\beta},\alpha) = \sum_{i=1}^{n} y_i log\left(\frac{exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}{1+\alpha exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}\right) + \sum_{i=1}^{n}\left(y_i - 1\right) log\left(1+\alpha y_i\right) - \\ - \sum_{i=1}^{n}\frac{exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})(1+\alpha y_i)}{1+\alpha exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})} - \sum_{i=1}^{n} log(y_i!) \tag{3.13}$$

Considerando,

$\ell(\boldsymbol{\beta},\alpha) = \ell_1(\boldsymbol{\beta},\alpha) + \ell_2(\boldsymbol{\beta},\alpha) + \ell_3(\boldsymbol{\beta},\alpha) + \ell_4(\boldsymbol{\beta},\alpha)$, onde

$$\ell_1(\boldsymbol{\beta},\alpha) = \sum_{i=1}^{n} y_i log\left(\frac{exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}{1+\alpha exp(\mathbf{x_i}^{\boldsymbol{T}}\boldsymbol{\beta})}\right),$$

$$\ell_2(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} (y_i - 1) \, log \, (1 + \alpha y_i) \, ,$$

$$\ell_3(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \frac{exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})(1 + \alpha y_i)}{1 + \alpha exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})},$$

$$\ell_4(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} log(y_i! \,),$$

os estimadores de máxima verosimilhança de $(\hat{\boldsymbol{\beta}}, \hat{\alpha})$, podem ser obtidos maximizando a função de log-verosimilhança $\ell(\boldsymbol{\beta}, \alpha)$ com respeito $\boldsymbol{\beta}$ e α. O sistema de equações da verosimilhança são

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} = 0, j = 1, 2, ..., p \\ \dfrac{\partial \ell(\boldsymbol{\beta}, \alpha)}{\partial \alpha} = 0 \end{cases} \tag{3.14}$$

Assim,

$$\frac{\partial \ell(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} = \frac{\partial \ell_1(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} + \frac{\partial \ell_2(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} + \frac{\partial \ell_3(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} + \frac{\partial \ell_4(\boldsymbol{\beta}, \alpha))}{\partial \beta_j}.$$

Onde

$$\frac{\partial \ell_1(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} = \frac{\sum_{i=1}^{n} \partial \left(y_i log \left(\dfrac{exp \, (\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})}{1 + \alpha exp \, (\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})} \right) \right)}{\partial \beta_j}$$

$$\frac{\partial \ell_1(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i \left[x_{ij} exp \, (\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta}) \, (1 + \alpha exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})) \right]}{(1 + \alpha exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta}))^2} \frac{(1 + \alpha exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta}))}{exp \, (\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})} -$$

$$- \sum_{i=1}^{n} \frac{y_i \left[exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta}) \alpha x_{ij} exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta}) \right]}{(1 + \alpha exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta}))^2} \frac{(1 + \alpha exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta}))}{exp \, (\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})} = \sum_{i=1}^{n} \frac{y_i x_{ij} (1 + \alpha exp(\mathbf{x_i^T} \boldsymbol{\beta})}{(1 + \alpha exp(\mathbf{x_i}^{T} \boldsymbol{\beta}))^2},$$

$$\frac{\partial \ell_2(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} = \frac{\partial \left(\sum_{i=1}^{n} (y_i - 1) \, log \, (1 + \alpha y_i) \right)}{\partial \beta_j} = 0;$$

$$\frac{\partial \ell_3(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} = \frac{\sum_{i=1}^{n} \partial \left(\dfrac{exp(\mathbf{x_i}^{\boldsymbol{T}} \boldsymbol{\beta})(1 + \alpha y_i)}{1 + \alpha exp(\mathbf{x_i}^{T} \boldsymbol{\beta})} \right)}{\partial \beta_j} =$$

$$= \sum_{i=1}^{n} \frac{[x_{ij} exp(\mathbf{x_i}^T \boldsymbol{\beta})(1+\alpha y_i)(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta})) - exp(\mathbf{x_i}^T \boldsymbol{\beta})(1+\alpha y_i)\alpha x_{ij} exp(\mathbf{x_i}^T \boldsymbol{\beta})]}{(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta}))^2}$$

$$= \sum_{i=1}^{n} \frac{[x_{ij} exp(\mathbf{x_i}^T \boldsymbol{\beta})(1+\alpha y_i)]}{(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta}))^2},$$

$$\frac{\partial \ell_4(\boldsymbol{\beta}, \beta_j)}{\partial \beta_j} = \frac{\partial \left(\sum_{i=1}^{n} log(y_i!)\right)}{\partial \beta_j} = 0.$$

Logo a derivada da função log-verosimilhança $\ell(\beta, \alpha)$ com respeito $\boldsymbol{\beta}$ é dada por:

$$\frac{\partial \ell(\beta, \alpha)}{\partial \beta} = \frac{\partial \ell_1(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} + \frac{\partial \ell_2(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} + \frac{\partial \ell_3(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} + \frac{\partial \ell_4(\boldsymbol{\beta}, \alpha)}{\partial \beta_j} =$$

$$= \sum_{i=1}^{n} \frac{y_i x_{ij}(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta})}{(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta}))^2} - \sum_{i=1}^{n} \frac{[x_{ij} exp(\mathbf{x_i}^T \boldsymbol{\beta})(1+\alpha y_i)]}{(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta}))^2}.$$

Como $exp(\mathbf{x_i}^T \boldsymbol{\beta}) = \mu_i$ vem

$$\frac{\partial \ell(\beta, \alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{(1+\alpha \mu_i)^2}. \tag{3.15}$$

E

$$\frac{\partial \ell(\boldsymbol{\beta}, \alpha)}{\partial \alpha} = \frac{\partial \ell_1(\boldsymbol{\beta}, \alpha)}{\partial \alpha} + \frac{\partial \ell_2(\boldsymbol{\beta}, \alpha)}{\partial \alpha} + \frac{\partial \ell_3(\boldsymbol{\beta}, \alpha)}{\partial \alpha} + \frac{\partial \ell_4(\boldsymbol{\beta}, \alpha)}{\partial \alpha}$$

Onde

$$\frac{\partial \ell_1(\boldsymbol{\beta}, \alpha)}{\partial \alpha} = \frac{\sum_{i=1}^{n} \partial \left(y_i log(\frac{exp(\mathbf{x_i}^T \boldsymbol{\beta})}{1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta})}) \right)}{\partial \alpha} =$$

$$= \sum_{i=1}^{n} \frac{y_i[-exp(\mathbf{x_i}^T \boldsymbol{\beta}) exp(\mathbf{x_i}^T \boldsymbol{\beta})]}{(1+\alpha exp(\mathbf{x_i}^T))^2} \times \frac{(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta}))}{exp(\mathbf{x_i}^T \boldsymbol{\beta})}$$

$$= \sum_{i=1}^{n} -\frac{y_i exp(\mathbf{x_i}^T \boldsymbol{\beta})(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta}))}{(1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta}))^2};$$

$$\frac{\partial \ell_2(\boldsymbol{\beta}, \alpha)}{\partial \alpha} = \frac{\sum_{i=1}^{n} \partial \left((y_i - 1) log(1+\alpha y_i) \right)}{\partial \alpha} = \frac{(y_i - 1) y_i}{1+\alpha y_i};$$

$$\frac{\partial \ell_3(\boldsymbol{\beta}, \alpha)}{\partial \alpha} = \frac{\sum_{i=1}^{n} \partial \left(\frac{exp(\mathbf{x_i} \boldsymbol{\beta})(1+\alpha y_i)}{1+\alpha exp(\mathbf{x_i}^T \boldsymbol{\beta})} \right)}{\partial \alpha} =$$

$$= \sum_{i=1}^{n} \frac{[exp(\mathbf{x_i}^T\boldsymbol{\beta})y_i(1+\alpha exp(\mathbf{x_i}^T\boldsymbol{\beta})) - (exp(\mathbf{x_i}^T\boldsymbol{\beta})(1+\alpha y_i)exp(\mathbf{x_i}^T\boldsymbol{\beta}))]}{(1+\alpha exp(\mathbf{x_i}^T\boldsymbol{\beta}))^2} =$$

$$= \sum_{i=1}^{n} \frac{exp(\mathbf{x_i}^T\boldsymbol{\beta})(y_i - exp(\mathbf{x_i}^T\boldsymbol{\beta}))}{(1+\alpha exp(\mathbf{x_i}^T\boldsymbol{\beta}))^2};$$

$$\frac{\partial \ell_4(\boldsymbol{\beta},\alpha)}{\partial \alpha} = \frac{\partial \left(\sum_{i=1}^{n} log(y_i!)\right)}{\partial \alpha} = 0.$$

Portanto a derivada da função log-verosimilhança $\ell(\beta,\alpha)$ com respeito a α é dada por:

$$\frac{\partial \ell(\boldsymbol{\beta},\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \left(-\frac{y_i exp(\mathbf{x_i}^T\boldsymbol{\beta})}{(1+\alpha exp(\mathbf{x_i}^T\boldsymbol{\beta}))} \right) + \frac{(y_i-1)y_i}{1+\alpha y_i} - \frac{exp(\mathbf{x_i}^T\boldsymbol{\beta})(y_i - exp(\mathbf{x_i}^T\boldsymbol{\beta}))}{(1+\alpha exp(\mathbf{x_i}^T\boldsymbol{\beta}))^2} \right\}.$$

Como $\mu_i = exp\left(\mathbf{x_i}^T\boldsymbol{\beta}\right)$ e igualando a zero vem

$$\frac{\partial \ell(\boldsymbol{\beta},\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \left(-\frac{y_i\mu_i}{1+\alpha\mu_i} \right) + \frac{y_i(y_i-1)}{(1+\alpha y_i)} - \frac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^2} \right\} = 0. \tag{3.16}$$

Portanto o sistema de equações da verosimilhança para o modelo de regressão de Poisson Generalizado I é dado por:

$$\begin{cases} \frac{\partial \ell(\boldsymbol{\beta},\alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i-\mu_i)x_{ij}}{(1+\alpha\mu_i)^2} = 0, j = 1, ..., p \\ \frac{\partial \ell(\boldsymbol{\beta},\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \left(-\frac{y_i\mu_i}{1+\alpha\mu_i} \right) + \frac{y_i(y_i-1)}{(1+\alpha y_i)} - \frac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^2} \right\} = 0 \end{cases} \tag{3.17}$$

O parâmetro de dispersão α, também pode ser estimado usando o método dos momentos, isto é, igualando a estatística de Pearson Generalizada a *n-p* graus de liberdade,

$$\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{(1+\alpha)y_i} = n - p. \tag{3.18}$$

O método dos momentos não será aplicado no trabalho, e será aplicado apenas o método da máxima verosimilhança (MMV) na estimação dos parâmetros.

3.3.2 Modelo de Regressão de Poisson Generalizado II

Para introduzir as covariáveis, a média é incluída através da equação (3.10), substituindo a equação (3.10) na função (3.9) obtém-se o modelo de regressão de Poisson Generalizado II, dada por:

$$P(Y_i = y_i|\boldsymbol{\beta}, \alpha) = \begin{cases} exp(\mathbf{x_i}^T\boldsymbol{\beta})(exp(\mathbf{x_i}^T\boldsymbol{\beta}) + (\alpha-1)y_i)^{y_i-1}\times \\ \times\alpha^{-y_i}\dfrac{exp(-\alpha^{-1}(exp(\mathbf{x_i}^T\boldsymbol{\beta}) + (\alpha-1)y_i))}{y_i!}, \\ 0, y_i > m, \alpha < 1 \end{cases} \tag{3.19}$$

portanto a função de verosimilhança do modelo de regressão de Poisson Generalizado II, definida em Consul e Famoye (1992), é dado por:

$$L(\beta, \alpha) = \prod_{i=1}^{n} exp(\mathbf{x_i}^T\boldsymbol{\beta})\left(exp(\mathbf{x_i}^T\boldsymbol{\beta}) + (\alpha-1)y_i\right)^{y_i-1} \times \\ \times\alpha^{-y_i}\frac{exp(-\alpha^{-1}(\mu_i + (\alpha-1))y_i}{y_i!}, \tag{3.20}$$

e a função de log-verosimilhança é:

$$\ell(\beta, \alpha) = \sum_{i=1}^{n} \mathbf{x_i}^T\beta + log\left(exp(\mathbf{x_i}^T\boldsymbol{\beta}) + (\alpha-1)y_i\right)^{y_i-1} - \\ -y_i log(\alpha) - \alpha^{-y_i}exp(-\alpha^{-1}(\alpha-1)y_i) - log(y_i). \tag{3.21}$$

Portanto, os estimadores de máxima verosimilhança $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$, podem ser obtidas como solução do sistema de equações de verosimilhança conforme na equação (3.14). Derivando a função (3.21) em ordem a $\boldsymbol{\beta}$ e α vem

$$\frac{\partial\ell(\boldsymbol{\beta}, \alpha)}{\partial\beta_j} = \sum_{i=1}^{n} x_{ij} + \frac{(y_i-1)x_{ij}exp(\mathbf{x_i}^T\boldsymbol{\beta})}{exp(\mathbf{x_i}^T\boldsymbol{\beta})} - \alpha^{-1}x_{ij}exp(\mathbf{x_i}^T\boldsymbol{\beta}) =$$

$$= \sum_{i=1}^{n}\left\{\frac{1}{exp(\mathbf{x_i}^T\boldsymbol{\beta})} + \frac{y_i-1}{exp(\mathbf{x_i}^T\boldsymbol{\beta}) + (\alpha-1)y_i} - \alpha^{-1}\right\}x_{ij}exp(\mathbf{x_i}^T\boldsymbol{\beta})$$

Como $\mu_i = exp\left(\mathbf{x_i}^T\boldsymbol{\beta}\right)$ obtém-se;

$$\frac{\partial\ell(\beta, \alpha)}{\partial\beta_j} = \sum_{i=1}^{n}\left\{\mu_i^{-1} - \alpha^{-1} + \frac{y_i-1}{\mu_i(\alpha-1)y_i}\right\}\mu_i x_{ij}, j = 1, 2, ...p \tag{3.22}$$

e

$$\frac{\partial \ell(\boldsymbol{\beta},\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \frac{(y_i-1)y_i}{exp(\mathbf{x_i}^T\boldsymbol{\beta})+(\alpha-1)y_i} - \frac{y_i}{\alpha} - \frac{[y_i\alpha-(exp(\mathbf{x_i}^T\boldsymbol{\beta})+(\alpha-1)y_i)]}{\alpha^2} =$$

$$= \sum_{i=1}^{n} \frac{(y_i-1)y_i}{exp(\mathbf{x_i}^T\boldsymbol{\beta})+(\alpha-1)y_i} - y_i\alpha^{-1} + (exp(\mathbf{x_i}^T\boldsymbol{\beta})-y_i)\alpha^{-2}$$

Como $\mu_i = exp\,(\mathbf{x_i}^T\boldsymbol{\beta})$ logo tem-se;

$$\frac{\partial \ell(\beta,\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \frac{y_i(y_i-1)}{\mu_i+(\alpha-1)y_i} - y_i\alpha^{-2} + (\mu_i-y_i)\alpha^{-2}. \tag{3.23}$$

O sistema de equações de verosimilhança para o modelo de regressão de Poisson Generalizada II é dado por:

$$\begin{cases} \frac{\partial \ell(\boldsymbol{\beta},\alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \mu_i^{-1} - \alpha^{-1} + \frac{y_i-1}{\mu_i(\alpha-1)y_i} \right\} \mu_i x_{ij} = 0 \\ \frac{\partial \ell(\boldsymbol{\beta},\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \frac{y_i(y_i-1)}{\mu_i+(\alpha-1)y_i} - y_i\alpha^{-2} + (\mu_i-y_i)\alpha^{-2} = 0 \end{cases} \tag{3.24}$$

3.3.3 Estimação do parâmetro β

Para a resolução numérica das equações de verosimilhança, utilizam-se o algorítmico de *Newton-Raphson* e o método de *scores* de *Fisher*. A equação iterativa do método de regressão dos mínimos quadrados ponderados pode ser escrito como,

$$\boldsymbol{\beta}_r = \boldsymbol{\beta}_{(r-1)} + I^{-1}_{(r-1)}\boldsymbol{s}_{(r-1)}, \tag{3.25}$$

onde $\boldsymbol{\beta}_r$ e $\boldsymbol{\beta}_{(r-1)}$ são os vectores do parâmetro β na iteração de ordem r e *(r-1)*, $I(\cdot)^{-1}$ é a inversa que se supõe exista da matriz de informação de *Fisher* negativa das segundas derivadas da função do logaritmo da verosimilhança avaliada em $\boldsymbol{\beta}_{(r-1)}$ e $\boldsymbol{s}_{(r-1)}$ é o vector das primeiras derivadas da função logaritmo da verosimilhança avaliada em $\boldsymbol{\beta}_{(r-1)}$.

