Universidade do Minho
Escola de Ciências

Lúria Constância Cavalata Fernando

Modelos Lineares Generalizados na Análise de Dados de Saúde

julho de 2021

Lúria Constância Cavalata Fernando Modelos Lineares Generalizados na Análise de Dados de Saúde

UMinho | 2021

Universidade do Minho
Escola de Ciências

Lúria Constância Cavalata Fernando

Modelos Lineares Generalizados na Análise de Dados de Saúde

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação das Professoras
Arminda Gonçalves
Susana Faria

julho de 2021

Dedicatória

Dedico este trabalho ao meu marido Estatela Fernando que sempre me apoiou e acreditou em mim, pela motivação mesmo estando distante, aos meus filhos Estância Fernando e Gedeão Fernando, que suportaram a minha ausência de quase três anos.

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho.

Atribuição

CC BY

https://creativecommons.org/licenses/by/4.0/

Declaração de integridade

Eu, Lúria Constância Cavalata Fernando, nº PG35967, aluna do Mestrado em Estatística na Escola de Ciências da Universidade do Minho, 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.

Agradecimentos

Agradeço a Deus pelas muitas bençãos que proporciona na minha vida, aos meus pais por todo apoio e dedicação;

Agradeço às minhas professoras e orientadoras Doutora Arminda Manuela e Doutora Susana Faria, pela paciência, incentivo, empenho e todo o apoio dado no decorrer deste estudo e também pelas sugestões e críticas que foram importantes para a realização deste trabalho;

Agradeço aos meus Colegas pelo apoio, durante todo o meu percurso acadêmico;

A todos que de alguma forma contribuíram para a realização deste trabalho, muito obrigada.

Resumo

Ao longo dos séculos surgiram surtos e epidemias de doenças infeciosas que provocaram milhões de mortes. Atualmente, a nova doença COVID-19 derivada de um vírus (nomeado de SARS-CoV-2 pela Organização Mundial da Saúde (OMS)) já provocou no mundo inteiro, até ao momento, mais de 3 milhões de mortes. Para compreender melhor o comportamento desta nova doença têm sido desenvolvidos e analisados inúmeros processos de modelação, em particular na área de Modelos Lineares Generalizados.

Os Modelos Lineares Generalizados têm vindo a ser amplamente utilizados nas mais diversas áreas de estudo, para a modelação de fenómenos. O objetivo principal do estabelecimento de modelos deste tipo é analisar a influência que as variáveis explicativas têm sobre uma variável de interesse (a variável resposta), cuja distribuição pertence à família exponencial.

O principal objetivo deste estudo é desenvolver modelos estatísticos, no contexto de Modelos Lineares Generalizados, para identificar os principais fatores associados à recuperação dos doentes contaminados com a COVID-19. Assim, numa primeira abordagem são estabelecidos modelos de Regressão Logística com o objetivo de se analisar o efeito de diferentes fatores na recuperação (ou não recuperação) de um doente com COVID-19. Os dados utilizados nesta abordagem referem-se a dados de Filipinas observados no mês de fevereiro do ano de 2020.

Numa segunda abordagem pretendeu-se identificar fatores que influenciaram o número de doentes recuperados da COVID-19 estabelecendo Modelos de Regressão de Poisson. No entanto, os modelos desenvolvidos apresentaram o problema de sobredispersão, tornando-se necessário recorrer a Modelos de Regressão Binomial Negativa. Os modelos foram desenvolvidos com aplicação a um conjunto de dados relativos ao número de casos com COVID-19 registados em 130 países em agosto do ano de 2020.

Palavras-chave: COVID-19, Modelos Lineares Generalizados, Regressão Logística, Modelo de Regressão de Poisson, Modelo de Regressão Binomial Negativa.

Abstract

Over the centuries, outbreaks and epidemics of infectious diseases have caused millions of deaths. Currently, the new disease COVID-19 derived from a virus named SARS-CoV-2 by the World Health Organization (WHO) has caused more than 3 million deaths worldwide so far. To better understand the behavior of this new disease, numerous modeling processes have been developed and analyzed, particularly in the area of Generalized Linear Models.

Generalized Linear Models have been widely used in various fields of study to model phenomena. The main purpose of establishing this type of models is to analyze the influence that explanatory variables have on a variable of interest (the response variable) whose distribution belongs to the exponential family.
The main objective of this study is to develop statistical models in the context of Generalized Linear Models to identify the main factors associated with the recovery of patients infected with COVID-19.

Thus, in a first approach, Logistic Regression models are established to analyze the effect of different factors on the recovery (or non-recovery) of a patient with COVID-19. The data used in this approach derives from data collected in the Philippines in the month of February 2020.

In a second stage, it was intended to identify factors influencing the number of patients recovered from COVID-19 by establishing Poisson Regression Models. However, the developed models presented the problem of overdispersion, making it necessary to use Negative Binomial Regression Models. The models were developed with application to a dataset concerning the number of cases with COVID-19 registered in 130 countries in August 2020.

Keywords: COVID-19, Generalized Linear Models, Logistic Regression, Poisson Regression Models, Negative Binomial Regression Models.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AIC- do inglês, Critério de informação de Akaike.

AUC-do inglês, Area Under the Roc Curve.

F.d.p- Função Densidade de Probabilidade.

F.m.p-Função Massa de Probabilidade.

OMS-Organização Mundial de Saúde.

ROC-do inglês, Receiver operating Characteristic.

IDS- Índice de Desigualdade Social.

GTSP- Gasto Total na Saúde por Pessoa.

Capítulo 1

Introdução

São incalculáveis as situações em que existe a necessidade de estudar possíveis relações entre variáveis e analisar a influência que uma ou mais variáveis explicativas têm sobre uma variável de interesse, a variável resposta. O estatístico usualmente aborda tal problema utilizando os Modelos de Regressão Linear que relacionam essa variável de interesse com as variáveis ditas explicativas.

Segundo Turkman (2000), o Modelo de Regressão Linear, "criado"no início do século XIX por Legendre e Gauss, dominou a modelação estatística até meados do século XX, embora vários modelos tenham entretanto sido desenvolvidos para fazer face às situações que não eram adequadamente explicadas por estes modelos.

Para dar resposta às situações em que a variável resposta não segue uma distribuição Normal, os Modelos Lineares Generalizados, que são uma extensão dos Modelos de Regressão Linear, permitem incluir outras distribuições da variável dependente, desde que pertencentes à família Exponencial.

Os Modelos Lineares Generalizados foram apresentados por Nelder (1972) e Robert Wedderburn (1972) e incluem vários modelos estatísticos, incluindo os modelos de Regressão Linear, de Regressão Logística (ou Multinomial), de Regressão de Poisson e de Regressão Binomial Negativa estes consistem em abordagens cuja variável resposta tem como característica comum pertencer à família Exponencial de distribuições e as mesmas englobam diversos tipos de dados, sejam eles quantitativos (discretos ou contínuos) ou qualitativos nominais.

Neste trabalho, pretende-se analisar os dados recolhidos no âmbito de uma investigação levada a cabo na área da saúde nas Filipinas, no mês de fevereiro de 2020, logo no início

da pandemia da COVID-19 naquele país, onde foram diagnosticados 143 doentes com a doença COVID-19, dos quais, nesse período, 93 doentes não recuperaram da doença e 50 recuperaram. A variável resposta a ser estudada é a variável **Status** (Estado do Doente): recuperado ou não recuperado da doença COVID-19.

Numa segunda abordagem analisou-se os dados recolhidos no âmbito da investigação levada a cabo na área da saúde a nível mundial efetuada pela Organização Mundial da saúde (OMS) no dia 3 de agosto do ano 2020, em que foram estudados dados da COVID-19 em 130 países. Neste estudo a variável resposta é quantitativa discreta, valores que correspondem ao número total de recuperados por COVID-19 naquele dia.

Para isso foram desenvolvidos modelos estatísticos na área dos Modelos Lineares Generalizado que permitem indicar fatores associados à recuperação dos doentes com esta patologia.

Assim, o principal objetivo deste trabalho é desenvolver modelos estatístico para identificar os principais fatores associados à recuperação dos doentes contaminados com a COVID-19 utilizando Modelos Lineares Generalizados, em particular Modelos de Regressão Logística e Modelos de Regressão de Poisson/Regressão Binomial Negativa.

O tratamento de dados será realizado através do software *Rstudio*. Todos os gráficos apresentados ao logo desta dissertação foram integralmente realizados pelo mesmo.

1.1 Composição do Trabalho

O presente trabalho está dividido em seis capítulos que, em seguida, serão descritos da seguinte forma.

No Capítulo 1 é descrita uma breve introdução ao tema em análise.

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

No Capítulo 3 apresenta em detalhe o modelo de Regressão Logística para dados binários (com a distribuição de Bernoulli).

No Capítulo 4 são descritos os conteúdos teóricos relacionados com os modelo de Regressão de Poisson e de Regressão Binomial Negativa para dados de contagem.

No Capítulo 5 apresentam-se os resultados da aplicação destes modelos a dados reais de Saúde. Inicia-se o estudo com uma análise descritiva das bases de dados, seguindo-se a formulação e a discussão dos modelos de regressão para dados binários e para dados de contagem de forma a inferir sobre quais as variáveis com poder explicativo sobre as variáveis resposta de interesse, selecionando os modelos que melhor se ajustam aos dados.