Para a Distribuição de Poisson Generalizada a inversa da matriz de informação de *Fisher* de dada por:

$$I^{-1}_{r-1} = (\boldsymbol{X}^T\boldsymbol{W}_{(r-1)}\boldsymbol{X})^{-1}, \tag{3.26}$$

onde, $\boldsymbol{X}$ é a matriz de especificação do modelo, $\boldsymbol{W}$ é a matriz diagonal de ordem n cujo i-ésimo elemento é:

i)Para o caso do modelo de regressão de Poisson Generalizado I

$$w_i^{DPGI} = \frac{\mu_i}{(1+\alpha\mu_i)^2}$$

ii)Para o caso do modelo de regressão de Poisson Generalizado II

$$w_i^{DPGII} = \mu_i$$

Tem-se assim que:

i) Para o caso do modelo de regressão de Poisson Generalizado I

$$I = i_{ij} = \sum_{i=1}^{n} \frac{\mu_i x_{ij} x_{is}}{(1+\alpha\mu_i)^2}$$

ii)Para o caso do modelo de regressão de Poisson Generalizado II

$$I = i_{ij} = \sum_{i=1}^{n} \mu_i x_{ij} x_{is}$$

como

$$\boldsymbol{s}_{r-1} = \boldsymbol{X}^T \boldsymbol{W}_{(r-1)} \boldsymbol{k}_{r-1}, \tag{3.27}$$

onde, $\boldsymbol{k}$ é o vector cujo i-ésima linha é $k_i = \dfrac{y_i - \mu_i}{\mu_i}$ substituindo as equações (3.26) e (3.27) na equação (3.25) obtém-se:

$$\boldsymbol{\beta}_r = \boldsymbol{\beta}_{(r-1)} + (\boldsymbol{X}^T \boldsymbol{W}_{(r-1)} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{W}_{(r-1)} \boldsymbol{k}_{r-1}). \tag{3.28}$$

A equação (3.28) é usada para obter as estimativas da máxima verosimilhança do parâmetro $\boldsymbol{\beta}$ do modelo de regressão de Poisson Generalizada I e do modelo de regressão de Poisson Generalizada II.

3.3.4 Estimação do parâmetro α

A estimativa da maxima verosimilhança do α pode ser resolvido aplicando o algoritmo de *Newton-Raphson*:

$$\alpha_{(r)} = \alpha_{(r-1)} - \frac{\dfrac{\partial \ell(\beta, \alpha)}{\partial \alpha}}{\dfrac{\partial^2 \ell(\boldsymbol{\beta}, \alpha)}{\partial \alpha^2}} \tag{3.29}$$

Para o modelo de regressão de Poisson Generalizado I, $\dfrac{\partial \ell(\beta,\alpha)}{\partial \alpha}$ é dada em (3.16) e

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\alpha)}{\partial \alpha^2} = \sum_{i=1}^{n} \left\{ \frac{y_i \mu_i^2}{(1+\alpha\mu_i)^2} - \frac{y_i^2(y_i-1)}{(1+\alpha y_i)^2} + \frac{2\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^3} \right\} \tag{3.30}$$

Substituindo as equações (3.16) e (3.30) na equação (3.29) obtém-se:

$$\alpha_{(r)} = \alpha_{(r-1)} - \frac{\sum_{i=1}^{n} \left\{ \left(-\dfrac{y_i\mu_i}{1+\alpha\mu_i} \right) + \dfrac{y_i(y_i-1)}{(1+\alpha y_i)} - \dfrac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^2} \right\}}{\sum_{i=1}^{n} \left\{ \dfrac{y_i\mu_i^2}{(1+\alpha\mu_i)^2} - \dfrac{y_i^2(y_i-1)}{(1+\alpha y_i)^2} + \dfrac{2\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^3} \right\}} \tag{3.31}$$

A equação (3.31) é utilizada para obter as estimativa de máxima verosimilhança do parâmetro α no modelo de regressão de Poisson Generalizado I.

O processo para encontrar a estimativa de $\hat{\alpha}$, para o modelo de regressão de Poisson Generalizada II, não envolve qualquer iteracção, a estimativa pode ser obtida directamente pela equação:

$$\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{\alpha^2\mu_i} = n-p$$

onde

$$\alpha = \sqrt{\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{\mu_i(n-p)}} \tag{3.32}$$

3.3.5 Matriz de variância-covâriancia de $\hat{\beta}$

A matriz de variância-covariância, $var(\hat{\beta})$, para o modelo de regressão de Poisson Generalizado I é igual ao modelo de regressão de Poisson, isto é,

$$var(\hat{\beta}) = (\mathbf{X_i}^T W \mathbf{X})^{-1}$$

Onde W é a matriz diagonal de ordem n, cujo i-ésimo elemento é w_i dada por $w_i^{DPGI} = \dfrac{\mu_i}{(1+\alpha\mu_i)^2}$.

Para o modelo de regressão de Poisson Generalizado II, a matriz variância-covâriancia é dado por:

$$var(\hat{\beta}) = (\alpha)^2 (\mathbf{X_i}^T W \mathbf{X})^{-1}$$

Onde W é a matriz diagonal de ordem n, cujo i-ésimo elemento é w_i dada por $w_i^{DPGII} = \mu_i$

3.3.6 Teste da Razão da Verosimilhança

Para efectuar o teste de comparação entre o modelo de regressão de Poisson e o modelo de regressão de Poisson Generalizado I, construimos as seguintes hipóteses:

$H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$,

Para o modelo de regressão de Poisson e modelo de regressão de Poisson Generalizado II, o teste de hipótese é:

$$H_0 : \alpha = 1 versus H_1 : \alpha \neq 1$$

.

A estatística de teste para os dois casos é dado por:

$$T = -2\left(\ell_0 - \ell_1\right),$$

onde ℓ_0 é a função de verosimilhança do modelo de regressão de Poisson e ℓ_1 é a função de verosimilhança do modelo de regressão de Poisson Generalizado I (ou do modelo de regressão de Poisson Generalizado II). Sob, hipótese nula, T segue uma distribuição qui-quadrado com 1 (um) grau de liberdade (Wang e Famoye, 1997).

3.4 Revisão da literatura sobre aplicações

O modelo de referência para dados de contagem é o modelo de regressão de Poisson. Esta distribuição foi desenvolvida por Siméon-Denis Poisson, em 1837, e publicada em seu trabalho *Recheches sur la probabilité des jugements em matières criminelles et matière civile* (Poisson, 1837).

Segundo Ramalho (1996), é comum, quando estamos a trabalhar com dados de contagem, iniciarmos a estimação dos parâmetros por meio de um modelo de regressão de Poisson, devido à sua simplicidade. Neste caso, a variável dependente de um modelo de regressão de Poisson deve seguir uma distribuição de Poisson com o valor esperado igual à variância. Entretanto, de acordo com de Souza Tadano *et al.* (2009), esta propriedade é frequentemente violada, já que é comum a existência de sobredispersão (ou subdispersão), ou, seja é frequente que a variância da variável aleatória dependente seja maior (ou menor) que o seu valor esperado. Nestes casos

trabalha-se com Modelos de Regressão Binomial Negativa, Modelo de Regressão de Poisson Generalizados ou Modelos de Quasi-verosimilhança.

É comum encontrarmos exemplos de aplicação destes modelos de regressão em economia, finanças, demografia, ecologia e meio-ambiente, ciências actuariais, medicina e veterinária entre outras áreas de conhecimentos.

Ismail e Jemain (2007) aplicaram o modelo de regressão de Poisson, o modelo de regressão Binomial Negativa e o modelo de regressão de Poisson Generalizado para estudar o número de sinistros do conjunto de dados do trabalho de Bailey e Simon, 1960 que apresentavam sobredispersão. Os mesmos autores concluíram que o modelo de regressão de Poissn Generalizado é o mais adequado para modelar os dados, por apresentar menor valor de AIC e BIC.

Cameron e Trivedi (1986) utilizaram o modelo de regressão Binomial Negativa para analisar o número de consultas médicas nas últimas 2 semanas para uma amostra de 5190 pacientes, no período de 1977 – 19778, à luz de um modelo económico de determinação conjunta da utilização do serviço de saúde e escolha do seguro de saúde. Os resultados demonstraram que o modelo de regressão Binomial Negativa é o modelo adequado para modelar esses dados que apresentavam sobredispersão.

Ozuna e Gomez (1995) aplicaram o Modelo de regressão de Poisson para estudar o número de passeios de barco recreativos para o lago Somerville, Texas, em 1980, com base num inquérito administrado a 2000 proprietários de barcos de lazer registrados em 23 municípios no leste do Texas. Concluíram que o modelo de regressão de Poisson não é adequado devido à sobredispersão dos dados e excesso de zeros. Portanto os mesmos autores aplicaram, os modelos de regressão Binomial Negativa e Binomial Negativa com inflação de zeros e concluíram que os dois modelos são adequados.

Hinde (1982) considerou um conjunto de dados que descreve o número de falhas em rolos de tecidos. Observou-se que os dados eram sobredispersos e, portanto, o uso do modelo de regressão de Poisson não era apropriado. Hinde aplicou o modelo de regressão de Poisson Generalizado para ajustar os dados e concluiu que esse modelo de regressão foi melhor que o modelo de regressão de Poisson na modelação dos dados.

Ismail e Zamani (2013) analisaram a ocorrências de acidentes rodoviários em carros particulares da base de dados Malásia, com objectivo de desenvolver modelos para descrever e estimar com precisão o número de acidentes. Os testes de dispersão indicaram a sobredispersão dos dados. Os mesmos autores utilizaram os modelos

de regressão de Poisson, de Poisson Generalizado e Binomial Negativa concluíram que embora as estimativas dos parâmetros são iguais nos três modelos, a qualidade de ajustamento dos modelos aos dados é melhor no modelo de regressão de Poisson Generalizado e no modelo de regressão Binomial Negativa.

Para modelar o número de casos de cancro do colo do útero, Melliana *et al.* (2013), utilizaram o modelo de regressão de Poisson e verificaram que há indícios de sobredispersão. Os mesmos autores aplicaram os modelos de regressão de Poisson Generalizado e de regressão Binomial Negativa para modelar os dados sobredispersos. No entanto, aplicando o critério de informação AIC na comparação dos dois modelos, verificaram que o modelo indicado para modelar os dados que analisaram era o modelo de Regressão Binomial Negativa, por apresentar menor valor AIC.

Capítulo 4

Análise e Modelação de Dados

Neste capítulo, o objetivo é aplicar modelos de regressão de Poisson Generalizado I, na análise de dados de contagem, em duas bases de dados diferentes. Na primeira base de dados estudam-se dados de contagem com subdispersão e na segunda analisam-se dados de contagem com sobredispersão.

Inicialmente serão apresentadas as duas bases de dados, de seguida, será apresentada e realizada uma análise descritiva para cada conjunto de dados e por último estimam-se modelos de regressão.

4.1 Base de Dados *Takeoverbids*

A base de dados *TakeoverBids*, obtida da biblioteca *countreg* do *software* R (Mott *et al.*, 1967), descreve as 126 empresas norte-americanas que foram alvo de ofertas públicas de aquisição durante o período de 1978-1985 e foram efetivamente adquiridas durante um período de 52 semanas após a oferta inicial. Esses dados, estudados originalmente em Cameron e Johansson (1997), são ainda usados em Cameron e Trivedi (2013), Jaggia e Thosar (1993) bem como em Sáez-Castillo e Conde-Sánchez (2013).

O conjunto de dados *TakeoverBids* é constituído por 126 observações e as seguintes variáveis :

- *Bids*: número de ofertas públicas de aquisição após a oferta inicial recebida pela empresa (variável resposta);
- *Legalrest*: indica se a gerência da empresa responde por ação judicial e está codificada como 1-sim, 2-não;
- *Realrest*: indica se administração da empresa dá metas e propõe mudanças na estrutura de ativos e está codificada como 1-sim, 2-não;

• *Finrest*: indica se a administração da empresa dá metas e propõe mudanças na estrutura de propriedades e está codificada como 1-sim, 2-não;

• *Whiteknight*: indica se a gerência da empresa convidou terceiros para oferta e está codificada como 1-sim, 2-não;

• *Insthold*: percentagem de ações detidas pelas empresas;

• *Size*: Valor total dos ativos em biliões de dólares;

• *Bidpremium*: preço do prémio dividido pelo preço 14 dias úteis antes da oferta;

• *Regulation*: indica se há intervenção de reguladores federais e está codificada como 1-sim, 2-não.

4.1.1 Análise Descritiva univariada

Nesta secção será apresentada a análise descritiva das variáveis quantitativas e das variáveis qualitativas presentes na base de dados. Para as variáveis quantitativas contínuas (*Insthold*, *Size* e *Bidpremium*) utilizam-se as medidas de tendência central e as medidas de dispersão, bem como os gráficos caixa com bigodes, e para as variáveis explicativas qualitativas (*Legalrest*, *Realrest*, *Finrest*, *Whiteknight* e *Regulation*) apresentam-se gráficos de barras e tabelas de frequências.

Número de ofertas públicas de aquisição (*Bids*)

A variável número de ofertas públicas de aquisição após a oferta inicial recebida pela empresa (*Bids*) é a variável resposta e na Tabela 4.1 é apresentada a tabela das frequências desta variável. Na Tabela 4.2 mostra as medidas de tendência central e as medidas de dispersão. A Figura 4.1 mostra o gráfico de barras da referida variável e observa-se que o número máximo de ofertas públicas de aquisição após a oferta inicial é 10 e a média das ofertas públicas após a oferta inicial recebida pela empresa durante o período de 1978-1985 é de 1,74. Observa-se 75% do número de ofertas públicas de aquisição após a oferta inicial recebida pela empresa são menores ou iguais a 2. Nota-se alta variabilidade desta variável ($CV = 82,39\%$).

Percentagem de ações detidas pelas empresas (*Insthold*)

A Figura 4.2 apresenta o gráfico da caixa com bigodes da variável *Insthold*. A Tabela 4.3 mostra as medidas de tendência central e as medidas de dispersão da referida variável. Observa-se que 25% da percentagem de acções detidas pelas instituições são menores ou iguais a $8,2\%$ e 25% são superiores ou iguais a $38,7\%$. A média da percentagem de acções detidas pelas instituições durante o período de 1978-1985 é de $25,2\%$.

Tabela 4.1: Tabela de frequências do número de ofertas públicas durante o período de 1978-1985

Bids	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
0	9	7,14	9	7,14
1	63	50,00	72	57,14
2	31	24,60	103	81,75
3	12	9,52	115	91,27
4	6	4,76	121	96,03
5	1	0,79	122	96,83
6	2	1,59	124	98,41
7	1	0,79	125	99,21
10	1	0,79	126	100,00

Tabela 4.2: Estatísticas descritivas da variável *Bids*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0	1,74	1	1	2	10	1,43	82,184

Tabela 4.3: Estatísticas descritivas da variável *Insthold*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação (%)
0,00	0,252	0,082	0,205	0,387	0,904	0,186	73,809

Figura 4.1: Gráfico de barras da variável *Bids*

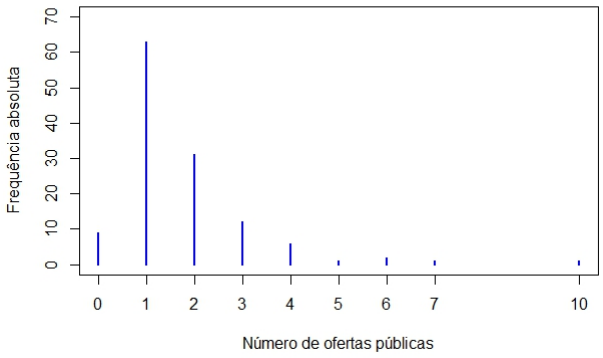
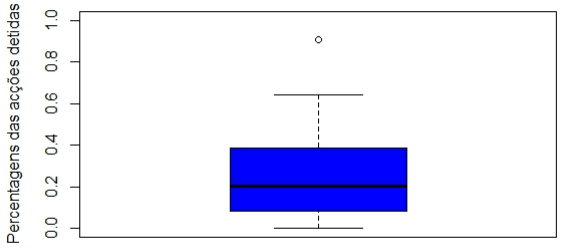


Figura 4.2: Caixa com bigodes da variável *Insthold*



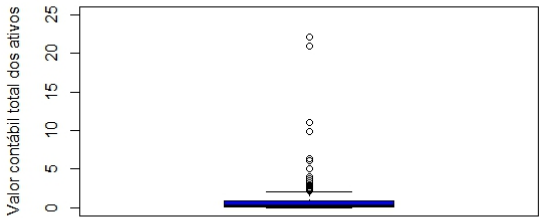
Valor total dos ativos (*Size*)

A Figura 4.3 apresenta o gráfico da caixa com bigodes da variável *Size* e na Tabela 4.4 é apresentado as estatísticas descritivas da referida variável. Observa-se que o mínimo e o máximo do valor total dos ativos é de 0,018 e 22,169 biliões de dólares, respectivamente. 25% do valor total dos ativos são menores ou iguais 0,108 biliões de dólares e 50% dos valores estão compreendidos entre 0,108 e 0,883 biliões de dólares. Da Figura 4.3 observa-se a existência de um grande número de observações outliers.

Tabela 4.4: Estatísticas descritivas da variável *Size*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação (%)
0,018	0,219	0,108	0,205	0,249	0,883	3,097	1413,984

Figura 4.3: Caixa com bigodes da variável *Size*



Preço do prémio dividido pelo preço 14 dias úteis antes da oferta (*Bidpremium*)

A Figura 4.4 apresenta o gráfico da caixa com bigodes da variável *Bidpremium*. A Tabela 4.5 mostra o resumo estatístico desta mesma variável. Observa-se que a mediana desta variável é de 1,328 e o valor mínimo e máximo é 0,943 e 2,066 respectivamente, e 50% dos preços dos prémios dividido pelo preço 14 dias úteis antes da oferta ao longo do período de 1978-1985 estão compreendidos entre 1,218 e 1,430, respectivamente.

Tabela 4.5: Estatísticas descritivas da variável *Bidpremium*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação (%)
0,943	1,347	1,218	1,328	1,430	2,066	0,189	14,031

Gerência da empresa responde por ação judicial (*Legalrest*)

A Figura 4.5 apresenta o gráfico de barras da variável *Legalrest* e a tabelas das frequências da variável *Legalrest* está apresentada na Tabela 4.6. Pode-se observar que há menos empresas a responder por acção judicial.

Tabela 4.6: Tabela de frequências da variável *Legalrest*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	54	42,86	54	42,86
2 - Não	72	57,14	126	100,00

Figura 4.4: Caixa com bigodes da variável *Bidpremium*

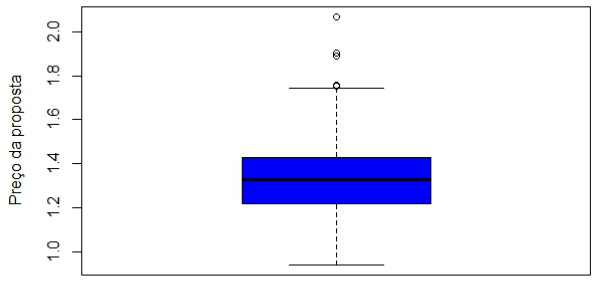
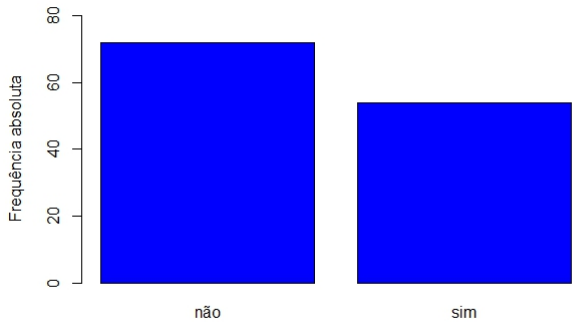


Figura 4.5: Gráfico de barras da variável *Legalrest*



Administração da empresa dà metas e propõe mudanças na estrutura de ativos (*Realrest*)

A Figura 4.6 apresenta o gráfico de barras da variável *Realrest*. A Tabela 4.7 mostra a tabela das frequências da referida variável. As empresas mais frequentes são as empresas que não propuseram mudanças na estrutura de ativos (103 empresas não propuseram mudanças).

Administração da empresa dà metas e propõe mudanças na estrutura de propriedades (*Finrest*)

A Figura 4.7 apresenta o gráfico de barras da variável *Finrest*. A Tabela 4.8 mostra a tabela das frequências da mesma variável. Pode-se observar que as empresas que propuseram mudanças nas estruturas de propriedades são apenas 13 empresas.

Tabela 4.7: Tabela de frequências da variável *Realrest*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	23	18,25	23	18,25
2 - Não	103	81,75	126	100,00

Figura 4.6: Gráfico de barras da variável *Realrest*

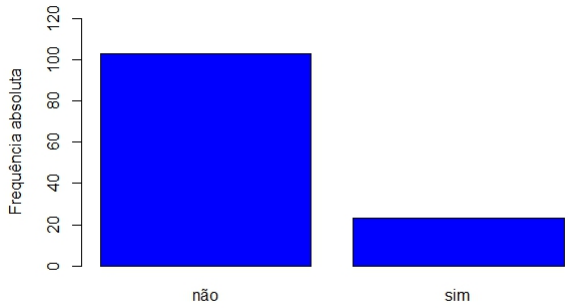


Tabela 4.8: Tabela de frequências da variável *Finrest*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	13	10,32	13	10,32
2 - Não	113	89,68	126	100,00

Gerência da empresa convidou terceiros (*Whiteknight*)

A Figura 4.8 apresenta o gráfico de barras da variável *Whiteknight* e a Tabela 4.9 mostra a tabela das frequências da referida variável. As empresas mais frequentes são as empresas que convidaram terceiros (75 empresas convidaram terceiros).

Tabela 4.9: Tabela de frequências da variável *Whiteknight*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	75	59,52	75	59,52
2 - Não	51	40,48	126	100,00

Figura 4.7: Gráfico de barras da variável *Finrest*

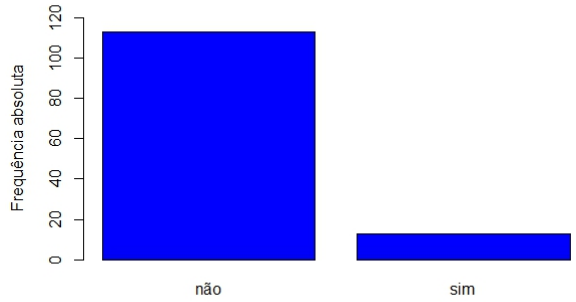
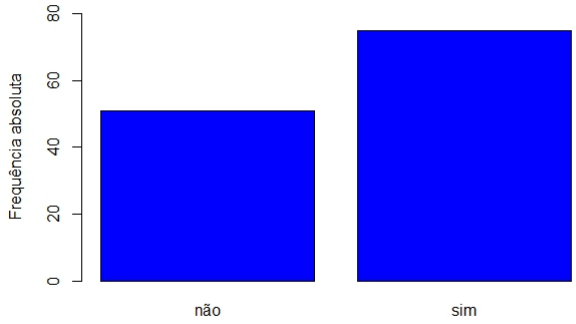


Figura 4.8: Gráfico de barras da variável *Whiteknight*



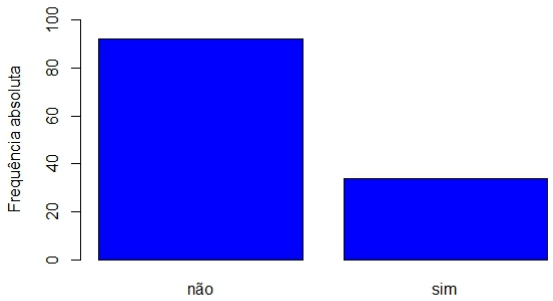
Intervenção dos reguladores federais (*Regulation*)

Por fim, a Figura 4.9 apresenta o gráfico de barras da variável *Regulation* e a Tabela 4.10 mostra a tabela das frequências da referida variável. Observa-se que há poucas com intervenções dos reguladores federais (34 empresas não têm).

Tabela 4.10: Tabela de frequências da variável *Regulation*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	34	26,98	34	26,98
2 - Não	92	73,01	126	100,00

Figura 4.9: Gráfico de barras da variável *Regulation*



Associação entre a variável resposta e as variáveis explicativas

Na Tabela 4.11 é apresentado o coeficiente de correlação de Spearman entre a variável resposta *Bids* e cada uma das variáveis explicativas quantitativas, bem como o p-valor associado ao teste de hipótese para avaliar se o coeficiente de correlação de *Spearman* é zero entre as variáveis.

No caso da associação entre a variável(*Bids*) e as variáveis explicativas quantitativas *Insthold* e *Size*, os p-valores associados ao teste de hipótese do coeficiente de correlação de Spearman das variáveis *Insthold* e *Size* são superiores ao nível de significância de 5%, portanto não se rejeita a hipótese de não existir associação entre a variável *Insthold* e a variável resposta, e entre a variável *Size* e a variável resposta. Para um nível de significância de 5%, rejeita-se a hipótese de não existir associação entre a variável *Bidpremium* e a variável resposta, ou seja, a variável explicativa quantitativa *Bidpremium* apresenta uma correlação significativa negativa com a variável *Bids* , isto é, existe uma tendência para número de ofertas públicas de aquisição após a oferta inicial diminuir com a diminuição dos preços dos prémios dividido pelo preço 14 dias úteis antes da oferta recebida pela empresa.

Tabela 4.11: Coeficiente de correlação de Spearman entre as variáveis explicativas e a variável *Bids*

Variável	Coeficiente	P-valor
Insthold	-0,029	0,741
Size	0,095	0,289
Bidpremium	-0,186	0,037

A Tabela 4.12 apresenta os valores das estatísticas de teste e os respectivos

p-valores do teste de Mann-Whitney quando se pretende comparar a distribuição da variável *Bids* nas categorias das variáveis explicativas. Para um nível de significância de 5% não se rejeita a hipótese nula sugerindo que não existe diferença entre as categorias das variáveis *Realrest*, *Finrest* e a variável *Regulation* em termos da distribuição do número de ofertas públicas da aquisição após a oferta inicial recebida pela empresa. Para um nível de significância de 5% rejeita-se a hipótese nula da igualdade das distribuições, pode se afirmar que existe evidencia estatística de que entre as categorias das variáveis *Legalrest*, *Whiteknight* ocorrem diferenças significativas em termos da distribuição do número de ofertas públicas de aquisição após a oferta inicial recebida pela empresa.

Tabela 4.12: Teste de Mann-Whitney entre as variáveis explicativas e a variável resposta *Bids*

Variável	Estatística de teste	P-valor
Legalrest	1546,5	0,035
Realrest	1301,5	0,427
Finrest	697,5	0,752
Whiteknight	1177,5	$< 0,001$
Regulation	1444	0,479

4.1.2 Escolha do modelo estatístico

Além das 8 variáveis explicativas descritas na Secção 4.1, inclui-se ainda a variável explicativa *Sizeq* (o quadrado do valor total dos ativos em biliões de dólares), de acordo com a sugestão de Jaggia e Thosar (1993).

Para explicar o número de ofertas públicas de aquisição após a oferta inicial recebida pela empresa, utiliza-se inicialmente o modelo de regressão de Poisson por ser o mais usado para modelar dados de contagem. A Tabela 4.13 apresenta o modelo de regressão de Poisson com todas as variáveis explicativas (*Modelo 1*).

Modelo de Regressão de Poisson (*Modelo 1*)

$$Bids_i \sim Poi\left(\mu_i\right), i=1,.....,126$$

$$log(\mu_i) = \beta_0 + \beta_1 * Legalrest_i + \beta_2 * Realrest_i + \beta_3 * Finrest_i +$$
$$+ \beta_4 * Whiteknight_i + \beta_5 * Bidpremium_i + \beta_6 * Insthold_i + \beta_7 * Size_i + \beta_8 * Sizeq_i +$$
$$+ \beta_9 * Regulation_i,$$

Tabela 4.13: Estimativas dos parâmetros do modelo de regressão de Poisson com todas variáveis explicativa

Variáveis Explicativas	Estimativas dos parâmetros	Erro padrão	Estatística de teste	P-valor
constante	0,986	0,534	1,847	0,065
Legalrest	0,260	0,151	1,723	0,085
Realrest	-0,196	0,193	-1,016	0,309
Finrest	0,074	0,217	0,342	0,732
Whiteknight	0,481	0,159	3,030	0,002
Bidpremium	-0,678	0,377	-1,799	0,072
Insthold	-0,362	0,424	-0,853	0,394
Size	0,179	0,060	2,974	0,003
Sizeq	-0,008	0,003	-2,425	0,020
Regulation	-0,029	0,161	-0,183	0,855

A multicolinearidade é um problema no ajuste do modelo que pode causar impactos nas estimativas dos parâmetros. Pode-se diagnosticar a existência de multicolinearidade usando a estatística VIF (*Variance Inflation Factor*). Na Tabela 4.14 são apresentadas os valores da estatística VIF para cada uma das variáveis existente no conjunto de dados *Takeoverbids*. Nota-se que o VIF de todas as variáveis são menores que 5 ($VIF < 5$), logo não há problemas de multicolinearidade.

Tabela 4.14: Valores do VIF do modelo de regressão de Poisson

Variáveis Explicativas	VIF (*Variance Inflation Factor*)
Legalrest	1,231
Realrest	1,123
Finrest	1,113
Whiteknight	1,091
Bidpremium	1,025
Insthold	1,268
Size	1,211
Regulation	1,214

Aplicando o método de seleção de variáveis *Backward*, isto é, retirando sucessivamente as variáveis, estatisticamente não significativas, obtém-se o *Modelo 2*. A Tabela 4.15 apresenta o modelo de regressão de Poisson ajustado com todas as variáveis estatisticamente significativas.