No Capítulo 6 são apresentadas as principais conclusões e algumas sugestões para trabalho futuro.

Capítulo 2

Modelos Lineares Generalizados

A formulação de um modelo linear generalizado compreende a escolha de uma distribuição de probabilidade para a variável resposta, das variáveis quantitativas e/ou qualitativas para representar a estrutura linear do modelo e de uma função de ligação. Para a melhor escolha da referida distribuição de probabilidade é aconselhável examinar os dados para observar algumas características, tais como a assimetria, a natureza discreta ou contínua, o intervalo de variação, etc. É importante salientar que os termos que compõem a estrutura linear do modelo podem ser de natureza discreta ou contínua, qualitativa ou mista, e devem dar uma contribuição significativa na explicação da variável resposta, Cordeiro (2004).

2.1 Família Exponencial

Uma variável aleatória Y tem distribuição pertencente à família Exponencial de dispersão (ou simplesmente família Exponencial) quando a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) for escrita na forma

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

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

Quando ϕ for conhecido tem-se uma distribuição da família Exponencial com parâmetro canónico θ.

Pode ser demonstrado, McCullagh e Nelder (1989) que, se Y é uma variável aleatória com uma distribuição pertencente à família Exponencial, então

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

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

onde b'(θ) e b''(θ) são a primeira e a segunda derivadas de b(θ), respectivamente. Assim, a variância de Y é o produto de duas funções a b''(θ) que depende apenas do parâmetro canónico θ que se designa por função de variância de μ e que se representa por, Var(μ), e outra, a(ϕ), que depende apenas do parâmetro de dispersão ϕ, Turkman (2000).

2.1.1 Exemplos de algumas Distribuições Conhecidas Pertencentes à Família Exponencial

Distribuição Binomial

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

$$f(y|n,\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \tag{2.4}$$

$$= \exp\left[y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}\right].$$

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

A média e a variância de Y são representadas por

$$E(Y) = \mu = b'(\theta) = n\pi,$$

$$Var(Y) = a(\phi)b''(\theta) = n\pi(1-\pi).$$

A função de variância é Var(μ)=$n\pi(1-\pi)$.

Distribuição de Bernoulli

Uma variável aleatória Y tem distribuição de Bernoulli com parâmetro π se sua função de probabilidades é dada por

$$= \exp\left[y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right].$$

$$f(y|\pi) = \pi^y (1-\pi)^{1-y} = \tag{2.5}$$

Obtém-se, $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, b($\theta$)=log(1 + exp($\theta$)), a($\phi$)=1 e c(y, ϕ)=0.

A média e a variância de Y são representadas por

$$E(Y) = \mu = b'(\theta) = \pi,$$

$$Var(Y) = a(\phi)b''(\theta) = \pi(1-\pi).$$

A função de variância é Var(μ)=π(1 − π).

Distribuição de Poisson

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

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!} = \tag{2.6}$$

$$= \exp\left[y \log(\mu) - \mu - log(y!)\right].$$

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

A média e a variância de Y são respetivamente

$$E(Y) = b'(\theta) = \exp(\theta) = \mu$$

e

$$Var(Y) = a(\phi)b''(\theta) = \exp(\theta) = \mu.$$

A função de variância Var(μ) = μ.

Distribuição Binomial Negativa

Seja Y uma variável aleatória que segue uma distribuição Binomial Negativa com parâmetros k e p, y ~ BN(k, p). A variável Y representa o número de insucessos anteriores a k sucessos, num conjunto de acontecimentos independentes e com a mesma probabilidade de sucesso, p. A função de probabilidade de Y é dada por

$$f(y|n,p) = \binom{y+k-l}{k-l} p^k (1-p)^y = \tag{2.7}$$

$$= \exp\left[y\log(1-p) + k\log(p) + \log\binom{y+k-l}{k-l} \right].$$

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

A média e a variância são expressas por

$$E(Y) = \mu = b'(\theta) = \frac{k(1-p)}{p},$$

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

A função de variância

$$Var(\mu) = \frac{k(1-p)}{p^2}.$$

Distribuição Normal

Seja Y uma variável aleatória que segue uma distribuição Normal com valor médio μ e variância σ^2, se a sua função densidade de probabilidade é dada por

$$f(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(y-\mu)^2 \right\}, y \in \mathbb{R}. \tag{2.8}$$

Tem-se, então,

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

Comparada com (2.1), tem-se $\theta = \mu$, b(θ)=$\frac{\mu^2}{2}$, a(ϕ)=σ^2 e c(y,ϕ)=-$\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right)$.
A média e a variância de Y são representadas por

$$E(Y) = b'(\theta) = \theta = \mu,$$

$$Var(Y) = a(\phi)b''(\theta) = \theta = \sigma^2.$$

A função de variância é Var (μ)=1.

O que mostra que a distribuição $N(\mu, \sigma^2)$ com μ desconhecido e $\sigma^2 > 0$, conhecido, pertence à família Exponencial na forma canónica.

Distribuição Gama

Se Y tem distribuição Gama com parâmetro de forma ν e de escala, ν/μ, ($Y \sim Ga(\nu, \nu/\mu)$), a sua função densidade de probabilidade é

$$f(y|\nu, \mu) = \frac{1}{\Gamma(\mu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu} y\right) = \tag{2.9}$$

$$= \exp\{\nu(-\frac{y}{\mu} - \log\mu) + (\nu - 1)\log y - \log\Gamma(\nu) + \nu\log\nu\} =$$

$$= \exp\left\{\nu(\theta y + \log(-\theta)) + (\nu - 1)\log y - \log\Gamma(\nu) + \nu\log\nu\right\},$$

com $y > 0$ e $\theta = -\frac{1}{\mu}$.
Logo para a função de probabilidade da forma (2.1) apresentada, tem-se

$\theta = -\frac{1}{\mu}$,
$b(\theta) = -\log(-\theta)$,
$a(\phi) = \frac{1}{\nu}$,
$c(y, \phi) = (\nu - 1)\log y + \nu\ln\nu - \log\Gamma(\nu)$.
A média e a variância de Y são representadas por

$$E(Y) = b'(\theta) = \frac{1}{\theta} = \mu,$$

$$Var(Y) = a(\phi)b''(\theta) = \frac{\mu^2}{\nu}.$$

A função de variância é Var(μ)=μ^2.

2.2 Descrição do Modelo Linear Generalizado

Nelder e Wedderburn (1972) deram o nome de Modelo Linear Generalizado como uma extensão do modelo linear clássico

$$Y = X\beta + \epsilon$$

onde X é uma matriz de dimensão n×(p+1) de especificação do modelo.

Em geral, a matriz das covariáveis de interesse X com um primeiro vetor unitário, associada a um vetor $\beta = (\beta_0, \beta_1 \ldots, \beta_p)^T$ de parâmetros, e ϵ é um vetor de erros aleatórios com distribuição que se supõe $N_p(0, \sigma^2 I)$.

Estas hipóteses implicam obviamente que o valor esperado da variável resposta é uma função linear das covariáveis

$$E(Y|X) = \mu = X\beta.$$

Portanto, a extensão é feita em duas vertentes: a distribuição considerada não tem de ser Normal, mas que pode pertencer a qualquer família Exponencial; a estrutura de linearidade mantém a função que associa o valor esperado e o vetor de covariáveis de interesse $X = (X_1, ..., X_p)$ pode ser qualquer função diferenciável.

As componentes fundamentais do Modelo Linear Generalizado são:

- Componente aleatória;
- Componente sistemática;
- Função de ligação.

Componente aleatória

Dado o vetor de covariáveis $X = (X_1, ..., X_p)$ as variáveis Y_i, com $i = 1, \ldots, n$, são (condicionalmente) independentes com distribuição pertencente à família Exponencial da forma (2.1), com $E[y_i|_{X_j}] = \mu_i = b'(\theta_i)$, para $i = 1, \ldots, n$ e $j = 1, \ldots, p$, possivelmente, um parâmetro de dispersão ϕ não dependente de i.

Componente sistemática

Consiste numa combinação linear de variáveis preditoras, ou seja, o valor esperado μ_i está relacionado como preditor linear $\eta_i = X_i^T\beta$ através da relação

$$\mu_i = h(\eta_i) = h(X_i^T\beta), \eta_i = g(\mu_i)$$

onde, h é uma função monótona e diferenciável;
β é um vetor de parâmetros de dimensão p;
x_j é o vector de especificação de dimensão p;
$g = h^{-1}$ é a função de ligação que relaciona a média de y_i ao preditor linear, ou seja,

$$g(\mu_i) = \eta_i, i = 1, \ldots, n.$$

Quando existem covariáveis qualitativas elas devem ser codificadas à custa de variáveis binárias mudas chamadas indicatrizes (ou *dummy*).

Função de ligação

Função diferenciável e monótona g(.) que associa a componente aleatória e sistemática, através duma relação da forma

$$g(\mu_i) = \eta_i = x_i^T\beta = \sum_{j=1}^{p} x_{ij}\beta_j \longleftrightarrow \mu_i = g^{-1}(x_i^T\beta).$$

Quando o preditor linear coincide com o parâmetro canónico, isto é, $\theta_i = \eta_i = X_i^T\beta$. A função de ligação correspondente diz-se então função de ligação canónica.

Em síntese, a estrutura de um Modelo Linear Generalizado é formada por três partes: uma componente aleatória composta de uma variável aleatória Y com n observações independentes, um vetor de médias μ e uma distribuição pertencente à família Exponencial; uma componente sistemática composta por variáveis explicativas $X_1, \ldots, X_p$ tais que produzem um preditor linear η; e uma função monótona diferenciável, conhecida como função de ligação, que relaciona estas duas componentes (Cordeiro (2004)).

2.3 Estimação dos Parâmetros

Num Modelo Linear Generalizado o parâmetro β é parâmetro de interesse, o qual é estimado pelo método da máxima verosimilhança. Nos Modelos Lineares Generalizados, os métodos de inferência estatística baseiam-se, essencialmente, na função de verosimilhança. O parâmetro de dispersão ϕ, quando existe, é considerado um parâmetro perturbador e é estimado pelo método dos momentos. Geralmente o método de máxima verosimilhança é também considerado como base fundamental no processo inferencial, no caso dos testes de hipóteses sobre os coeficientes estimados e da qualidade do ajustamento.

A função de verosimilhança do modelo, em função de β, é dada por Turkman (2000)

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi) \; = \tag{2.10}.$$

$$= \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} = \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)\right\}.$$

O logaritmo da verosimilhança é dado por

$$\log(L(\beta)) = \ell(\beta) =$$

$$= \sum_{i=1}^{n} \left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} = \sum_{i=1}^{n} \ell_i(\beta),$$

onde, ℓ_i é a contribuição de cada observação y_i para a verosimilhança.

O estimador de máxima verosimilhança para β são obtidos como solução do sistema de equações de verosimilhança e as mesmas são dadas por

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

A equação (2.11) é a derivada do logaritmo da verosimilhança em relação ao parâmetro β e pode-se chamar de *Score*.

Para obter estas equações, segundo Turkman (2000), utiliza-se a regra de cadeia

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

Sabendo que $b'(\theta_i) = \mu_i$ e $\text{Var}(Y_i) = \phi b''(\theta_i)$, então

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

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

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

Assim,

$$\frac{\partial \ell_i(\beta)}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_{ij}; \tag{2.12}$$

e a equação de verosimilhança para β são

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_{ij} = 0, j = 1, \ldots, p. \tag{2.13}$$

A função Score $S(\beta)$ é o vector p-dimensional formado pelas derivadas parciais de primeira ordem do logaritmo da função de verosimilhança.

Logo a função *score* é obtida por

$$S(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta), \tag{2.14}$$

onde $s_i(\beta)$ é o vetor de componente $\left(\frac{\partial \ell_i(\beta)}{\partial \beta_j}\right)$ obtidas na equação(2.11).

Quando se considera um Modelo Linear Generalizado com a função de ligação canónica, a matriz Hessiana da log-verosimilhança não depende da variável resposta Y, pelo que a *Hessiana* e o seu valor esperado coincidem. Logo, neste caso os métodos de *Fisher* e *Newton-Raphson* coincidem. Esta é uma das razões que confere às ligações canónicas, Cadima (2018).

A matriz de covariância da função *score* é denominada por *Matriz de Informação de Fisher* que é representada por

$$I(\beta) = E\left[-\frac{\partial s(\beta)}{\partial(\beta)}\right]. \tag{2.15}$$

Para se obter a *Matriz de Informação de Fisher* temos que considerar o valor esperado das segundas derivadas $\ell_i(\beta)$.
Tem-se, para família regulares, que

$$-E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = E\left(\frac{\partial \ell_i}{\partial \beta_j}\frac{\partial \ell_i}{\partial \beta_k}\right) =$$

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

$$= \frac{\tilde{x}_{ij}\tilde{x}_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Assim, o elemento genérico de ordem (j, k) da *Matriz de Informação de Fisher* é dado por

$$-\sum_{i=1}^{n} E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{n} \frac{\tilde{x}_{ij}\tilde{x}_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{2.16}$$

2.4 Teste de Hipóteses

Os métodos de inferência nos Modelos Lineares Generalizados baseiam-se, fundamentalmente, na teoria de máxima verosimilhança. Uma vez feita a estimativa para os coeficientes de regressão, vamos agora analisar a sua significância, isto é, determinar se as variáveis explicativas introduzidas no modelo estão significativamente associadas à variável resposta. Os testes de hipóteses que se vão apresentar são:

- *O Teste de Wald*,
- *O Teste de Razão de Verosimilhanças*.

2.4.1 Teste de Wald

O teste de Wald, baseada na normalidade assintótica do estimador de máxima verosimilhança. É, em geral, mais usada para testar hipóteses nulas sobre componentes individuais, ou seja o teste Wald utiliza-se para testar a hipótese nula de que o parâmetro β_j estimado é igual a zero..
As hipóteses a testar são:

$H_0 : \beta_j = 0,\ \textit{versus}\ H_1 : \beta_j \neq 0, j = 0, ..., p$

que indica que o coeficiente independente β_0 =0 é irrelevante para o modelo (j = 0) ou se a variável explicativa X_j não é estatisticamente significativa para o modelo de regressão (j≠ 0)

A estatística de teste para grandes amostras é

$$ET = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1).$$

2.4.2 Teste de Razão de Verosimilhanças

O teste da razão de verosimilhanças é amplamente utilizado em Modelos Lineares Generalizados para testar a nulidade conjunta de $q < p$ parâmetros a partir de modelos encaixados, ou seja, o teste é aplicado para a comparação dos modelos em que um deles é submodelo do outro e de igual modo para avaliar a significância dos coeficientes estimados simultaneamente.

Dados dois modelos aninhados, M_p e M_q, com um número de variáveis p e q respetivamente, tal que $q < p$, para comparar a qualidade de ajustamento de dois modelos aplica-se o teste da razão de verosimilhanças, sob as hipóteses
H_0 : Os dois modelos têm a mesma qualidade de ajustamento;
H_1 : Os dois modelos não têm a mesma qualidade de ajustamento.

A estatística de teste é definida por

$$\begin{aligned} G &= -2(\log(\ell_{M_q}(\beta)) - \log(\ell_{M_p}(\beta))), \\ &= -2\log\left[\frac{\ell_{M_q}(\beta)}{\ell_{M_p}(\beta)}\right]. \end{aligned} \tag{2.17}$$

A estatística de teste segue uma distribuição de Qui-quadrado com (p-q) graus de liberdade, i.e.,

$$G \sim \chi^2_{p-q}$$

onde $\log(L_{M_p}(\beta))$ é a função logaritmo da verosimilhança do modelo M_p com p variáveis e $\log(L_{M_q}(\beta))$ é a função logaritmo da verosimilhança do modelo M_q com q variáveis.

2.5 Seleção do Modelo

Uma vez que a etapa de seleção do modelo final é muito importante no processo de modelação estatística, quando o processo envolve muitas covariáveis temos interesse em saber qual é modelo mais adequado, isto é, com o menor número de variáveis independentes (parcimonioso), que ofereça uma boa interpretação do problema posto e que ainda se ajuste bem aos dados.

Existem três métodos de seleção para ajustar um modelo:

Forward: parte do modelo nulo, isto é, sem variáveis explicativas, testa-se a adição de cada variável no modelo, começa-se por incluir a variável que apresenta o menor valor de prova (essa variável é aquela à qual corresponde a estatística de teste com valor absoluto mais alto), até que não existam mais variáveis cuja adição ao modelo produza alterações significativas ao mesmo, isto é, até não existirem variáveis que passem um pré-determinado critério de inclusão. O modelo nulo é um modelo simples com apenas um parâmetro que representa o mesmo valor médio μ para todas as observações y_i.

Backward: ao contrário do método *forward*, este parte do modelo com todas as variáveis explicativas, isto é, o modelo completo, as variáveis vão sendo retiradas sucessivamente do modelo, até que todas as variáveis sejam estatisticamente significativas no modelo.

Stepwise (both): é a combinação dos dois modelos anteriores, que testa em cada passo as variáveis que devem ser incluídas ou excluídas, partindo do modelo nulo. No entanto, em cada passo é feita uma análise das variáveis já introduzidas até aí, por forma a garantir que permaneçam relevantes após a introdução da nova variável.

2.6 Qualidade de Ajustamento

2.6.1 Função Desvio

A Função Desvio é utilizada para testar a significância dos coeficientes estimados do modelo. Ou seja é uma medida de avaliação da qualidade do ajustamento, em que quanto menor é o Desvio, melhor o ajustamento do modelo. Esta medida de discrepância é baseada no teste de Razão de Verosimilhanças.

Comecemos por relembrar que o modelo nulo é o modelo mais simples onde somente o parâmetro constante é estimado, apresentando por isso um menor valor da função de verosimilhança. Por sua vez o modelo completo ou saturado é o modelo que estima um parâmetro para cada observação, isto é, as estimativas de máxima verosimilhança são as próprias observações, $\hat{\mu}_i = y_i$, Santos (2013).