Modelo de Regressão de Poisson ajustado (*Modelo 2*)

$$Bids_i \sim Poi\,(\mu_i)\,,\ i{=}1{,}.....126$$

$$log(\mu_i) = \beta_0 + \beta_1 * Whiteknight_i + \beta_2 * Size_i + \beta_3 * Sizeq_i$$

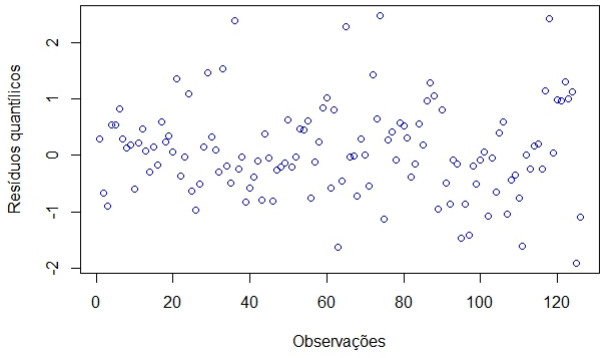
Tabela 4.15: Estimativas dos parâmetros do modelo de regressão de Poisson com todas variáveis estatisticamente significativas

Variáveis Explicativas	Estimativas dos parâmetros	Erro Padrão	Estatística de teste	P-valor
Constante	0,045	0,135	0,333	0,739
Whiteknight	0,553	0,152	3,633	$< 0,001$
Size	0,158	0,049	3,160	0,002
Sizeq	-0,007	0,003	-2,416	0,016

Análise de resíduos

O gráfico apresentado na Figura 4.10 representa o resíduos quantílicos *versus* índice das observações. Observa-se que os resíduos quantílicos se distribuem de forma uniforme em torno de zero pelo que sugere a independência dos erros.

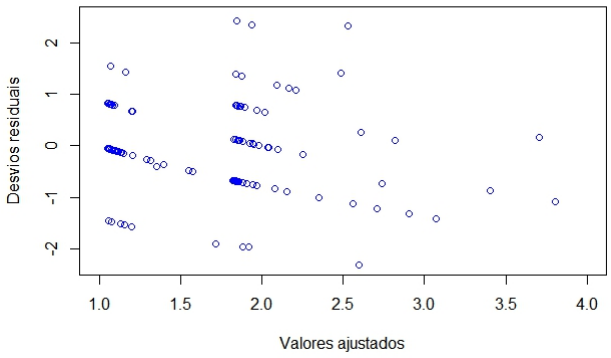
Figura 4.10: Gráfico dos resíduos quantílicos *versus* índices das observações do *Modelo 2*



A Figura 4.11 apresenta o gráfico desvios residuais *versus* valores ajustados. Pode-se observar que os resíduos estão aleatoriamente distribuídos em torno do resíduo zero, garantindo a homocedasticidade dos erros.

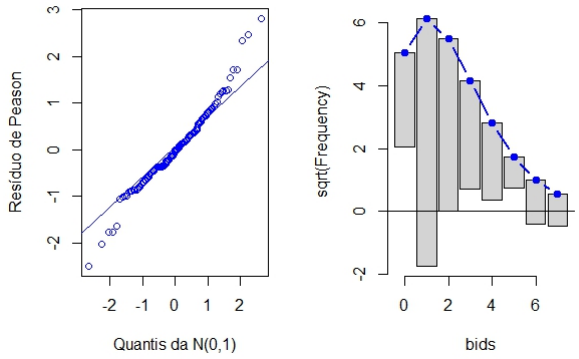
Na Figura 4.12 apresentam-se os gráficos dos resíduos de Pearson *versus* quantis da distribuição normal e o *rootogram*. Pode se constatar no gráfico dos resíduos de Pearson *versus* quantís da *N(0,1)*, o afastamento dos resíduos em torno da linha recta, sugerindo a plausibilidade da suposição dos resíduos não seguirem uma

Figura 4.11: Gráfico dos desvios residuais *versus* valores ajustados do *Modelo 2*



distribuição normal. Relativamente ao *rootogram* observa-se que a primeira barra da distribuição do número de oferta públicas de aquisição após a oferta inicial em relação a recta das ordenadas inicia-se em 2 notando o afastamento do eixo das abcissas, ou seja, ocorre uma distorção maior de resíduos nesta faixa. No entanto, também se verifica que à medida que as contagens aumentam, as barras não se aproximam do eixo das abcissas.

Figura 4.12: Gráficos dos resíduos de Pearson *versus* quantís da N(0,1) e o *rootogram* referentes ao *Modelo 2*



Sobredispersão e Subdispersão

Para verificar se há indícios de sobredispersão no modelo ajustado, pode-se utilizar a estatística baseada no desvio residual (D), dada pela seguinte expressão (Zuur

et al., 2009):

$$\hat{\alpha} = \frac{D}{n-p},$$

onde n é o número total de observações que constitui a amostra e p é o número de variáveis explicativas. No *Modelo 2* ajustado com a regressão de Poisson, verifica-se que $\hat{\alpha} = 0.787$, logo não há indícios de sobredispersão.

Segundo Cameron e Trivedi (2005), para identificar a sobredispersão ou subdispersão dos dados, podem-se utilizar os testes de hipóteses especificado no Capítulo 2, na Secção 2.8.3. O teste da dispersão testa a hipótese nula de equidispersão num modelo de regressão de Poisson contra a alternativa de sobredispersão ou subdispersão. Para o *Modelo 2* ajustado, observa-se que o parâmetro $\hat{\alpha} = -0.252 < 0$ e para um nível de significância de 5%, rejeita-se a hipótese nula (p-valor=0,01), concluindo que há indícios de subdispersão.

No entanto, embora o modelo de regressão Binomial Negativa seja adequado em situações de sobredispersão, optou-se por ajustar um modelo de regressão Binomial Negativa para comparar com o modelo de regressão de Poisson. Os resultado do modelo Binomial Negativa (*Modelo 3*) com todas as variáveis explicativas estão apresentados na Tabela 4.16

Modelo de regressão Binomial Negativa (*Modelo 3*)

$$Bids_i \sim BinNeg\left(\mu_i, \alpha\right), i{=}1,.....126$$

$$log(\mu_i) = \beta_0 + \beta_1 * Legalrest_i + \beta_2 * Realrest_i + \beta_3 * Finrest_i +$$
$$+ \beta_4 * Whiteknight_i + \beta_5 * Bidpremium_i + \beta_6 * Insthold_i + \beta_7 * Size_i + \beta_8 * Sizeq_i +$$
$$+ \beta_9 * Regulation_i$$

Aplicando o método de seleção de variáveis *Backward*, retirando sucessivamente as variáveis estatisticamente não significativas, obtém-se o *Modelo 4*.

Modelo de regressão Binomial Negativa (*Modelo 4*)

$$Bids_i \sim BinNeg\left(\mu_i, \alpha\right), i{=}1,.....126$$

$$log(\mu_i) = \beta_0 + \beta_1 * Whiteknight_i + \beta_3 * Size_i + \beta_4 * Sizeq_i$$

Na Tabela 4.17 apresentam-se as estimativas dos parâmetros do modelo de regressão Binomial Negativa, com as variáveis estatisticamente significativas. Ajustado o modelo de regressão Binomial Negativa (*Modelo 4*), observa-se que não houve melhoria nos dois gráficos representados na Figura 4.13. No *rootogram* onde houve distorção dos resíduos não melhorou o ajuste no modelo de regressão Binomial Negativa (*Modelo 4*). O gráfico dos resíduos de Pearson *versus* quantís também

Tabela 4.16: Estimativas dos parâmetros do modelo de regressão Binomial Negativa com todas as variáveis explicativas

Variáveis explicativas	Estimativas dos parâmetros	Erro padrão	Estatística de teste	P-valor
Constante	0,986	0,534	1,847	0,065
Legalrest	0,260	0,151	1,723	0,085
Realrest	-0,196	0,193	-1,016	0,309
Finrest	0,074	0,217	0,342	0,732
Whiteknight	0,481	0,159	3,030	0,002
Bidpremium	-0,678	0,377	-1,799	0,072
Insthold	-0,362	0,424	-0,853	0,394
Size	0,179	0,060	2,974	0,003
Sizeq	-0,008	0,003	-2,425	0,015
Regulation	-0,029	0,161	-0,183	0,855
$\hat{\alpha} = 0.778$				

não melhorou, verificando novamente o afastamento dos resíduos em torno da linha recta. Relativamente aos gráficos representados na Figura 4.14, no primeiro gráfico dos resíduos quantílicos *versus* índice das observações observa-se que os resíduos quantílicos distribuem-se de forma uniforme em torno da reta do resíduo zero e no segundo gráfico dos desvios residuais *versus* valores ajustados os resíduos estão aleatoriamente distribuídos em torno da reta do resíduo zero, e verifica-se o pressuposto da homocedasticidade.

Tabela 4.17: Estimativas dos parâmetros do modelo de regressão da Binomial Negativa com todas variáveis estatisticamente significativas

Variáveis Explicativas	Estimativas dos parâmetros	Erro padrão	Estaticista de teste	P-valor
Constante	0,045	0,135	0,333	0,739
Whiteknight	0,553	0,152	3,633	0,002
size	0,158	0,049	3,160	0,002
Sizeq	-0,007	0,003	-2,416	0,016
$\hat{\alpha} = 0.778$				

Realizou-se o teste de *Voung*, entre o modelo de regressão de Poisson e o modelo de regressão Binomial Negativa e Conclui-se que não existe diferença significativa entre os modelos. Os resultados do teste estão apresentados na Tabela 4.18.

De seguida, estuda-se o modelo de regressão de Poisson Generalizado, uma vez que os dados apresentam subdispersão.

Figura 4.13: Gráficos dos resíduos de Pearson *versus* quantís da N(0,1) e o *rootogram* referentes ao *Modelo 4*

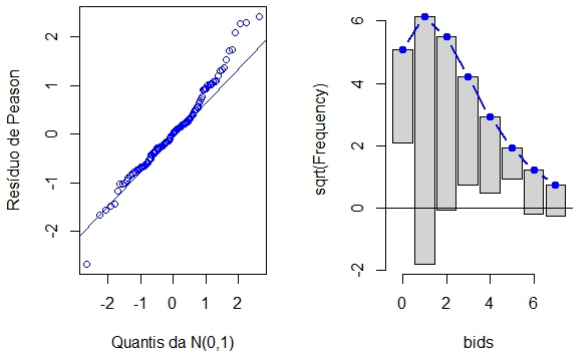


Figura 4.14: Gráficos dos resíduos quantílicos *versus* observações e dos desvios residuais *versus* valores ajustados do *Modelo 4*

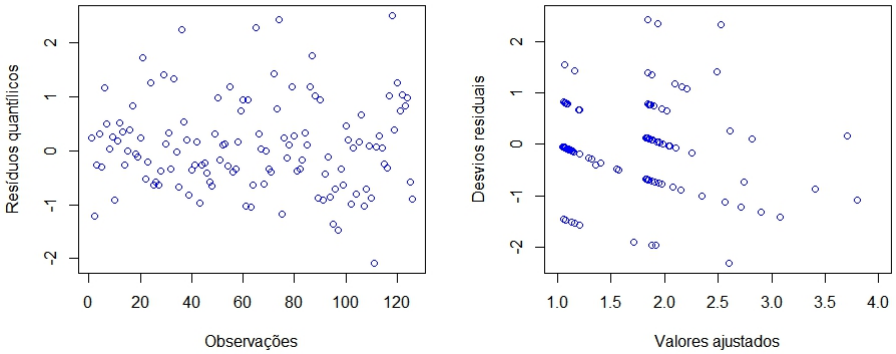


Tabela 4.18: Teste de Voung entre o modelo de regressão de Poisson e o modelo de regressão Binomial Negativa

Modelo Poisson (*Modelo 2*) vs	Modelo Binomial Negativa (*Modelo 4*)
Estatística de teste=-0,0009	P-valor= 0,5

Modelo de regressão de Poisson Generalizado (*Modelo 5*)

$$Bids_i \sim PG\left(\mu_i, \alpha\right), i=1,.....126$$

$$log(\mu_i) = \beta_0 + \beta_1 * Legalrest_i + \beta_2 * Realrest_i + \beta_3 * Finrest_i +$$
$$+ \beta_4 * Whiteknight_i + \beta_5 * Bidpremium_i + \beta_6 * Insthold_i + \beta_7 * Size_i + \beta_8 * Sizeq_i +$$

$+ \beta_9 * Regulation_i$

O modelo de regressão de Poisson Generalizado (*Modelo 5*) com todas as variáveis explicativas está apresentado na Tabela 4.19. Aplicando o método de seleção de variáveis (*Backward*), isto é, retirando sucessivamente as variáveis estatisticamente não significativas, obtém-se o *Modelo 6* com todas as variáveis estatisticamente significativas (Tabela 4.20).

Modelo de regressão de Poisson Generalizado (*Modelo 6*)

$$Bids_i \sim PG\left(\mu_i, \alpha\right), i=1,.....126$$

$log(\mu_i) = \beta_0 + \beta_1 * Whiteknight_i + \beta_3 * Size_i + \beta_4 * Sizeq_i + \beta_5 * Bidpremium_i$

Pelos valores apresentados na Tabela 4.20, para um nível de significância de 5% concluí-se que as variáveis *Whiteknight*, *Bidpremium*, *Size* e *Sizeq* são estatisticamente significativas. Comparando os três modelos, isto é, modelo de regressão de Poisson (*Modelo 2*), modelo de regressão Binomial Negativa (*Modelo 4*) e o modelo de regressão de Poisson Generalizado (*Modelo 6*), sugere-se o modelo de regressão de Poisson Generalizado (*Modelo 6*) por apresentar menor valor de *AIC*. As estatisticas de ajustamento dos referidos modelos apresentam-se na Tabela 4.21.

Tabela 4.19: Estimativas dos parâmetros do modelo de regressão de Poisson Generalizada com todas as variáveis explicativas

Variáveis Explicativas	Estimativas dos parâmetros	Erro padrão	Estatística de teste	P-valor
Constante	-0,367	0,139	-2,640	0,008
Legalrest	0,248	0,126	1,964	0,049
Realrest	-0,156	0,158	-0,984	0,325
Finrest	0,146	0,176	0,828	0,408
Whiteknight	0,541	0,134	4,043	$< 0,001$
Bidpremium	-0,779	0,315	-2,471	0,013
Insthold	-0,404	0,355	-1,141	0,254
Size	0,189	0,049	3,806	$< 0,001$
Regulation	-0,001	0,133	-0,011	0,991
Sizeq	-0,008	0,003	-3,120	0,002
$\hat{\alpha} = -0,181$				

Na Figura 4.15, observa-se o comportamento dos resíduos não revelam um bom ajustamento deste modelo. Quanto à Figura 4.16, o primeiro gráfico representa os resíduos quantílicos *versus* o índice das observações e no segundo gráfico representa-

Tabela 4.20: Estimativas dos parâmetros do modelo de regressão de Poisson Generalizada com todas as variáveis estatisticamente significativas

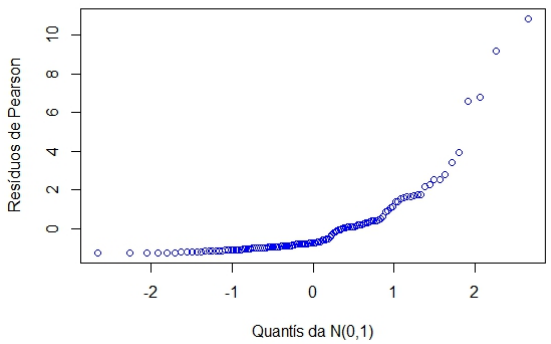
Variáveis Explicativas	Estimativas dos parâmetros	Erro Padrão	Estatística de teste	P-valor
Constante	-0,325	0,137	-2,373	0,017
Whiteknight	0,609	0,131	4,650	$< 0,001$
Bidpremium	-0,691	0,326	-2,122	0,034
Size	0,160	0,043	3,755	0,0001
Sizeq	-0,007	0,002	-2,858	0,004
$\hat{\alpha} = -0,181$				

Tabela 4.21: Estatísticas de ajustamento dos modelos (*Modelo 2*,*Modelo 4* e *Modelo 6*)

Estatísticas	Modelo Poisson (*Modelo 2*)	Modelo Binomial Negativo (*Modelo 4*)	Modelo PGeneralizado (*Modelo 6*)
AIC	384,616	386,617	380,033
BIC	395,961	400,799	397,051
ℓ	-188,308	-188,309	-184,016

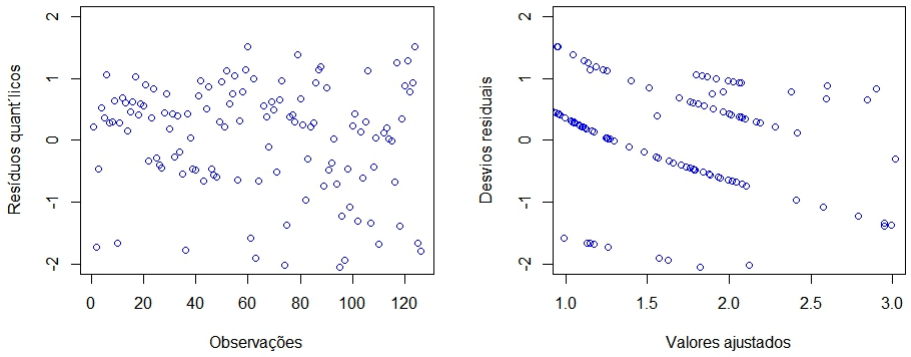
se resíduos *versus* os valores ajustados. Especificado cada gráfico observa-se que no primeiro gráfico os resíduos quantílicos distribuem-se de forma uniforme em torno da reta do resíduo zero. Enquanto que o segundo os resíduos estão bem distribuídos em torno de zero e não apresentam evidências de heteroscedasticidade.