O modelo completo ou saturado é útil para avaliar a qualidade do ajustamento de um determinado modelo ajustado aos dados, através da introdução de uma medida de

distância que representa o desvio do modelo ajustado em relação ao modelo saturado. Quanto mais próximo o modelo ajustado, $\hat{\mu}$, estiver dos valores observados, y, menor será o valor dessa distância, Turkman (2000).

Se assumirmos que não existem pesos associados às observações, isto é, $a(\phi) = \phi$, o critério da razão de verosimilhanças entre o modelo ajustado (Mq) e o modelo completo (Mp) é dada por

$$-2(\ell_q(\beta) - \ell_p(\beta)) = -2\left[\sum_{i=1}^{n} \frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{\phi} + c(y_i, \phi) - \left(\sum_{i=1}^{n} \frac{y_i\bar{\theta}_i - b(\bar{\theta}_i)}{\phi} + c(y_i, \phi)\right)\right] =$$

$$= 2\sum_{i=1}^{n} \frac{y_i(\bar{\theta}_i - \hat{\theta}_i) - b(\bar{\theta}_i) + b(\hat{\theta}_i)}{\phi} =$$

$$= \frac{D(y, \hat{\mu})}{\phi},$$

onde, ℓ_q corresponde ao logaritmo da função de verosimilhança do modelo ajustado (M_q), ℓ_p ao logaritmo da função de verosimilhança do modelo completo (M_p), y_i é o valor ajustado da observação i dada pelo modelo (M_q), $\hat{\theta}_i$ os parâmetros estimados pelo modelo (M_q), $\bar{\theta}_i$ os parâmetros do modelo (M_p). A quantidade D(y,$\hat{\mu}$), também conhecida por Desvio, é dada por

$$D(y, \hat{\mu}) = -2\phi(\ell_q(\beta) - \ell_p(\beta)) =$$

$$= 2\sum_{i=1}^{n} y_i(\bar{\theta}_i - \hat{\theta}_i) - b(\bar{\theta}_i) + b(\hat{\theta}_i).$$

A função Desvio de um modelo avalia, portanto, a discrepância entre os valores ajustados pelo modelo completo e os valores ajustados pelo modelo em estudo. O valor de D é sempre maior ou igual a zero e será tanto maior, quanto maior for a discrepância entre o modelo ajustado e os valores observados. Para se avaliar se um determinado modelo M_q, se ajusta bem aos dados, considera-se o teste de hipóteses com a seguinte hipótese nula:
H_0 : O ajustamento do modelo M_q é igual ao ajustamento do modelo M_p.

Sob H_0 e para amostras grandes, D apresenta uma distribuição assimptótica Qui-quadrado com χ^2_{j-q-1} graus de liberdade,

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

onde J é o número de parâmetro de covariáveis diferente existente nos dados, q é o número de parâmetros do modelo M_q a menos da constante.

2.6.2 Critério de Informação de Akaike

Quando dois modelos não são aninhados, não é possível utilizar a razão de verosimilhanças, pelo que se torna aconselhável utilizar um critério que meça a quantidade de informação que o modelo recolhe dos dados, por oposição ao ruído que o modelo não consegue explicar.

O Critério de Informação de Akaike (AIC) é uma medida de qualidade de ajuste baseada na log-verosimilhança e penaliza o modelo com muitas variáveis. A sua construção é amplamente utilizada no processo de seleção de modelos, na busca por um modelo bem ajustado aos dados, mas com poucos parâmetros. O AIC para um modelo é definido como

$$AIC = -2\log(L) + 2p, \tag{2.18}$$

sendo $\log(L)$ a log da função de verosimilhança do modelo e p o número de parâmetros do modelo. Assim, quanto menor o valor do AIC, melhor o modelo. Salienta-se que o AIC não traduz, no sentido absoluto, se um modelo faz um bom ajustamento aos dados ou não. Apenas revela se um determinado modelo é preferível relativamente a outro.

2.6.3 Análise de Resíduos

Segundo Turkman (2000), a análise de resíduos é útil, não só para uma avaliação local da qualidade de ajustamento de um modelo no que diz respeito à escolha da distribuição, da função de ligação e de termos do preditor linear, como também para ajudar a identificar observações mal ajustadas.

As técnicas para a análise de resíduos nos Modelos Lineares Generalizados são semelhantes ao do modelo clássico de Regressão.

Resíduo de Pearson

O resíduo de Pearson é calculado por

$$r_{ip} = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}, i = 1, \ldots, n \tag{2.19}$$

onde, $Var(\hat{\mu}_i)$ representa a função de variância estimada para a distribuição do modelo em estudo.

Outra medida da adequabilidade de modelos é a estatística de Pearson generalizada que é dada por

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{Var(\hat{\mu}_i)}. \tag{2.20}$$

Resíduo do Desvio

O resíduo da função Desvio correspondente à i-ésima observação é definido por

$$R_i^D = \delta_i \sqrt{d_i}, \tag{2.21}$$

onde, $\delta_i = sinal(y_i - \hat{\mu}_i)$, d_i é a contribuição da i-ésima observação para a medida do desvio definida em 2.6.1.

Resíduos Padronizados

O resíduo de Pearson padronizado é definido por

$$r_{iE} = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)(1 - h_{ii})}}. \tag{2.22}$$

sendo, h_{ii} o i-ésimo elemento da diagonal da matriz H definida por

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2},$$

sendo, W definida quando da discussão do Método de Fisher, Turkman (2000). O resíduo de Desvio padronizado é dado por

$$r_{DE} = \frac{r_D}{\sqrt{(1 - h_{ii})}}. \tag{2.23}$$

2.6.4 Tipos de Observações

Pretende-se detetar observações atípicas do conjunto de dados. Essas observações são classificadas em:

Outliers: são observações mal ajustados com resíduos altos e elas podem ser, ou não, influentes;

Pontos influentes: são observações com influência desproporcional nas estimativas dos coeficientes do modelo, isto é, quando excluídos do modelo mudam de forma substancial algumas propriedades do modelo ajustado.

Pontos alavanca: são observações posicionados em regiões remotas no espaço das variáveis explicativas com grande impacto na determinação das propriedades do modelo de regressão.

As observações *outliers*, influentes e alavanca são identificados através da análise de resíduos e da medida h_{ii}, onde h_{ii} é o elemento da diagonal da matriz de projeção.

Segundo Cordeiro (2004), é muito razoável utilizar h_{ii} como uma medida da influência da i-ésima observação sobre o próprio valor ajustado. Supondo que todos os pontos exerçam a mesma influência sobre os valores ajustados, podemos esperar que h_{ii} esteja próximo de $\frac{p}{n}$, em que p é o número de parâmetro no modelo e n é o total das observações. Portanto, convém examinar as observações correspondentes aos maiores valores de h_{ii}. Alguns autores sugerem $h_{ii} > \frac{2p}{n}$ como um indicador de pontos influentes. Para avaliar de uma forma mais geral a influência da i-ésima observação nas estimativas dos coeficientes da regressão utiliza-se a medida de distância de Cook que é dada por

$$DC_i = \frac{h_{ii} r_{ip}^2}{p(1 - h_{ii})}.$$

Logo, DC_i será elevado quando o valor de h_{ii} é diferente de zero e resíduos elevados, então para valores elevados de DC_i considera-se a respetiva observação como influente.

2.6.5 Tipos de Gráficos

A representação gráfica é uma forma informal de avaliar a qualidade de ajustamento de um modelo. Existem vários tipos de gráficos para analisar a qualidade de ajustamento, mas neste trabalho vai-se usar três tipos:

1)-Gráfico dos resíduos padronizados *versus* valores ajustados, este gráfico pode ser útil na detecção de observações que divergem da tendência geral das demais observações, observações que estão fora do limite considerado para a distribuição dos resíduos, indicando possíveis *outliers.*

2)- Gráfico Normal de probabilidades para resíduos com envelope, analisa o pressuposto da normalidade dos resíduos e da escolha da distribuição para a variável resposta. Se o modelo ajustado é o correto, existe grande probabilidade de que todos os pontos estejam dentro do envelope.

3)- Gráficos de h_{ii} e DC_i *versus* a ordem da observação, são geralmente úteis na identificação de pontos alavanca e pontos influentes.

Capítulo 3

Modelo de Regressão Logística

A Regressão Logística tem-se constituído num dos principais métodos de modelação estatística de dados. Mesmo quando a resposta de interesse não é originalmente do tipo binário, alguns investigadores têm dicotomizado a resposta de modo que a probabilidade de sucesso possa ser ajustada através da Regressão Logística (Paula (2004)).

O modelo de Regressão Logística é um caso particular do Modelo Linear Generalizado, que geralmente é usado quando a variável resposta é qualitativa com dois resultados possíveis denominados de "fracasso"e "sucesso".

Seja x_{ij} a variável explicativa e y_i o número de ocorrências de um determinado evento, em que $i = 1, 2, \ldots, n$ representa o número de observações e $j = 1, 2, \ldots, p$ representa o número de covariáveis. Assume-se ainda, que a variável resposta tem distribuição de Bernoulli (π_i) com $\pi_i = P(y_i = 1) = E(Y_i)$ e cuja função de probabilidade é dada por

$$f(y_i|\pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, y_i = 0, 1, i = 1, \ldots, n \tag{3.1}$$

onde, π_i é a probabilidade de ocorrência de um evento, que significa a probabilidade do sucesso $P(y_i = 1) = \pi_i$ e a probabilidade do fracasso $P(y_i = 0) = 1 - \pi_i$.

O objetivo é formular um modelo para a probabilidade de um objeto ou indivíduo caracterizado por um vetor de variáveis explicativas tomar o valor 1.

No entanto, no modelo de regressão clássico o valor esperado é dado por

$$E(Y|\mathbf{x_i} = x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, i = 1, \ldots, n.$$

Assim sendo, é possível que a média assuma qualquer valor, quando x varia entre

$-\infty$ e $+\infty$, existindo assim incorrespondência ao contradomínio do modelo de Regressão Logística, isto é, quando
$g(x) \to +\infty$, então $P(Y = 1) \to 1$;
$g(x) \to -\infty$, então $P(Y = 1) \to 0$.

Para que haja correspondência, a Regressão Logística reformula o modelo linear de modo a conceder que o valor da variável resposta varie entre 0 e 1. A mesma é obtida pela seguinte equação

$$\pi_i = P(Y_i = 1|\mathbf{x}) \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}. \tag{3.2}$$

Para descrever a relação linear entre a variável resposta e as variáveis explicativas faz-se o uso da função logit que é o logaritmo da razão entre a probabilidade de sucesso e a probabilidade de insucesso. A equação da função logit é

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \tag{3.3}$$

onde π_i é uma proporção de Bernoulli, x_j (j=1, ..., p) é a variável explicativa e $\beta_0, \beta_1, \ldots, \beta_p$ são parâmetros do modelo de Regressão Logística, chamada de razão de desigualdade.

De acordo com Agresti (2013), uma das principais estatísticas utilizadas na análise de dados binários é a razão de chances $\left(\frac{\pi_i}{1-\pi_i}\right)$, que é definida como a razão entre a chance de um evento ocorrer num grupo, sendo que a chance é a probabilidade de ocorrência deste evento dividida pela probabilidade da não ocorrência do mesmo evento.

3.1 Estimação dos Coeficientes de Regressão

Para a estimação dos coeficientes de regressão, quando a variável resposta é binária, partindo do pressuposto que existe independência dos valores observados, utiliza-se o método de máxima verosimilhança descrito na Secção 2.3. Neste caso a função de verosimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)}.$$

Na prática usa-se o logaritmo da função verosimilhança (ou log-verosimilhança)

para simplificar a tarefa de obtenção dos estimadores e é dado por

$$\log(\beta) = l(\beta) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right].$$

Substituindo na expressão (3.2) fica

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}}{1-\frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}} \right) + \log\left(1 - \frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}\right) \right] =$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\exp\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) + \log\left(\frac{1}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})} \right) \right) \right].$$

Logo,

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) - \log\left(1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})\right) \right].$$

O valor de β que maximiza $\ell(\beta)$ é obtido após derivar $\ell(\beta)$ em relação aos parâmetros $(\beta_0, \beta_1, \ldots, \beta_p)$. Caso não seja possível uma solução analítica serão necessários métodos iterativos para a sua resolução.

3.2 Qualidade de Ajustamento

Para avaliar o ajustamento do modelo são utilizados alguns testes como o teste de *Wald* e o teste de Razão de Verosimilhança que têm como objetivo avaliar a significância de cada variável explicativa incluída no modelo. O teste normalmente usado para avaliar o ajustamento de modelos de Regressão Logística é o teste de *Hosmer-Lemeshow*, que verifica se existe uma associação estatisticamente significativa entre as variáveis preditoras e a variável resposta.

3.2.1 Teste de Hosmer e Lemeshow

Este teste avalia o modelo ajustado, comparando as frequências observadas e as esperadas. O teste associa os dados às suas probabilidades estimadas da mais baixa à mais alta, e então aplica um teste de Qui-quadrado para determinar se as frequências estimadas estão próximas das frequências observadas, Hosmer (1989).

A hipótese a testar é H_0: "O modelo encontrado explica bem os dados"

$$HL = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{n_k \pi_k (1 - \pi_k)} \sim \chi^2_{p-1} \tag{3.4}$$

onde,
k - é o número de grupos (exemplo 10 grupos compostos pelos decis do valor ajustado da probabilidade);
n_k - número de indivíduos em cada grupo;
o_k - número de respostas positivas dentro de cada grupo;
e_k - valor esperado do número de casos dentro de cada grupo assumindo que o modelo está correto.

O teste de diagnóstico de *Hosmer-Lemeshow* modificado verifica se as probabilidades estimadas a partir do modelo são consistentes com a resposta binária observada. Para tal ordena as probabilidades estimadas para cada observação, divide-as em (10) grupos de sensivelmente o mesmo número de probabilidades, calcula as probabilidades médias dentro de cada grupo e, multiplicando-as pelo número de observações do grupo, obtém o número esperado de sucessos nesse grupo. Esse número é então comparado com o número efetivo de sucessos observados no grupo através de um teste de Qui-quadrado de Pearson.

3.2.2 Curva ROC

Seja $\hat{Y} = 1$ se um indivíduo selecionado na população em estudo for classificado como acontecimento de interesse e $\hat{Y} = 0$ se classificado como não acontecimento. Para esta classificação é necessário estabelecer um ponto de corte que determina a probabilidade de um dado indivíduo ser classificado numa determinada classe. O ponto de corte mais utilizado é C = 0,5, significando que para um valor $\hat{Y}$ ser maior ou igual a 0,5 o indivíduo será classificado na classe 1, caso contrário será classificado na classe 0. A curva *Receiver Operating Characteristic* (ROC) é um método utilizado para medir a capacidade de predição do modelo. Esta curva representa a sensibilidade, a probabilidade de se detetar os verdadeiros positivos, contra a especificidade, probabilidade de se detetar os verdadeiros negativos, permitindo estudar a sua variação para diferentes pontes de corte. A área sob a curva ROC, denomina-se de *Area Under the ROC Curve* (AUC), que indica a capacidade do modelo discriminar corretamente as variáveis, variando entre 0 e 1, segundo Hein (2010):

- Se AUC = 0,5 não há discriminação;

- Se 0,6 $\leq$ AUC $< 0,7$ o modelo apresenta uma discriminação limitada;
- Se 0,7 $\leq$ AUC $< 0,8$ o modelo apresenta uma discriminação aceitável;
- Se 0,8 $\leq$ AUC $< 0,9$ o modelo apresenta uma excelente discriminação;
- Se AUC $\geq$ 0,9 o modelo apresenta uma discriminação quase perfeita.

3.2.3 Matriz de Confusão

Uma maneira prática de Avaliar o ajustamento de um modelo de Regressão Logística é pela projeção do modelo na Tabela de Classificação (ou Matriz de Confusão). Para isto, precisa-se criar uma tabela com o resultado da classificação cruzada da variável resposta, de acordo com uma variável dicotómica em que os valores se derivam das probabilidades logísticas estimadas na regressão, Mair (2008) .

Para a classificar os "eventos"dos "não eventos", elabora-se a Matriz de Confusão ou Tabela de Classificação, com as observações de Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN) (Tabela 3.1).

Tabela 3.1: Matriz de confusão

		Valor Observado	
		1	0
Valor estimado	1	VP	FP
	0	FN	VN

Sensibilidade: representa a proporção de verdadeiros positivos, ou seja, a capacidade do modelo em avaliar o evento dado Y=1, isto é,

$$SENS = \frac{VP}{VP + FP}.$$

Especificidade: fornece a proporção de verdadeiro negativos, valor este que é determinado pela probabilidade de prevermos a não ocorrência do evento entre os indivíduos em que este não foi observado, o não evento Y=0, isto é,

$$ESPE = \frac{VN}{VN + FN}.$$

Um bom modelo é aquele em que a sensibilidade e a especificidade são superiores a 80%, razoável se estes dois valores estiverem entre 50% e 80% e medíocre se ambos forem inferiores a 50%.

3.3 Interpretação dos Coeficientes

A interpretação dos parâmetros num modelo de Regressão Logística baseia-se em razões de chances *(odds ratios)*. O *odds ratio* é dado pelo quociente entre a *odds* do acontecimento de interesse ocorrer ($Y = 1$) nos indivíduos com $x = 1$ e a *odds* desse acontecimento ocorrer ($Y = 1$) nos indivíduos com $x = 0$.

Uma vez ajustado o modelo e avaliada a significância dos coeficientes estimados, é necessário interpretar os valores associados aos coeficientes do modelo. Para interpretarmos os valores associados aos coeficientes do modelo de Regressão Logística, é conveniente proceder à análise de acordo com a natureza das variáveis explicativas e as mesma podem ser categóricas ou não.

Variável independente dicotómica

O *odds ratio* é dado pelo quociente entre a *odds* do acontecimento de interesse ocorrer ($Y = 1$) nos indivíduos com $x = 1$ e a *odds* desse acontecimento ocorrer ($Y = 1$) nos indivíduos com $x = 0$.