Figura 4.15: Gráficos dos resíduos de Pearson *versus* quantis da N(0,1) referentes ao *Modelo 6*



A Tabela 4.22 apresenta as estimativas dos parâmetros e os respectivos erros

Figura 4.16: Gráficos dos resíduos quantílicos *versus* observações e dos desvios residuais *versus* valores ajustados do *Modelo 6*



padrão de todas as variáveis estatisticamente significativas do modelo de regressão de Poisson (*Modelo 2*), modelo de regressão Binomial Negativa (*Modelo 4*) e o modelo de regressão de Poisson Generalizado (*Modelo 6*). Observa-se que o *Modelo 2* e o *Modelo 4* apresentam estimativas dos parâmetros similares e por sua vez diferente das estimativas dos parâmetros do *Modelo 6*. As estimativas dos parâmetros do *Modelo 6* são maiores em valores absolutos, relativamente ao *Modelo 2* e do *Modelo 4*. O *Modelo 6* apresenta menores desvios padrão relativamente ao *Modelo 2* e o *Modelo 4*. O parâmetro de dispersão (α) no modelo de regressão Binomial Negativa é maior em relação no modelo de regressão de Poisson Generalizado.

Tabela 4.22: Estimativas dos parâmetros de regressão dos modelos com todas as variáveis estatisticamente significativas

Variáveis Explicativas	Modelo Poisson (*Modelo 2*)	Modelo Binomial Negativa (*Modelo 4*)	Modelo PGeneralizada I (*Modelo 6*)
Constante	0,045 (0,135)	0,045 (0,135)	-0,325 (0,137)
Whiteknight	0,552 (0,152)	0,553 (0,152)	0,609 (0,131)
Bidpremium	–	–	-0,691 (0,326)
Size	0,158 (0,049)	0,158 (0,049)	0,160 (0,042)
Sizeq	-0,007 (0,003)	-0.007 (0,003)	-0,007 (0,002)
$\hat{\alpha}$	–	0,778	- 0,181

No *Modelo 6*, o coeficiente positivo associado à variável explicativa *Whiteknight*, indica que o número de ofertas públicas de aquisição em empresas cuja gerência convidou terceiros é 1,83 vezes maior quando comparado com empresas cuja gerência

não convidou terceiros. Relativamente o coeficiente associado a variável *Size*, indica que o número esperado de ofertas públicas de aquisição após a oferta inicial recebida pela empresa aumenta cerca de $17,36\%$, por cada aumento de um bilião de dólares no valor total de ativos da empresa. Quanto ao coeficiente negativo associado a variável explicativa *Bidpremium* indica que o número esperado de ofertas públicas de aquisição recebida pela empresa diminui cerca de $49,87\%$ por um aumento de uma unidade no prémio da oferta.

O parâmetro de dispersão do modelo de regressão de Poisson Generalizado (*Modelo 6*) é negativo, isto é, $\hat{\alpha} = -0,181$, sugerindo a existência da subdispersão dos dados.

4.2 Base de Dados *Creditcard*

A base de dados *Creditcard* utilizado nesta secção extraído na biblioteca *AER* do*software* R (Kleiber e Zeileis, 2008), refere-se ao o histórico de crédito de uma amostra de clientes que solicitaram um tipo de cartão de crédito. Greene (1998) analisou o número de relatórios com avaliação negativa, de uma conta de crédito, de 1319 indivíduos que solicitaram um cartão de crédito. A variável resposta *Reports* é o número de relatórios com avaliação negativa.

O conjunto de dados da referida base de dados é constituída por 1319 observações. Para analisar o número de relatórios com avaliação negativa foram estabelecidas as seguintes 12 variáveis:

- *Reports*: número de relatórios com avaliação negativas (variável resposta);
- *Card*: indica se o pedido de cartão de crédito foi aceite e está codificada em 1-não, 2-sim;
- *Age*: idade em anos.
- *Income*: salário anual (em 10.000 dólares);
- *Share*: relação entre as despesas mensais com cartão de crédito e a receita anual;
- *Expenditure*: despesa média mensal com cartão de crédito (em dólares);
- *Owner*: indica se o indivíduo tem casa própria, e está codificada em 1-sim, 2-não;
- *Selfemp*: indica se o indivíduo é trabalhador independente, e está codificada em 1-sim , 2- não;
- *Dependents*: número de dependentes;
- *Months*: número de meses a viver no endereço atual;

• *Majorcards*: indica se o cartão de crédito está retido, e está codificada em 0-retido, 1-não retido ;

• *Active*: número de contas de crédito ativas.

4.2.1 Análise descritiva da base de dados *Creditcard*

Para análise descritiva referente à base de dados *Creditcard*, no caso das variáveis explicativas qualitativas (*Card*, *Owner*, *Selfemp* e *Majorcards*) utilizam-se gráficos de barras e tabelas de distribuição de frequências. Para as variáveis explicativas quantitativas contínuas, apresentam-se as tabelas das medidas de tendência central e das medidas de dispersão, assim como, os gráficos caixa com bigodes e os respectivos histogramas. Para as variáveis explicativas quantitativas discreta são apresentados gráfico de barras, tabelas de distribuição de frequências e as tabelas de medidas de dispersão e tendência central.

Variavel resposta (*Reports*)

O número de relatórios com avaliação negativa é a variável resposta. O comportamento desta variável está apresentado no gráfico da Figura 4.17. Na Tabela 4.23 são apresentadas as estatísticas descritivas da variável *Reports* e a Tabela 4.24 mostra a tabela das frequências desta variável. Nota-se que o número médio de relatórios com avaliação negativa é 0,456, o número mínimo e o número máximo de relatórios com avaliação negativa é de 0 à 14, respectivamente.

Tabela 4.23: Estatísticas descritivas da variável *Reports*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0	0,456	0	0	0	14	1,345	295

Pedido de cartão de crédito foi aceite(*Card*)

A Figura 4.18 mostra o gráfico de barras da variável *Card* e a tabela das frequências da referida variável está apresentada na Tabela 4.25 . Pode-se observar que há um maior número de cartões de crédito que foram aceites (296 pedidos não foram aceite).

Indivíduo tem casa própria(*Owner*)

A Figura 4.19 apresenta o gráfico de barras da variável *Owner*. Na Tabela 4.26 apresenta-se a tabela das frequências desta variável. Pode-se observar que há muitas pessoas sem casa própria (738 indivíduos não tem casa própria).

O indivíduo é trabalhador independente(*Selfemp*)

Figura 4.17: Gráfico de barras da variável *Reports*

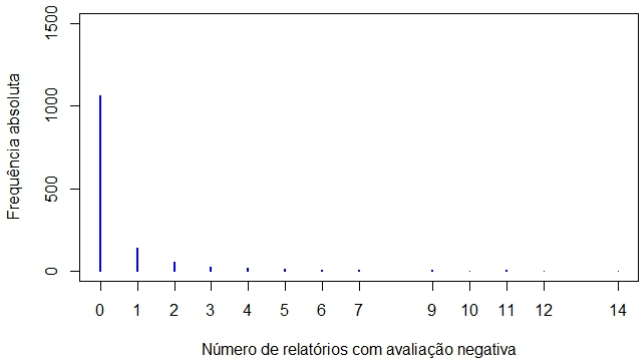


Figura 4.18: Gráfico de barras da variável *Card*

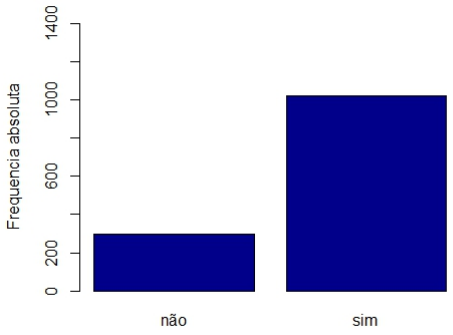


Tabela 4.24: Tabela de frequências do número de relatórios com avaliação negativa

Reports	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
0	1060	80,36	1060	80,36
1	137	10,39	1197	90,75
2	50	3,79	1247	94,54
3	24	1,82	1271	96,36
4	17	1,29	1288	97,65
5	11	0,83	1299	98,48
6	5	0,38	1304	98,86
7	6	0,45	1310	99,32
9	2	0,15	1312	99,47
10	1	0,08	1313	99,55
11	4	0,30	1317	99,85
12	1	0,08	1318	99,92
14	1	0,08	1319	100,00

Tabela 4.25: Tabela de frequências da variável *Card*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	1023	77,56	1023	77,56
2 - Não	296	22,44	1319	100,00

Tabela 4.26: Tabela de frequências da variável *Owner*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	581	44,05	581	44,05
2 - Não	738	55,95	1319	100,00

A Figura 4.20 apresenta o gráfico de barras da variável *Selfemp*. Na Tabela 4.27 são apresentados a tabela das frequências desta mesma variável. Nota-se que há poucos indivíduos que são trabalhadores independentes (91 indivíduos são trabalhadores independentes).

Cartão de crédito está retido (*Majorcards*)

A Figura 4.21 apresenta o gráfico de barra da variável *Majorcards* e a Tabela 4.28 mostra a tabela das frequências da referida variável. Observa-se que há poucos

Figura 4.19: Gráfico de barras da variável *Owner*

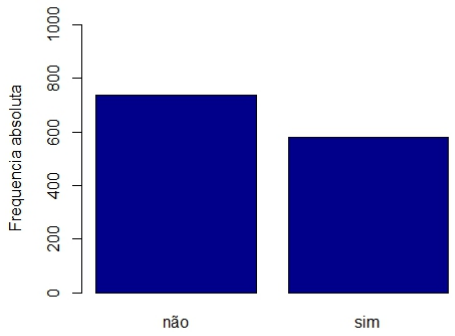
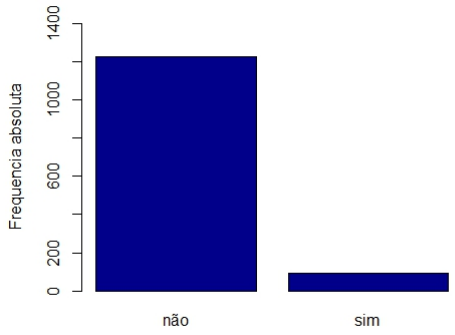


Tabela 4.27: Tabela de frequências da variável *Selfemp*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
1 - Sim	91	0,68	91	0,68
2 - Não	1228	93,100	1319	100,00

Figura 4.20: Gráfico de barras da variável *Selfemp*



cartões de crédito retido (241 cartões de créditos não retidos).

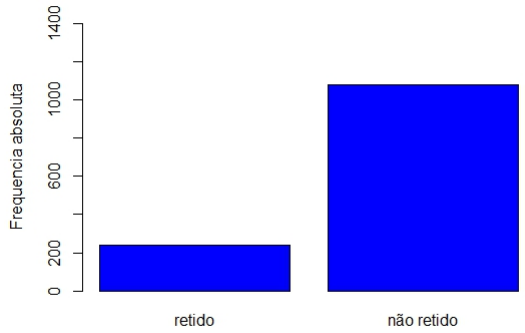
Idade em anos (*Age*)

A Figura 4.22 apresenta o histograma e o gráfico caixa com bigodes da variável *Age*. A Tabela 4.29 apresenta as medidas de tendência central e de dispersão da referida variável. O histograma da variável idade em anos sugere um enviesamento

Tabela 4.28: Tabela de frequências da variável *Majorcards*

Categoria	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
0 - Retido	241	18,27	241	18,27
1- Não retido	1078	81,73	1319	100,00

Figura 4.21: Gráfico de barras da variável *Majorcards*



positivo, a maior número de frequências das idades localiza-se no intervalo de classe $[20; 40]$. A Tabela 4.29 mostra que 50% das idades estão compreendidas entre 25,42 e 39,42 anos.

Tabela 4.29: Estatísticas descritivas da variável *Age*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0,167	33,213	25,417	31,250	39,417	83,500	10,143	30,539

Salário anual (*Income*)

O histograma e o gráfico caixa com bigode apresentados na Figura 4.23, dizem respeito à variável salário anual. A Tabela 4.30 mostra a estatística descritiva da variável *Income* e observa-se que o valor mínimo e o máximo do salário anual é de $0,210 \times 10^4$ e $13,50 \times 10^4$ dólares respectinamente, 25% dos indivíduos têm salário anual menores ou iguais à $2,244 \times 10^4$ dólares e 25% têm salário anual maiores ou iguais à 4×10^4 dólares. Também observa-se no gráfico caixa com bigodes, a existência de um grande número de observações *outliers*.

Relação entre as despesas mensais com cartão de crédito e a receita

Figura 4.22: Caixa com bigodes e o histograma da variável *Age*

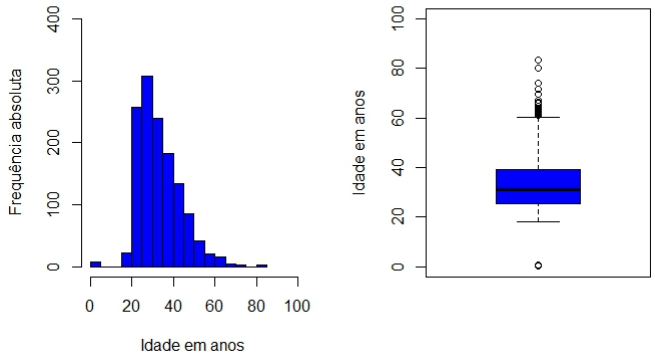


Tabela 4.30: Estatística descritiva da variável *Income*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0,210	3,365	2,244	2,900	4	13,500	1,694	50,339

anual (*Share*)

A Figura 4.24 mostra o histograma e o gráfico caixa com bigodes da variável *Share*. A Tabela 4.31 apresenta as medidas estatísticas desta variável. Observa-se que 50% da relação entre as despesas mensais com cartão de crédito e a receita anual são menores ou iguais a 0,039. Observa-se ainda existência de um grande número de observações *outliers* no gráfico caixa com bigodes. O histograma apresenta um enviesamento positivo e a maior frequência do valores relação entre as despesas mensais com cartão de crédito e a receita anual estão no intervalo de classe de $[0; 0, 1]$.