Assim, quando x é uma variável independente e a mesma pode assumir dois valores 0 ou 1, na qual pode-se construir a tabela de contingência com a probabilidade que se pretende estimar, isto é, quando Y tem a seguinte distribuição de probabilidade $\pi_1 = P(Y = 1|X = 1)$ e $\pi_0 = P(Y = 1|X = 0)$:

	x=1	x=0
y=0	1-π_1	1-π_0
y=1	π_1	π_0

Considerando a expressão (3.2) para se obter o π_0 e π_1 calcula-se

$$\pi_1 = \frac{\exp(\beta_1 + \beta_2)}{1 + \exp(\beta_1 + \beta_2)},$$

$$\pi_0 = \frac{\exp(\beta_1)}{\exp(\beta_1)},$$

em que as chances representam-se da seguinte maneira

$$\frac{\pi_1}{1 - \pi_1} = \exp(\beta_1 + \beta_2)$$

e

$$\frac{\pi_0}{1-\pi_0} = \exp(\beta_1).$$

Quando a variável resposta assumir o valor 1 com $x = 1$ a razão de chance é $\frac{\pi_1}{1-\pi_1}$, da mesma forma a chance da variável resposta assumir valor 1 com $x = 0$, esse *odds* é $\frac{\pi_0}{1-\pi_0}$.

Aplicando a função logit fica

$$logit[P(Y = 1|X = 1)] = \log\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_1 + \beta_2, \tag{3.5}$$

e

$$logit[P(Y = 1|X = 0)] = \log\left(\frac{\pi_0}{1-\pi_0}\right) = \beta_1. \tag{3.6}$$

A razão de chance designada por *odds ratio* é estimada da seguinte forma

$$odds\ ratio = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}. \tag{3.7}$$

Substituindo a expressão (3.5) na (3.6) fica,

$$odds\ ratio = \frac{\exp(\beta_1 + \beta_2)}{\exp(\beta_1)} = \exp(\beta_2).$$

O valor da razão de chance representa o risco para a variável resposta Y tomar o valor 1, quando a variável explicativa $X = 1$, em relação a x = 0. O intervalo de $100\%(1-\alpha)$ de confiança para a estimativa e $\exp(\beta_2)$ é dado por

$$\left]\exp(\hat{\beta_2} - z_{1-\frac{\alpha}{2}}\hat{SE}(\hat{\beta_2})); \exp(\hat{\beta_2} + z_{1-\frac{\alpha}{2}}\hat{SE}(\hat{\beta_2}))\right[$$

onde $Z_{1-\frac{\alpha}{2}}$ é o quantil de probabilidade da distribuição Normal de valor médio zero e variância unitária.

Se o intervalo de confiança incluir o valor 1, não existe relação significativa entre as variáveis X (variável explicativa) e Y (variável resposta).

Capítulo 4

Modelos de Regressão para Dados de Contagem

Dados na forma de contagens aparecem com muita frequência em aplicações estatísticas. São exemplo disso estudos sobre o número de acidentes, o número de elementos numa fila de espera, etc. O modelo de Poisson, como se sabe, desempenha um papel fundamental na análise deste tipo de dados. Como também já se referiu na Secção 1.2, este é um modelo que pertence à família Exponencial que tem a particularidade de o valor médio ser igual à variância, Turkman (2000).

A Regressão de Poisson, isto é, o modelo Linear Generalizado com resposta de Poisson, é o método padrão usado para analisar dados de contagem. No entanto, muitas situações de dados da vida real violam as suposições nas quais o modelo de Poisson se baseia. Por exemplo, o modelo de Poisson assume que a média e a variância da variável resposta são idênticas. Isso significa que os eventos ocorrem dentro de um período de observação a uma taxa constante; um evento é igualmente provável em qualquer ponto do período. Um fenómeno que ocorre com frequência nas aplicações é o fenómeno de sobredipersão. Tal sobredispersão é indicada se a variância da variável resposta for superior ao valor da sua média. Designando por ϕ o parâmetro de sobredispersão, tal que $Var(Y) = \phi E[Y] = \phi\mu$, isto é, a variância é proporcional à média. Em termos de estimação, as estimativas pontuais são iguais às da situação em que não existe sobredispersão, mas a variância dos estimadores $\hat{\beta}$ é inflacionada pelo fator ϕ. Muitos autores sugerem a estimação de ϕ dada pelo quociente entre a estatística Qui-quadrado de Pearson e o número de graus de liberdade correspondente

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

sendo $\phi > 1$ indica sobredispersão. Quando este rácio é inferior a um, pode-se assumir a não existência de sobredispersão, prosseguindo-se com o processo de validação do modelo.

Uma ferramenta gráfica adicional para determinar se o modelo é adequado ou se existe sobredispersão nos dados é o *envelope plot*. Este gráfico é parte do gráfico normal quantil-quantil (ou seja, o *Q-Q plot*), para o qual os resíduos obtidos do modelo ajustado, contra os resíduos teóricos obtidos da distribuição Normal, são projetados. Se o gráfico for significativamente diferente de uma linha reta, há indícios claros de que os resíduos não seguem uma distribuição Normal, o que implica que o modelo ajustado não é adequado para os dados, (Turkman (2000)). O *envelope plot* simula intervalos de confiança empíricos para determinar se os resíduos diferem significativamente da linha reta. Se houver sobredispersão, a projeção dos resíduos cairá fora dos intervalos.

Para resolver o problema de sobredispersão utiliza-se, usualmente, o modelo de Regressão Binomial Negativa, Gauss Moutinho Cordeiro (2013).

4.1 Modelo de Regressão de Poisson

A distribuição de Poisson é uma distribuição discreta que assume um valor de probabilidade apenas para inteiros não negativos; essa característica da distribuição de Poisson torna-a uma escolha excelente para modelar resultados de contagem, que assumem apenas valores inteiros de 0 ou mais, Coxe (2009) .

Considere que Y_i é uma variável aleatória que segue uma distribuição de Poisson com parâmetro μ_i, denotado por $Y_i \sim P_0(\mu_i)$, então a função de probabilidade é dada por

$$f(y_i) = \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i), \tag{4.1}$$

com $y_i = 0, 1, 2, \ldots$, onde μ_i representa o número médio de ocorrências de um determinado acontecimento, $\mu_i > 0$. A média e a variância são dadas por,

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

O que implica que qualquer fator que afete a média afetará a variância.

No contexto de Modelos Lineares Generalizados, considerem-se $y_1, \ldots, y_n$ variáveis aleatórias independentes, com

$$y_i|x_i \sim P(\mu_i), i = 1, \ldots, n.$$

Seja $x_i = (x_{i1}, \ldots, x_{ip}), i = 1, \ldots n$ vetores de covariáveis correspondentes a cada observação na amostra.

Contudo, o modelo de Regressão de Poisson não pode ser utilizada para calcular o valor médio, uma vez que o preditor linear pode assumir qualquer valor real, enquanto que a média de Poisson, μ_i, só assume valores não negativos. Para solucionar este problema, utiliza-se a função de ligação logarítmica da seguinte forma

$$y_i|x_i \sim P(\mu_i)$$

$$\log(\mu(x_i)) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{4.2}$$

Os coeficientes de regressão $\beta_j, j = 1, \ldots, p$ representam a variação esperada no logaritmo do valor médio, por unidade de variação na variável explicativa x_j.

4.1.1 Estimação dos Coeficientes do Modelo

Utiliza-se o método de máxima verosimilhança para estimar os coeficientes do modelo. A função de verosimilhança para o modelo de Regressão de Poisson é dada por

$$L(\beta) = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i), \tag{4.3}$$

então o logaritmo da verosimilhança para o modelo de Poisson fica

$$\log(L(\beta)) = \sum_{i=1}^{n} ((y_i \log(\mu(x_i)) - \mu(x_i) - \log(y_i!)) \tag{4.4}$$

onde $\log(\mu(x_i))$ que está na equação (4.2) e $\mu(x_i)$ depende do vetor de covariáveis x_i e $\log(y_i!)$ é uma constante.

4.1.2 Qualidade de Ajustamento

Considerando que y_i são os valores observados e que $\hat{\mu}_i$ são os valores estimados pelo modelo de Regressão de Poisson, a função desvio para respostas de Poisson é

dada,

$$D = 2\sum_{i=1}^{n}(y_i \log(y_i) - y_i - \log(y_i!) - y_i \log(\hat{\mu}_i) + \hat{\mu}_i + \log(y_i!)),$$

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

O primeiro termo representa duas vezes a soma dos valores observados multiplicada pelo logaritmo do quociente entre os valores observados e estimados pelo modelo. Já o segundo termo diz respeito à soma das diferenças entre os valores observados e estimados, que usualmente é zero.

Assim, para modelos com termo constante, a função desvio reduz-se a

$$D = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)\right).$$

Pela equação (2.18) definida na secção (2.5.3), substitui-se a função de variância, $Var(\mu_i) = \mu_i$ e a estatística de Pearson generalizada é dada por

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

A qualidade do ajustamento de um modelo de Regressão de Poisson pode também ser avaliada por esta estatística.

4.2 Modelo de Regressão Binomial Negativa

A Regressão Binomial Negativa é mais eficazmente usada para modelar dados de contagem que violam a suposição de Poisson da igualdade de média e variância, por considerar um parâmetro adicional de dispersão α no cálculo da variância condicional. Este parâmetro é não negativo, portanto a variância condicional é, ou pode ser, maior que o valor médio.