Tabela 4.31: Estatísticas descritivas da variável *Share*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0,0001	0,069	0,002	0,039	0,094	0,906	0,095	137,717

Despesa média mensal com cartão de crédito (*Expenditure*)

A Figura 4.25 apresenta o gráfico caixa com bigode e o histograma da variável *Expenditure*. Na Tabela 4.32 apresenta-se as estatísticas descritivas da referida variável. observa-se que o número máximo de despesa média com cartão de crédito é 3099,505 dólares e 25% da despesas medias são menores ou iguais a 4,583 dólares e 25% maior

Figura 4.23: Caixa com bigodes e o histograma da variável *Income*

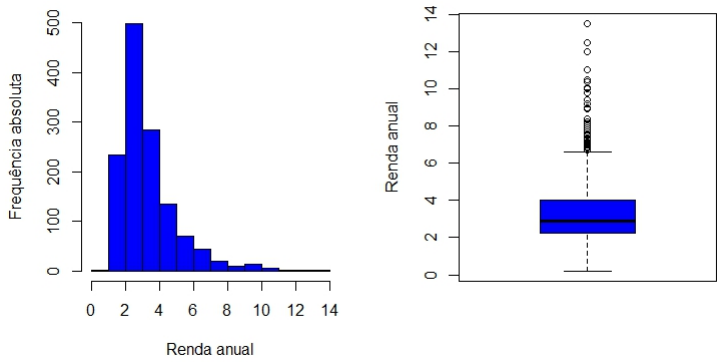


Figura 4.24: Caixa com bigodes e o histograma da variável *Share*

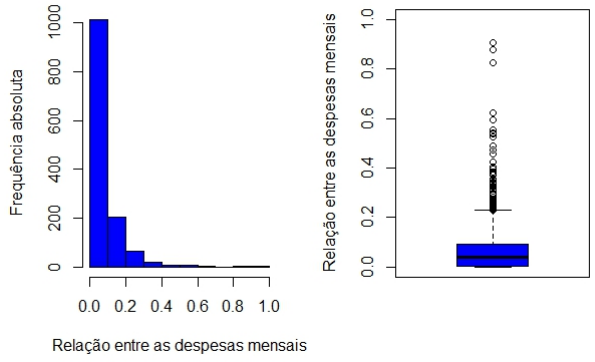
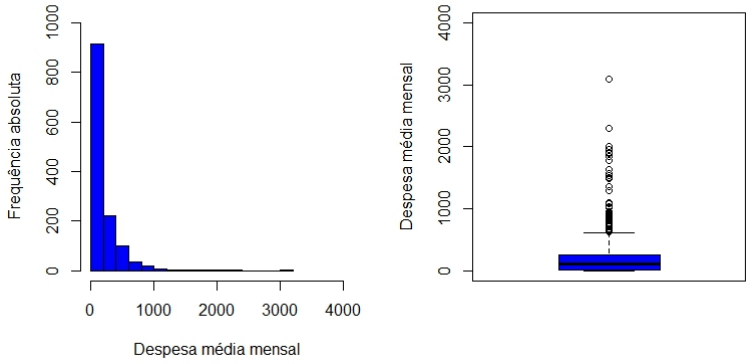


Figura 4.25: Caixa com bigodes e o histograma da variável *Expenditure*



ou iguais a 249,036 dólares. Observa-se através do histograma que a distribuição da variável não é simétricos.

Tabela 4.32: Estatísticas descritivas da variável *Expenditure*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0	185,057	4,583	101,298	249,036	3099,505	272,219	147,1

Meses a viver no endereço atual(*Months*)

A Figura 4.26 e a Tabela 4.33 apresentam as estatísticas descritivas da variável *Months*. Nota-se uma alta variabilidade desta variável cerca de ($CV = 119.905\%$), o valor mínimo e o valor máximo é de 0 e 540 respectivamente. Observa-se que 50% dos individuos estão a viver no endereço atual à menos de 30 meses. A partir do histograma conclui-se que a variável *Months* tem uma distribuição não simétrica.

Tabela 4.33: Estatísticas descritivas da variável *Months*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0	55,27	12	30	72	72	66,272	119,905

Número de dependentes(*Dependents*)

A Figura 4.27 apresenta o gráfico de barra da variável *Dependents*. Na Tabela 4.34 apresenta-se as medidas de tendência central e de dispersão e as frequências são apresentadas na Tabela 4.35. A média do número de dependentes é de 0,994 e a variabilidade desta variável é alta cerca de ($CV = 125,54\%$).

Figura 4.26: Caixa com bigodes e o histograma da variável *Months*

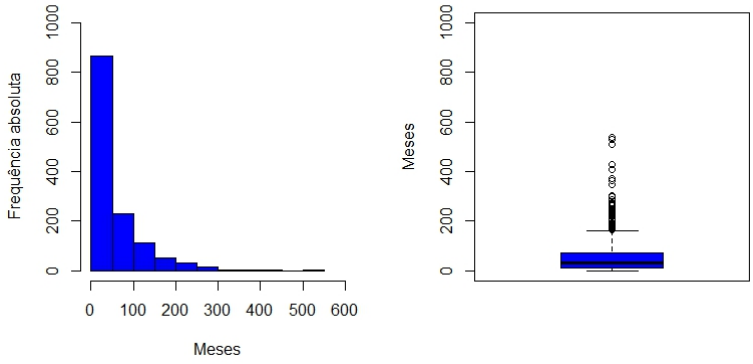


Tabela 4.34: Tabela de frequências da variável *Dependents*

Dependents	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
0	659	49,96	659	49,96
1	267	20,24	926	70,20
2	218	16,53	1144	86,73
3	115	8,72	1259	95,45
4	44	3,34	1303	98,79
5	9	0,68	1312	99,45
6	7	0,53	1319	100,00

Número de contas de crédito ativas(*Active*)

A Figura 4.28, Tabela A.1 (em anexo) e a Tabela 4.36 mostram as estatísticas descritivas da variável *Active*. Nota-se que o número médio de contas de créditos ativas é de 6,997, o número mínimo é 0 e o número máximo é 46. 50% do número de contas de crédito ativas estão compreendidos entre 2 e 11.

Associação entre a variável resposta e as variáveis explicativas

A Tabela 4.37 apresenta os coeficientes de correlação de *Spearman* entre a variável resposta *Reports* e cada uma das variáveis explicativas quantitativas, assim como o

Tabela 4.35: Estatísticas descritivas da variável *Dependents*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0	0,994	0	1	2	6	1,248	125,54

Figura 4.27: Gráfico de barras da variável *Dependents*

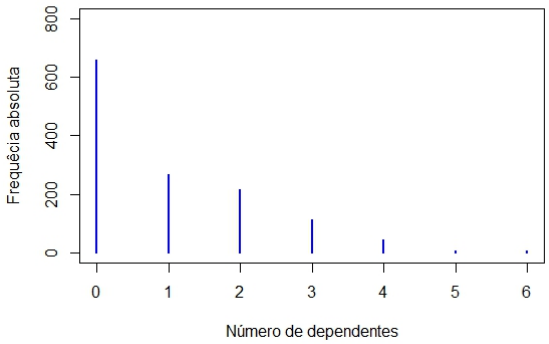


Tabela 4.36: Estatística descritiva da variável *Active*

Mínimo	Média	1º Quartil	Mediana	3º Quartil	Máximo	Desvio padrão	Coeficiente variação(%)
0	6,997	2	6	11	46	6,306	90,122

p-valor associado ao teste de hipótese do coeficiente de correlação de *Spearman* ser zero entre as variáveis.

Os p-valores associados ao teste de hipóteses do coeficiente de correlação de Spearman das variáveis *Income* e *Dependents* são superiores a um nível de significância de 5%, portanto apresentam uma correlação não significativa, indicando a não associação destas variáveis com a variável resposta. Para um nível de significância de 5%, rejeita-se a hipótese de não existir associação entre as variáveis explicativas quantitativas *Age*, *Months*, *Active* e a variável resposta, e existe uma correlação significativa positiva. Enquanto as variáveis *Share* e *Expenditure* apresentam uma correlação significativa negativa com a variável resposta.

Na Tabela 4.38 apresentam-se os valores das estatísticas de teste e os respectivos p-valores do teste de Mann-Whitney, quando se pretende comparar a distribuição da variável *Reports* nas categorias das variáveis explicativas. Para um nível de significância 5%, não se rejeita a hipótese nula sugerindo que não existe diferença entre as categorias das variáveis *Owner*, *Selfemp* e *Majorcards* em termos da distribuição do número de relatórios com avaliação negativa. Para um nível de significância de 5%, rejeita-se a hipótese nula da igualdade das distribuições sugerindo evidências estatística que ocorrem diferença significativa entre as categorias da variável *Card* em termos de número de relatórios com avaliação negativa.

Figura 4.28: Gráfico de barras da variável *Active*

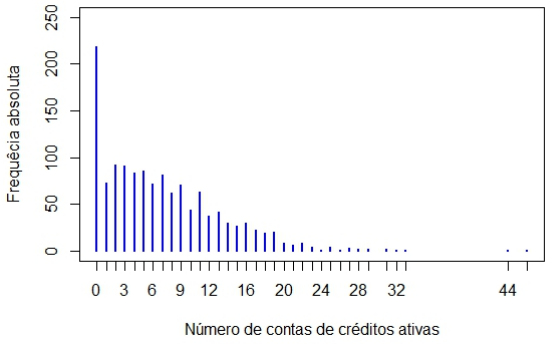


Tabela 4.37: Coeficiente de correlação de *Spearman* entre as variáveis explicativas e a variável *Reports*

Variável	Coeficiente	P-valor
Age	0,104	0,001
Income	0,048	0,081
Share	-0,302	$< 0,001$
Expenditure	-0,274	$< 0,001$
Dependents	0,029	0,285
Months	0,134	$< 0,001$
Active	0,246	$< 0,001$

Tabela 4.38: Teste de Mann-Whitney para as variáveis explicativas e a variável resposta *Reports*

Variável	Estatística de teste	P-valor
Card	217360	$< 0,001$
Owner	219790	0,256
Selfemp	53282	0,286
Majorcards	131770	0,613

4.2.2 Modelo estatístico

Para explicar o número de relatórios com avaliação negativa apresenta-se inicialmente o modelo de regressão de Poisson. A Tabela 4.39 apresenta o modelo de regressão de Poisson com todas as variáveis explicativas (*Modelo 7*).

Modelo de regressão de Poisson (*Modelo 7*)

$$Reports_i \sim Poi\left(\mu_i\right), i=1,....,1319$$

$$log(\mu_i) = \beta_0 + \beta_1 * Card_i + \beta_2 * Age_i + \beta_3 * Income_i +$$
$$+ \beta_4 * Share_i + \beta_5 * Expenditure_i + \beta_6 * Owner_i + \beta_7 * Selfemp_i + \beta_8 * Dependents_i +$$
$$+ \beta_9 * Months_i + \beta_{10} * Majorcards_i + \beta_{11} * Active_i$$

Tabela 4.39: Estimativas do modelo de regressão de Poisson com todas variáveis explicativas

Variáveis Explicativas	Estimativas dos parâmetros	Erro padrão	Estatística de teste	P-valor
Constante	-0,382	0,182	-2,092	0,036
Card	-2,726	0,127	-21,499	$< 0,001$
Age	-0,001	0,005	-0,033	0,974
Income	0,036	0,027	1,312	0,189
Share	0,779	1,080	0,721	0,471
Expenditure	0,001	0,001	1,516	0,129
Owner	-0,383	0,102	-3,759	$< 0,001$
Selfemp	-0,166	0,149	-1,111	0,267
Dependents	0,026	0,036	0,730	0,465
Months	0,002	0,001	3,493	$< 0,001$
Majorcards	0,245	0,106	2,314	0,021
Active	0,065	0,004	15,825	$< 0,001$

A Tabela 4.40 apresenta-se os valores da estatística VIF para cada uma das variáveis existente no conjunto de dados *Creditcard*. Logo concluí-se que não existem

problemas de multicolinearidade ($VIF < 5$).

Tabela 4.40: Valores da estatística do VIF do modelo de regressão de Poisson

Variáveis explicativas	VIF
Card	1,657
Age	1,532
Income	1,456
Share	3,848
Expenditure	3,682
Owner	1,445
Selfemp	1,027
Dependents	1,184
Months	1,315
Majorcards	1,042
Active	1,203

Aplicando o método de seleção de variáveis *Backward*, e retirando sucessivamente as variáveis estatisticamente não significativas, obtém-se o *Modelo 8*. Os valores ajustados deste modelo apresentam-se na Tabela 4.41.

Modelo de regressão de Poisson (*Modelo 8*)

$$Reports_i \sim Poi\left(\mu_i\right), i=1,....,1319$$

$log(\mu_i) = \beta_0 + \beta_1 * Card_i + \beta_2 * Expenditure_i + \beta_3 * Owner_i + \beta_4 * Months_i + \beta_5 * Majorcards_i + \beta_6 * Active_i$

Tabela 4.41: Estimativas dos parâmetros do modelo de regressão de Poisson com todas variáveis estatisticamente significativas

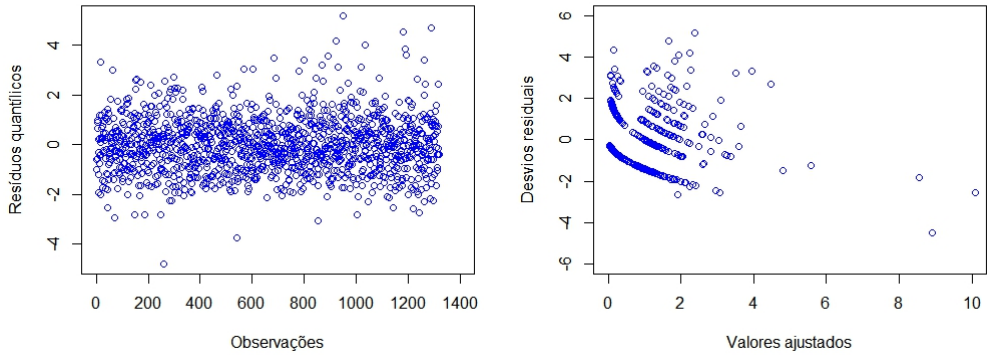
Variáveis explicativas	Estimativas dos parâmetros	Erro padrão	Estatística de teste	P-valor
Termo constante	-0,299	0,109	-2,723	0,006
Card	-2,704	0,117	-23,068	$< 0,001$
Expenditure	0,0007	0,0002	3,785	$< 0,001$
Owner	-0,344	0,093	-3,710	$< 0,001$
Months	0,002	0,0005	4,006	$< 0,001$
Majorcards	0,274	0,105	2,622	0,008
Active	0,065	0,004	16,367	$< 0,001$

Análise de resíduos

Na Figura 4.29 são apresentados os gráficos dos resíduos quantílicos *versus* observações, assim como os desvios residuais *versus* valores ajustados do modelo de

regressão de Poisson (*Modelo 8*). Relativamente ao primeiro gráfico observa-se que os resíduos quantílicos se distribuem de forma uniforme em torno de zero, ou seja, a distribuição dos resíduos quantílicos fornece indicativo da independência dos erros. Quanto ao segundo gráfico pode-se observar que os resíduos não se distribuem de forma uniforme em torno de zero, ou seja, há indícios de que o pressuposto da homogeneidade das variâncias dos erros não é verificado. Na Figura 4.30, apresenta-se os gráficos dos resíduos de Person *versus* quantis da distribuição normal e o *rootgram*. Observa-se que no primeiro gráfico, ocorre uma distorção dos resíduos em torno da linha recta, portanto os resíduos não seguem uma distribuição normal. No segundo gráfico percebe-se que as barras do número de relatórios com avaliação negativa mais baixas (de 0 a 2) se distanciam muito da reta das abcissas. À medida que o número de relatórios com avaliação negativa aumentam, as barras se aproximam mais do eixo das abcissas o valor ajustado (comprimento das barras) e o observado (linha azul) se tornam mais próximos.