Com efeito, o modelo é baseado na premissa de que os eventos entram num período de observação com uma distribuição Gama. Observando que as variâncias de Poisson e Gama são μ e μ^2 , respectivamente, a Binomial Negativa é considerado como uma distribuição de mistura Poisson-Gama com uma variância de $\mu + \mu^2$. Esta função pode ser reparametrizada como $\mu + \alpha\mu^2$ permitir uma relação linear no segundo termo. O parâmetro α pode ser considerado um fator de heterogeneidade e é inserido como uma constante conhecida, Hilbe (2011).

Seja Y uma variável aleatória com distribuição Binomial Negativa de parâmetros μ e α, designada Y $\sim$ BN(μ, α), se a função de probabilidades é dada por

$$f(y;\mu,\alpha) = \frac{\Gamma(\alpha+y)}{\Gamma(y+1)\Gamma(\alpha)}\left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}\left(\frac{\mu}{\mu+\alpha}\right)^{y}, \alpha > 0, \mu > 0 \tag{4.5}$$

onde $\Gamma(.)$ é a função Gama, α é denominado por parâmetro de heterogeneidade. O valor médio e a variância são dadas, respetivamente por

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

$$Var(Y) = \mu + \alpha\mu^2.$$

Se o parâmetro α for conhecido, verifica-se que a distribuição Binomial Negativa pertence à família Exponencial de distribuições e a teoria de Modelos Lineares Generalizados aplica-se.

Se o parâmetro α for desconhecido (situação mais frequente), deve ser estimado via máxima verossimilhança juntamente com os demais parâmetros do modelo.

Seja Y uma variável aleatória, representando o número de ocorrências de um determinado acontecimento com n observações que representa o número de ocorrências de um determinado acontecimento num certo período de tempo ou espaço, $X = (X_1, \ldots, X_p)$ um vetor de covariáveis e $x_i^T = (x_{i1}, \ldots, x_{ip})$ uma observação do indivíduo i e assume-se que

$$Y|X = x_i \sim BN(\mu(x_i), \alpha)$$

onde $\mu_i = \mu(x_i)$ representa o número médio de ocorrências de um dado acontecimento dada a observação x_i.

O modelo de Regressão Binomial Negativa é então expresso por Hilbe (2001), Santos (2013), por

$$Y|X = x_i \sim BN(\mu(x_i), \alpha),$$

$$\log(\mu_i(x_i)) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

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

Naturalmente, tem-se que

$$\mu_i = E(Y|X = x_i)$$

e

$$Var(Y|X = x_i) = \mu_i + \alpha\mu_i^2.$$

4.2.1 Estimação dos Coeficientes do Modelo

Para estimar os coeficientes de regressão aplica-se o método de máxima verossimilhança. A função do logaritmo da verosimilhança para n observações da distribuição Binomial Negativa é dada por

$$\log(L(\beta)) = \sum_{i=1}^{n}\left(y_i \log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \left(\frac{1}{\alpha}\right)\log(1+\alpha\mu_i) + \log\left(\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\right)\right), \tag{4.6}$$

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

O desvio é duas vezes a diferença entre o log de máxima verosimilhança e o log de verosimilhança do modelo ajustado. No modelo de Regressão Binomial Negativa as estimativas de máxima verosimilhança para β e α são obtidas através do método de mínimos quadrados apresentado na Secção 2.5.3.

4.2.2 Qualidade de Ajustamento

Para avaliar a qualidade de ajustamento de um modelo de Regressão Binomial Negativa com p parâmetros e as observações $(y_1, \ldots, y_n)$, utilizam-se as mesmas estatísticas que para o modelo de Regressão de Poisson.

A expressão de cálculo da função desvio para o modelo de Regressão Binomial Negativa é dada por

$$D = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(\frac{1}{\alpha} + y_i\right)\log\left(\frac{1+\alpha y_i}{1+\alpha\hat{\mu}_i}\right)\right). \tag{4.7}$$

Substituindo a função da variância $Var(\hat{\mu}_i) = \mu + \alpha\mu^2$, a estatística de Pearson generalizada é dada por

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

Capítulo 5

Aplicação a Dados Reais

Neste capítulo iremos aplicar Modelos Lineares Generalizados, em particular a Regressão Logística, a Regressão de Poisson e a Regressão Binomial Negativa, na análise de dados reais.

Uma das bases de dados utilizada refere-se ao número de doentes contaminados com o COVID-19 e que já recuperaram ou não, nas Filipinas. A amostra foi obtida no mês de fevereiro de 2020 e contém 143 observações. A variável resposta ***Status*** (Estado do Doente) é uma variável binária, que pode tomar dois valores: 0 se o doente recuperou, e 1 se o doente não recuperou.

A base de dados é ainda constituída por quatro variáveis: **Idade**, **Sexo**, **Nacionalidade** e **Transmissão** que são variáveis explicativas do modelo de regressão.
Na Tabela 5.1 está apresentada a descrição das variáveis.

Tabela 5.1: Variáveis em Estudo

Variável	Descrição
Status	A variável indica se o doente recuperou ou não;
Idade	A variável representa a idade dos doentes;
Sexo	A variável representa o sexo dos doentes;
Nacionalidade	A variável representa a nacionalidade dos doentes;
Transmissão	A variável representa o local de transmissão da doença.

A análise exploratória tem como objetivo obter informação proveniente dos dados a tratar. Para uma melhor compreensão das variáveis qualitativas iremos utilizar o gráfico de barras. Na descrição da variável quantitativa contínua irá ser apresentada o

histograma e caixa com bigode

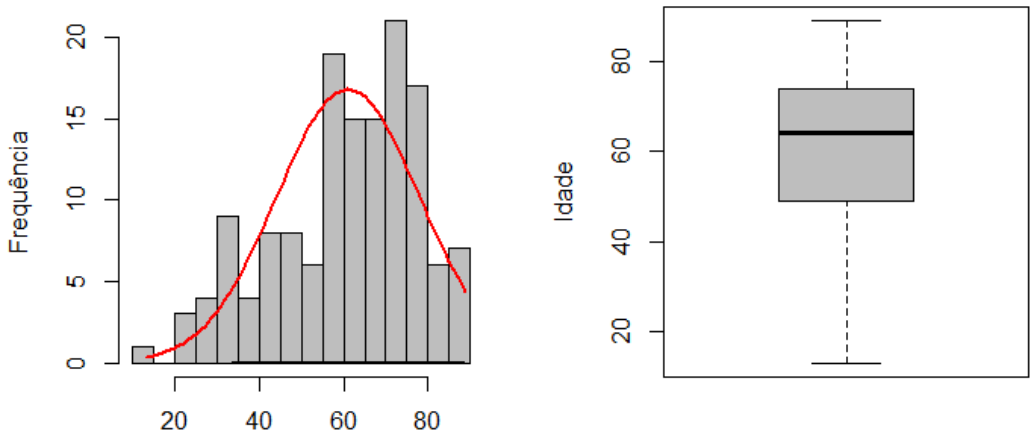

Figura 5.1: Histograma e diagrama em caixa de Bigodes para a variável **Idade**

A Figura 5.1 sugere que a variável idade tem distribuição enviesada à esquerda. Aplicou-se o teste de *Shapiro Wilk* para verificar se a variável **Idade** segue uma distribuição Normal e obteve-se uma estatística de teste de 0,9750 e o valor de prova de 0,0002, rejeitando-se a hipótese nula, ao nível de significância de 5%. Ou seja, há evidência estatística que a variável **Idade** não segue uma distribuição Normal. A caixa com bigode sugere que 75% dos doentes infetados com o COVID-19 estão entre as idade 13 e 75 anos.

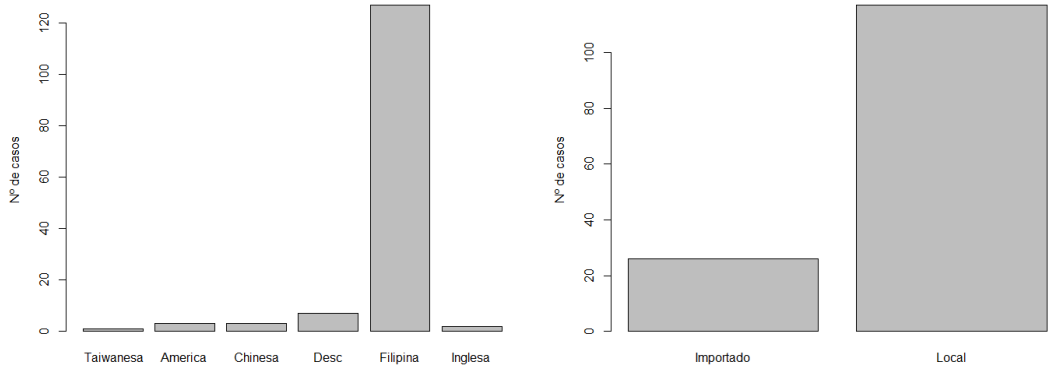

Figura 5.2: Gráfico de barras para a variável **Nacionalidade** e **Transmissão**

Na Figura 5.2 o gráfico à esquerda indica que o maior número de doentes é de nacionalidade Filipina. O gráfico à direita podemos observar que o maior número de doentes (117) teve a doença por transmissão local.