Figura 4.29: Gráficos dos resíduos quantílicos *versus* observações e dos desvios residuais *versus* valores ajustados do *Modelo 8*



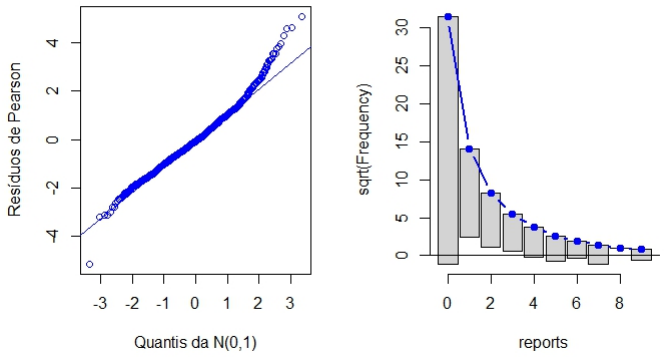
Sobredispersão e Subdispersão

Para o *Modelo 8* ajustado nota-se que o parâmetro $\hat{\alpha} = 0.989 > 0$, logo concluí-se que não há indícios de sobredispersão.

Aplicando o teste de Cameron e Trivedi (2005), para identificar a sobredispersão dos dados, para o *Modelo 8* ajustado, observa-se que a estimativa do parâmetro de dispersão $\hat{\alpha} = 0.973 > 0$ e para um nível de significância de 5% rejeita-se a hipótese nula da equidispersão em modelo de regressão de Poisson contra a alternativa de sobredispersão ($p - valor < 0.001$). Concluindo que há indícios de sobredispersão.

Visto que o modelo de regressão de Poisson apresenta sobredispersão, percebe-se

Figura 4.30: Gráficos dos resíduos de Pearson *versus* quantís da N(0,1) e o rootogram referente ao *Modelo 8*



que é necessário encontrar um modelo que consiga melhorar o ajuste. Para isso, a distribuição Binomial Negativa deve conseguir ajustar melhor pela sua capacidade de modelar dados que apresentam sobredispersão e também, de certa forma, o modelo de regressão de Poisson Generalizado, mas antes disso ajusta-se inicialmente o modelo de regressão Binomial Negativa.

Modelo de regressão Binomial Negativa (*Modelo 9*)

$$Reports_i \sim BinNeg\left(\mu_i, \alpha\right), i{=}1,.....1319$$

$log(\mu_i) = \beta_0 + \beta_1 * Card_i + \beta_2 * Age_i + \beta_3 * Income_i +$
$+ \beta_4 * Share_i + \beta_5 * Expenditure_i + \beta_6 * Owner_i + \beta_7 * Selfemp_i + \beta_8 * Dependents_i +$
$+ \beta_9 * Months_i + \beta_{10} * Majorcards_i + \beta_{11} * Active_i$

Na Tabela 4.42 apresenta-se os resultados do *Modelo 9* com todas as variáveis explicativas. Aplicando o método de seleção de variáveis *Backward*, retirando sucessivamente as variáveis estatisticamente não significavas obtém-se o *Modelo 10*.

Modelo de regressão Binomial Negativa (*Modelo 10*)

$$Reports_i \sim BinNeg\left(\mu_i, \alpha\right), i{=}1,.....1319$$

$log(\mu_i) = \beta_0 + \beta_1 * Card_i + \beta_2 * Income_i + \beta_3 * Share_i + \beta_4 * Owner_i + \beta_5 * Months_i + \beta_6 * Active_i$

A Tabela 4.43 apresenta-se as estimativas dos parâmetros do modelo de regressão Binomial Negativa (*Modelo 10*) com todas as variáveis estatisticamente significativas. Ajustado o modelo de regressão Binomial Negativa (*Modelo 10*), percebe-se

Tabela 4.42: Estimativas dos parâmetros do modelo de regressão Binomial Negativa com todas variáveis explicativa

Variáveis Explicativas	Estimativas dos parâmetros	Erro Padrão	Estatística de teste	P-valor
Termo constante	-1,023	0,282	-3,620	0,001
Card	-2,907	0,161	-18,000	$< 0,001$
Age	0,004	0,008	0,589	0,556
Income	0,079	0,042	1,864	0,062
Share	1,982	1,368	1,448	0,147
Expenditure	0,001	0,001	0,511	0,609
Owner	-0,436	0,155	-2,818	$< 0,001$
Selfemp	-0,089	0,234	-0,383	0,702
Dependents	0,002	0,054	0,043	0,966
Months	0,003	0,001	2,802	0,005
Majorcards	0,205	0,167	1,229	0,219
Active	0,112	0,009	12,646	$< 0,001$
$\hat{\alpha} = 0,838$				

uma melhoria significativa nos dois gráficos da Figura 4.31. O número de relatórios com avaliação negativa mais baixos que não tiveram um ajuste muito bom melhoraram significativamente o ajuste no modelo de regressão Binomial Negativa. O gráfico dos resíduos de Pearson *versus* quantís da distribuição normal também teve uma melhoria significativa, sugerindo a plausibilidade da suposição da normalidade dos resíduos. Relativamente os dois gráficos da Figura 4.32, observando o gráfico dos desvios residuais *versus* valores ajustados não melhorou o seu ajuste, há indícios que a variância dos resíduos não é heteroscedastica enquanto que o gráfico resíduos quantílicos *versus* índice das observações os resíduos distribuem-se uniformemente em torno de 0.

Aplicou-se o teste de *Voung*, entre o modelo de regressão Binomial Negativa (*Modelo 10*) e o modelo de regressão de Poisson (*Modelo 8*), concluí-se que o modelo de regressão Binomial Negativa (*Modelo 10*) é preferível ao modelo de regressão de Poisson (*Modelo 8*). Os resultados do teste está apresentado na Tabela 4.44.

Modelo de Regressão de Poisson Generalizado(*Modelo 11*)

$$Reports_i \sim PG\left(\mu_i, \alpha\right), i=1,.....1319$$

$$log(\mu_i) = \beta_0 + \beta_1 * Card_i + \beta_2 * Age_i + \beta_3 * Income_i +$$
$$+ \beta_4 * Share_i + \beta_5 * Expenditure_i + \beta_6 * Owner_i + \beta_7 * Selfemp_i + \beta_8 * Dependents_i +$$
$$+ \beta_9 * Months_i + \beta_{10} * Majorcards_i + \beta_{11} * Active_i$$

Figura 4.31: Gráficos dos resíduos de Pearson *versus* quantís da N(0,1) e o *rootogram* referente ao *Modelo 10*

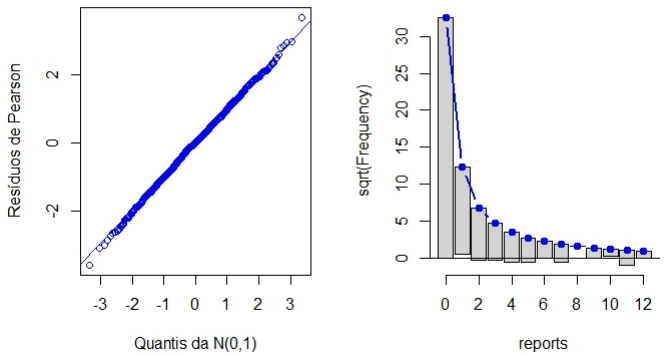


Figura 4.32: Gráficos dos resíduos quantílicos *versus* observações e dos desvios residuais *versus* valores ajustados referente ao *Modelo 10*

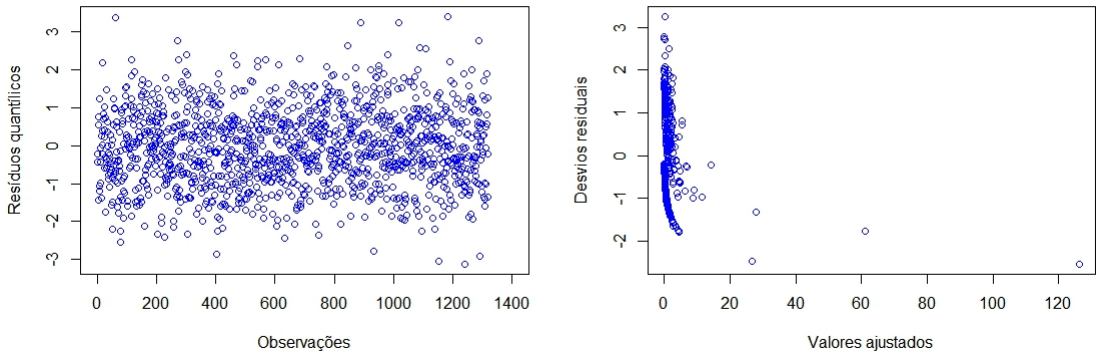


Tabela 4.43: Estimativas dos parâmetros do modelo de regressão Binomial Negativa (*Modelo 10*) com todas variáveis estatisticamente significativas

Variáveis Explicativas	Estimativas dos parâmetros	Erro Padrão	Estatística de teste	P-valor
Termo constante	-0,791	0,151	-5,235	$< 0,001$
Card	-2,893	0,160	-18,029	$< 0,001$
Income	0,091	0,036	2,507	0,012
Share	2,536	0,793	3,200	0,001
Owner	-0,404	0,146	-2,766	0,006
Months	0,003	0,0008	3,272	0,001
Active	0,114	0,009	12,970	$< 0,001$
$\hat{\alpha} = 0,825$				

Tabela 4.44: Teste de Voung entre o modelo de regressão de Poisson e o modelo de regressão Binomial negativa

Modelo Poisson (*Modelo 8*) vs	Modelo Binomial Negativa (*Modelo 10*)
Estatística de teste= 252.722	$P - valor < 0.001$

Na Tabela 4.45 são apresentados os resultados do modelo ajustado de regressão de Poisson Generalizado (*Modelo 11*) com todas as variáveis explicativas. Aplicando o método de seleção de variáveis (*Backward*), isto é, retirando sucessivamente as variáveis não significativas obtém-se o *Modelo 12* com todas as variaveis estatisticamente significativas (Tabela 4.46).

Modelo de Regressão de Poisson Generalizado(*Modelo 12*)

$$Reports_i \sim PG\left(\mu_i, \alpha\right), i=1,.....1319$$

$log(\mu_i) = \beta_0 + \beta_1 * Card_i + \beta_2 * Income_i + \beta_3 * Share_i + \beta_4 * Owner_i + \beta_5 * Months_i + \beta_6 * Active_i$

As variáveis apresentadas na Tabela 4.46, para um nível de significância 5% são estatisticamente significativas. Portanto comparando os três modelos, isto é, Modelo de Regressão de Poisson (*Modelo 8*), Modelo de Regressão Binomial Negativa (*Modelo 10*) e o Modelo de Regressão de Poisson Generalizado (*Modelo 12*), nota-se que o *Modelo 12* é preferível por apresentar menor valor de *AIC* e *BIC*. As estatísticas de ajustamento dos referidos modelos apresentam-se na Tabela 4.47.

Na Tabela 4.48 apresentam-se as estimativas dos parâmetros de todas as variáveis estatisticamente significativas do modelo de regressão de Poisson (*Modelo 8*), modelo de regressão Binomial Negativa (*Modelo 10*) e o modelo de regressão de Poisson

Tabela 4.45: Estimativas dos parâmetros do modelo de regressão de Poisson Generalizada com todas variáveis explicativa

Variáveis Explicativas	Estimativas dos parâmetros	Erro Padrão	Estatística de teste	P-valor
Constante	-1,121	0,299	-3,749	$< 0,001$
Card	-3,019	0,176	-17,168	$< 0,001$
Age	0,005	0,007	0,630	0,528
Income	0,088	0,046	1,899	0,058
Share	2,457	1,747	1,406	0,159
Expenditure	0,0001	0,0006	0,157	0,875
Owner	-0,459	0,157	-2,917	0,004
Selfemp	-0,051	0,246	-0,208	0,835
Dependents	-0,007	0,055	-0,125	0,900
Months	0,003	0,001	2,421	0,015
Majorcards	0,166	0,170	0,977	0,328
Active	0,130	0,013	9,949	$< 0,001$
$\hat{\alpha} = 0.429$				

Generalizado (*Modelo 12*), e os valores em parênteses são os respectivos erros padrão. Observa-se que o parâmetro de dispersão do modelo de regressão Binomial Negativa (*Modelo 10*) é maior em relação ao modelo de de regressão de Poisson Generalizado (*Modelo 12*). As estimativas dos parâmetros do *Modelo 12* são maiores em valor absoluto e apresenta maior desvio padrão do que as do *Modelo 8* e o *Modelo 10*.

Como foi sugerido o modelo de regressão de Poisson Generalizado por apresentar menor valor de *AIC* e *BIC*, portanto será feita a interpretação dos resultados das variáveis estatisticamente significativas do *Modelo 12*. O coeficiente positivo asso-

Tabela 4.46: Estimativas dos parâmetros do modelo de regressão de Poisson Generalizada com todas variáveis estatisticamente significativas

Variáveis Explicativas	Estimativas dos parâmetros	Erro Padrão	Estatística de teste	P-valor
Constante	-0,889	0,167	-5,3291	$< 0,001$
Card	-3,015	0,174	-17,3007	$< 0,001$
Income	0,093	0,041	2,2893	0,022
Share	2,716	0,831	3,2672	$< 0,001$
Owner	-0,434	0,149	-2,9037	0,004
Months	0,003	0,001	2,7355	0,006
Active	0,133	0,013	10,2260	$< 0,001$
$\hat{\alpha} = 0.430$				

Tabela 4.47: Estatísticas de ajustamento dos modelos (*Modelo 8*,*Modelo 10* e *Modelo 12*)

Estatísticas	Modelo Poisson (*Modelo 8*)	Modelo Binomial Negativo (*Modelo 10*)	Modelo PGeneralizado (*Modelo 12*)
AIC	1940,663	1695,644	1688,043
BIC	1976,955	1737,121	1729,52
ℓ	-963,331	-839,822	-836,022

Figura 4.33: Gráficos dos resíduos quantílicos *versus* observações e dos desvios residuais *versus* valores ajustados referente ao *Modelo 12*

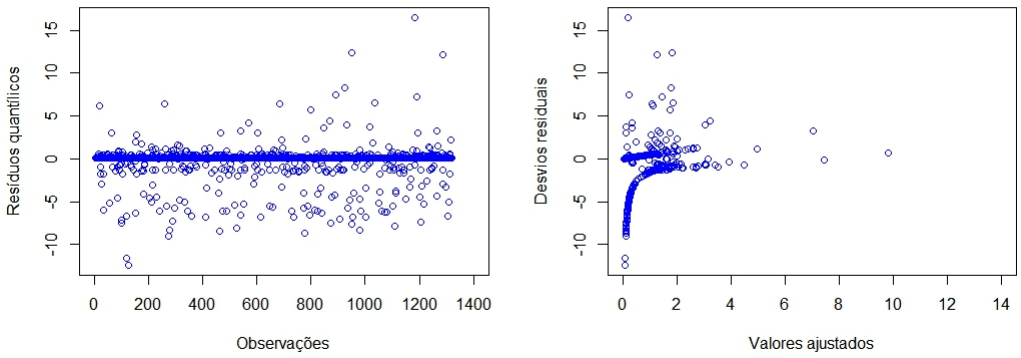


Tabela 4.48: Estimativas dos parâmetros dos modelos de regressão com todas as variáveis estatisticamente significativas

Variáveis Explicativas	Modelo Poisson (*Modelo 8*)	Modelo Binomial Negativa (*Modelo 10*)	Modelo PGeneralizada I (*Modelo 12*)
Constante	-0,299 (0,109)	-0,791 (0,151)	-0,889 (0,167)
Card	-2.702 (0,117)	-2,893 (0,160)	-3,015 (0,174)
Income	–	0,091 (0,036)	0,093 (0,041)
Share	–	2,536 (0,792)	2,716 (0,831)
Expenditure	0,0007 (0,0002)	–	–
Owner	-0,344 (0,093)	-0,404 (0,146)	-0,434 (0,149)
Months	0,002 (0,0005)	0,003 (0,0008)	0,003 (0,001)
Majorcards	0,274 (0,105)	–	–
Active	0,065 (0,004)	0,114 (0,009)	0,133 (0,013)
$\hat{\alpha}$	–	0,825	0,430

ciado a variável *Income*, indica que o número esperado de relatórios com avaliação negativa aumenta cerca de, por cada aumento em unidade o salário anual. O coeficiente associado a variável *Months* indica que o número esperado de relatórios com avaliação negativa aumenta cerca de por cada mês a viver no endereço atual. De igual modo, o número esperado de relatórios com avaliação negativa aumenta cerca de 14.178% por cada aumento do número de contas de crédito ativo. Relativamente ao coeficiente negativo associado à variável*Card* , indica que o número esperado de relatórios com avaliação negativa em clientes cujo pedido de cartão de crédito foi aceite é cerca de 89.71% menor do que clientes cujo cartão de crédito não foi aceite. O coeficiente associado a variável *Owner* , indica que o número esperado de relatórios com avaliação negativa em clientes com casa própria é cerca de 27.524% menor do que clientes que têm casa própria.

Capítulo 5

Conclusões

Nesta dissertação foram estudados o modelo de regressão de Poisson, o modelo de regressão Binomial Negativa e o modelo de regressão de Poisson Generalizado para modelar dados de contagem que apresentam sobredispersão ou subdispersão.

Estes modelos foram aplicados em dois conjuntos de dados. A primeira análise foi feita com o conjunto de dados *Takeoverbids*, que descreve o número de ofertas públicas de aquisição a oferta inicial recebida por 126 empresas Norte Americanas. Inicialmente fez-se uma análise descritiva dos dados, para melhor compreensão do comportamento estatístico das variáveis existente na referida base de dados. O modelo inicial para ajustar esses dados foi o modelo de regressão de Poisson, na qual revelou-se inadequado por apresentar indícios de subdispersão. De seguida, ajustou-se o modelo de regressão Binomial Negativa, não houve melhoria no ajuste.

Por fim, foi ajustado o modelo de regressão de Poisson Generalizado e chegou-se as seguintes conclusões:

• Com base o critério de informação (AIC e BIC), compararam-se os três modelos, isto é, o modelo de regressão de Poisson (*Modelo 2*), o modelo de regressão Binomial Negativa (*Modelo 4*) e o modelo de regressão de Poisson Generalizado (*Modelo 6*). O modelo de regressão de Poisson Generalizado (*Modelo 6*) é o modelo com menor valor de AIC.

• As estimativas dos parâmetros do modelo de regressão de Poisson Generalizado (*Modelo 6*) são menores em relação ao modelo de regressão Binomial Negativa (*Modelo 4*) e do modelo de regressão de Poisson (*Modelo 2*).

• O parâmetro de dispersão do modelo de regressão de Poisson Generalizado (*Modelo 6*) é menor do que o parâmetro de dispersão do modelo de regressão Binomial Negativa (*Modelo 4*).

• As estimativas do parâmetros do modelo de regressão de Poisson Generalizado

(*Modelo 6*) apresentam menores erros Padrão do que as estimativas dos dois outros modelos.

Quanto às estimativas dos parâmetros, as variáveis que tiveram maior impacto sobre o número de ofertas públicas de aquisição após a oferta inicial recebida pela empresa alvo foram as variáveis *Whiteknigth*, *Bippremium*, *Size* e *Sizeq*, observou-se por exemplo que o número esperado de ofertas públicas de aquisição após a oferta inicial recebida pela empresa aumenta cerca de $17,36\%$, por cada vez que se aumenta um bilião de dólares no valor total de ativos da empresa.

Relativamente ao conjunto de dados de uma amostra de clientes candidatos a cartão de credito (*Creditcard*), onde a variável de interesse foi o número de relatórios com avaliação negativa *Reports*, fez-se inicialmente uma análise exploratória dos dados, onde estudou-se o comportamento estatístico de cada variável existente no referido conjunto de dados.

Para modelar o número de relatórios com avaliação negativa ajustou-se o modelo de regressão de Poisson, que se revelou inadequado devido à sobredispersão dos dados. De seguida ajustou-se o modelo de regressão Binomial Negativa, que mostrou um bom ajuste aos dados. Mediante o teste de *Voung* para dados de contagem, verificou-se que o modelo de regressão Binomial Negativa (*Modelo 10*) é preferível que o modelo de regressão de Poisson (*Modelo 8*).

Por fim, ajustou-se então o modelo de regressão de Poisson Generalizado e concluí-se que:

• O modelo de regressão de Poisson Generalizado (*Modelo 12*) é preferível ao modelo de regressão Binomial Negativa (*Modelo 10*) e do modelo de regressão de Poisson (*Modelo 8*) por apresentar menor valor de *AIC* e *BIC*.

• O parâmetro de dispersão do modelo de regressão de Poisson Generalizado (*Modlo 12*) é menor do que o parâmetro de dispersão do modelo de regressão Binomial Negativa (*Modelo 10*).

Quanto às estimativas dos parâmetros, as variáveis que tiveram maior impacto sobre o número de relatórios com avaliação negativa foram as variáveis *Card*, *Income*, *Share*, *Owner*, *Majorcards* e *Active*, observou-se por exemplo que o número esperado de relatórios com avaliação negativa aumenta cerca de $14,178\%$, por cada vez que se aumenta o número de clientes com cartão de crédito ativo e o número esperado de relatórios com avaliação negativa diminui cerca de 89.708% à medida que o número de clientes que pediram cartão de crédito e foram aceite diminui quando comparado com os clientes que pediram e não foram aceite.

5.1 Trabalho Futuro

Para modelar dados de contagem com sobredispersão, várias metodologias têm sido desenvolvidas nomeadamente o modelo Quasi-verosimilhança. Um trabalho futuro será comparar essas metodologias com as estudadas neste trabalho.

Uma outra análise interessante seria estudar os modelos de regressão de Poisson Generalizados com efeitos aleatórios associado a um determinado individuo.

Bibliografia

[1] H. Akaike. A new look at the statistical model identification. Em *Selected Papers of Hirotugu Akaike*, 215–222. Springer, 1974.

[2] R. A. Bailey e L. J. Simon. Two studies in automobile insurance ratemaking. *ASTIN Bulletin: The Journal of the IAA*, 1(4):192–217, 1960.

[3] A. C. Cameron e P. K. Trivedi. Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of applied Econometrics*, 1(1):29–53, 1986.

[4] A. C. Cameron e P. K. Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, 46(3):347–364, 1990.

[5] A. C. Cameron e P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[6] A. C. Cameron e P. K. Trivedi. *Regression analysis of count data*. Cambridge University press, 2013.

[7] A. C. Cameron e P. K. Trivedi. Microeconometrics using stata. *Indicator*:47, 2009.

[8] Cameron e Johansson. Count data regression using series expansions: with applications. *Journal of Applied Econometrics*, (3):203–223, 1997.

[9] P. Consul. A model for distributions of injuries in auto-accidents. *Insurance: Mathematics and Economics*, 13(2):147, 1993.

[10] P. Consul e F. Famoye. The truncated generalized poisson distribution and its estimation. *Communications in Statistics-Theory and Methods*, 18(10):3635–3648, 1989.

[11] P. Consul e F. Famoye. Generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 21(1):89–109, 1992.

[12] P. Consul e G. C. Jain. A generalization of the poisson distribution. *Technometrics*, 15(4):791–799, 1973.

[13] P. Consul e L. Shenton. Some interesting properties of lagrangian distributions. *Communications in Statistics-Theory and Methods*, 2(3):263–272, 1973.

[14] P. Consul e M. Shoukri. Maximum likelihood estimation for the generalized poisson distribution. *Communications in Statistics. Theory and Methods*, 13(12):1533–1547, 1984.

[15] P. Consul e M. Shoukri. The generalized poisson distribution when the sample mean is larger than the sample variance. *Communications in Statistics-Simulation and Computation*, 14(3):667–681, 1985.

[16] G. M. Cordeiro e C. Demétrio. *Modelos lineares generalizados*. Univ. estadual de Campinas. Dep. de estat@ Wistica Campinas, 1986.

[17] Y. de Souza Tadano, C. Ugaya e A. Franco. Método de regressão de poisson: metodologia para avaliação do impacto da poluição atmosférica na saúde populacional. *Ambiente & Sociedade*, 12(2):241–255, 2009.

[18] H. Demirtas. On accurate and precise generation of generalized poisson variates. *Communications in Statistics-Simulation and Computation*, 46(1):489–499, 2017.

[19] P. K. Dunn e G. K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

[20] F. Famoye. Restricted generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 22(5):1335–1354, 1993.

[21] J. Ferlay. Cancer incidence, mortality and prevalence worldwide. *GLOBOCAN2002*, 2004.

[22] W. Greene. Sample selection in credit-scoring models. *Japan and The World Economy*, 10(3):299–316, 1998.

[23] A. Hackman, Y. Abe, W. Insull Jr, H. Pownall, L. Smith, K. Dunn, A. M. Gotto Jr e C. M. Ballantyne. Levels of soluble cell adhesion molecules in patients with dyslipidemia. *Circulation*, 93(7):1334–1338, 1996.

[24] J. M. Hilbe. *Modeling count data*. Springer, 2011.

[25] J. Hinde. Compound poisson regression models. Em *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, páginas 109–121. Springer, 1982.

[26] J. Hinde e C. G. Demétrio. Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170, 1998.

[27] N. Ismail e A. A. Jemain. Handling overdispersion with negative binomial and generalized poisson regression models. Em *Casualty actuarial society forum*, páginas 103–158. Citeseer, 2007.

[28] N. Ismail e H. Zamani. Estimation of claim count data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models. Em *Casualty Actuarial Society E-Forum*, 1–18, 2013.

[29] S. Jaggia e S. Thosar. Multiple bids as a consequence of target management resistance: a count data approach. *Review of Quantitative Finance and Accounting*, 3(4):447–457, 1993.

[30] K. G. Janardan, H. W. Kerster e D. J. Schaeffer. Biological applications of the lagrangian poisson distribution. *Bioscience*, 29(10):599–602, 1979.

[31] N. Jonhson, S. Kotz e A. Kemp. Discrete distributions: distributions in statistics. *Citado na*:32, 1969.

[32] C. Kleiber e A. Zeileis. *Applied econometrics with R*. Springer Science & Business Media, 2008.

[33] J. Marôco. *Análise de equações estruturais: Fundamentos teóricos, software & aplicações*. ReportNumber, Lda, 2010.

[34] P. McCullagh e J. Nelder. Generalized linear models., 2nd edn.(chapman and hall: london). *Standard book on generalized linear models*, 1989.

[35] A. Melliana, Y. Setyorini e H. Eko. Purhadi: the comparison of generalized poisson regression and negative binomial regression methods in overcoming overdispersion. *IJSTR*, 2:255–258, 2013.

[36] L. E. Mott, R. B. Lounsbury, B. C. Thompson, B. S. Peter, J. J. L. Murtaugh e A. A. Sardinas. Data processor module for a modular data processing system for operation with a time-shared memory in the simultaneous execution of multi-tasks and multi-programs, mai. de 1967. US Patent 3,319,226.

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

[38] Ozuna e Gomez. Specification and testing of count data recreation demand functions. *Empirical Economics*, 20(3):543–550, 1995.

[39] S. D. Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile precédées des règles générales du calcul des probabilités par SD Poisson*. Bachelier, 1837.

[40] J. Ramalho. *Modelos de regressão para dados de contagem*. Tese de mestrado, 1996.

[41] A. Sáez-Castillo e A. Conde-Sánchez. A hyper-poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis*, 61:148–157, 2013.

[42] C. Schmidt. *Modelo de regressão de Poisson aplicado à área da saúde*. Tese de doutoramento, Dissertação de Mestrado em Modelagem Matemática. Universidade Regional do ..., 2003.

[43] G. Schwarz *et al.* Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[44] K. F. Sellers e D. S. Morris. Underdispersion models: models that are "under the radar". *Communications in Statistics-Theory and Methods*, 46(24):12075–12086, 2017.

[45] S. M. Stigler. An attack on gauss, published by legendre in 1820. *Historia Mathematica*, 4(1):31–35, 1977.

[46] J. W. Tukey e W. S. Cleveland. *The collected works of John W. Tukey*, volume 1. Taylor & Francis, 1984.

[47] M. A. A. Turkman e G. L. Silva. Modelos lineares generalizados da teoria a prática. Em *VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa*, 2000.

[48] Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*:307–333, 1989.

[49] W. Wang e F. Famoye. Modeling household fertility decisions with generalized poisson regression. *Journal of Population Economics*, 10(3):273–283, 1997.

[50] J. S. Williams e H. K. Iyer. *Scientific Theories, Observational Studies and Statistical Methods: A Rejoinder to AS Heath*. Department of Statistics, Colorado State University, 1985.

[51] Z. Yang, J. W. Hardin e C. L. Addy. A score test for overdispersion in poisson regression based on the generalized poisson-2 model. *Journal of Statistical Planning and I nference*, 139(4):1514–1521, 2009.

[52] H. Zamani e N. Ismail. Functional form for the generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 41(20):3666–3675, 2012.

[53] A. Zeileis, C. Kleiber e S. Jackman. Regression models for count data in r. *Journal of statistical software*, 27(8):1–25, 2008.

[54] W. M. Zeviani, E. E. Ribeiro Jr e C. A. Taconeli. Modelos de regressão para dados de contagem com r, 2016.

[55] A. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev e G. M. Smith. *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media, 2009.

Apêndice A

Anexos

Tabela A.1: Número de contas de crédito ativas

Active	Frequência absoluta	Frequência relativa (%)	Frequência absoluta acumulada	Frequência relativa acumulada (%)
0	219	16,60	219	16,60
1	73	5,53	292	22,14
2	92	6,97	384	29,11
3	91	6,89	475	36,01
4	84	6,37	559	42,38
5	86	6,52	645	48,90
6	72	5,46	717	54,36
7	82	6,22	799	60,58
8	62	4,70	861	65,28
9	71	5,38	932	70,66
10	44	3,34	976	73,99
11	63	4,78	1039	78,77
12	38	2,88	1077	81,65
13	42	3,18	1119	84,84
14	30	2,27	1149	87,11
15	27	2,05	1176	89,16
16	30	2,27	1206	91,43
17	23	1,74	1229	93,18
18	20	1,52	1249	94,69
19	21	1,59	1270	96,29
20	9	0,68	1279	96,97
21	7	0,53	1286	97,49
22	9	0,68	1295	98,18
23	5	0,38	1300	98,56
24	1	0,08	1301	98,64
25	4	0,30	1305	98,94
26	1	0,08	1306	99,01
27	3	0,23	1309	99,24
28	2	0,15	1311	99,39
29	2	0,15	1313	99,55
31	2	0,15	1315	99,69
32	1	0,08	1316	99,77
33	1	0,08	1317	99,85
44	1	0,08	1318	99,92
46	1	0,08	1319	100,00

Apêndice B