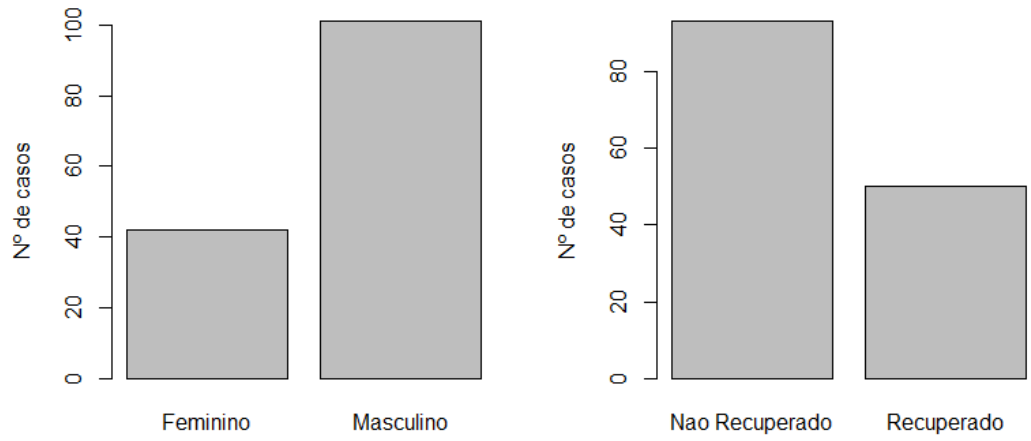

Figura 5.3: Gráfico de barras para a variável **Sexo** e ***Status***

Na Figura 5.3 apresenta a distribuição dos doentes recuperados ou não da COVID-19 nas Filipinas a distribuição dos doentes segundo o género. Os doentes que foram mais infectados são do sexo masculino e há um maior número de doentes não recuperados.

Associação entre a variável resposta e a Idade

Para verificar a existência de diferenças significativas entre as idade dos dois grupos de doentes (recuperados e não recuperados), aplica-se o teste de *Mann- Whitney*, uma vez que o pressuposto de Normalidade não foi verificado.

O objectivo do teste é o de avaliar a igualdade das medianas das idades dos dois grupos de doentes. As hipóteses a testar são:

H0: A mediana das idades para os doentes recuperados é igual a mediana das idades para os doentes não recuperados.

H1: A mediana das idades para os doentes recuperados é diferente a mediana das

idades para os doentes não recuperados.

Na Tabela 5.2 apresenta-se a distribuição da idade dos doentes recuperados e não recuperados.

Tabela 5.2: Distribuição da idade entre doentes Recuperado e não Recuperado

Idade	Recuperado	Não Recuperado
Média	48	68
Mediana	46	69
Desvio Padrão	17	12
Máximo	88	89
Mínimo	13	34

Em relação aos doentes recuperados, a média da idade é 48, com desvio padrão 17, a mediana das idade é 46, a idade mínima é 13, e a idade máxima é 88.

Em relação aos doentes não recuperados, a média da idade é 68, com desvio padrão 12, a mediana das idade é 69, a idade mínima é 34, e a idade máxima é 89.

Aplicando o teste não paramétrico *Mann- Whitney*, obtêm-se a estatística de teste 10296 e o valor de prova é 0,0001. Conclui-se que há evidencias estatísticas para afirmar que existem diferenças significativas entre as idades dos dois grupo de doentes.

Associação entre a variável resposta e as variáveis explicativas categóricas

Uma tabela de contingência é uma tabela de tabulação cruzada de duas ou mais variáveis aleatórias, normalmente qualitativas. O objetivo principal da análise da tabela de contingência é averiguar se existe ou não alguma relação entre as variáveis aleatórias de qualquer tipo que se representam agrupados numa tabela de contingência.

Para averiguar a existência dessas relações pode realizar-se o teste de independência de Qui-quadrado.

Este teste compara as frequências dos valores observados com as frequências dos valores esperados das diferentes categorias de uma variável aleatória e a hipótese nula é rejeitada quando o valor da estatística de teste for maior que o valor crítico da

distribuição Qui-quadrado.

Este teste não deve ser utilizado se mais do que 20% das frequências esperadas, sob a hipótese nula, forem inferiores a 5 ou se algumas delas for igual 0.

As hipóteses a testar são:
H_0: Não há associação entre as duas variáveis

H_1: Há associação entre as duas variáveis

Para testar a hipótese nula de que não existe associação entre as duas variáveis, usamos a seguinte estatística de teste

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} = \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

onde r é número de linhas, c é o número de colunas da tabela de contingência, n_{ij} representa a frequência observada na célula (i, j), e e_{ij} representa a frequência esperada na célula (i, j) que é estimada por

$$e_{ij} = \frac{t_i \times t_j}{t_a}$$

onde t_i total de linha i, t_j total de coluna j e t_a total da amostra.

Nas Tabelas 5.3, 5.4 e 5.5 estão apresentados as frequências observadas dos pacientes da amostra, recuperados e não recuperados da COVID-19 em cada uma das categorias das covariáveis utilizadas neste estudo. Quanto maior for a diferença entre as frequências observadas e as frequência esperadas, maior será a associação entre as variáveis.

Tabela 5.3: Frequência da variável **Status** segundo a **transmissão**

Status	Importado	Local	Total
Recuperado	19	31	50
Não Recuperado	7	86	93
Total	26	117	143

Tabela 5.4: Frequência da variável **Status** segundo o **Sexo**

Status	Feminino	Masculino	Total
Recuperado	15	35	50
Não Recuperado	27	66	93
Total	42	101	143

Tabela 5.5: Frequência da variável ***Status*** segundo a **Nacionalidade**

Status	Americana	Inglesa	Chinesa	Tawainesa	Filipina	Desconhecida	Total
Recuperado	1	0	2	1	46	0	50
Não Recuperado	2	2	1	0	81	7	93
Total	3	2	3	1	127	7	143

Na Tabela 5.6 apresentam-se os valores da estatística de teste e os respetivos valores de prova do teste de independência de Qui-quadrado entre as variáveis explicativas e a variável resposta.

Na base de dados em estudo, há evidência estatística que existe uma associação significativa entre a variável resposta ***(Status)*** e a variável **Transmissão** e que não existe associação significativa com as variáveis **Nacionalidade** e **Sexo**, ao nível de significância de 5%.

Tabela 5.6: Teste de independência de Qui-quadrado entre as variáveis explicativas e a variável resposta

Variáveis	Estatística de teste (X^2)	valor-p	Graus de liberdade
Transmissão	18,302	<0,001	1
Nacionalidade	8,116	0,152	5
Sexo	0,098	0,754	1

5.1 Estimação do modelo

Este modelo foi desenvolvido com objectivo de estimar a probabilidade de um determinado doente não recuperar da COVID-19 recorrendo-se a Regressão Logística.

Inicialmente ajustou-se um modelo de Regressão Logística simples a cada uma das varáveis explicativas, com o objetivo de estudar a importância de cada variável explicativa tem para a variável resposta, a Tabela 5.7 esta a Regressão Logística simples.

Tabela 5.7: Modelo de Regressão Logística simples

Variáveis explicativas	Estimativas dos coeficientes	*Odds ratio*s (OR)	Intervalo de confiança (95% OR)	Desvio padrão	Teste Wald	Valor-p
Constante	4,626	102	(2,94 ; 6,59)	0,926	4,997	0,001
Idade	-0,088	0,92	(-0,06; -0,12)	0,015	-5,743	$< 0,001$
Constante	-0,485	0,61	(-1,13 ;0,13)	0,317	-1,528	0,121
Sexo						
Masculino	-0,193	0,82	(-0,94; 0,56)	0,382	-0,506	0,613
Constante	17,570	0,01	(0,04 ; 21,96)	3956	0,004	0,996
Nacionalidade						
América	-18,260	1,27	(0,00 ; 0,00)	3956	-0,005	0,996
Inglesa	-35,130	0,00	(0,00 ; 1,24)	4845	-0,007	0,994
Chinesa	-16,870	0,01	(-0,01 ; 9,95)	3956	-0,004	0,997
Filipina	-18,130	0,06	(-0,04 ; 1,81)	3956	-0,005	0,996
Desconhecido	-35,130	0,00	(0,00 ; 2,66)	4229	-0,008	0,993
Constante	0,998	2,71	(0,17 ; 1,94)	0,442	2,258	0,024
Transmissão						
Local	-2,019	0,13	(-3,04 ;-1,09)	0,489	-4,126	0,001

De seguida procedeu-se à construção dos modelos de regressão logística múltipla. Na seleção das variáveis usa-se o método de seleção *backward, stepwise e forward* .

Modelo inicial com todas as variáveis explicativas (Modelo Completo)

Status ~ Bernoulli(p), onde p é a probabilidade de um doente não recuperar da COVID-19.

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times$ **Idade** $+ \beta_2 \times$ **Sexo** $+ \beta_3 \times$ **Nacionalidade**$+ \beta_4 \times$ **Nacionalidade**$+ \beta_5 \times$ **Nacionalidade**$+ \beta_6 \times$ **Nacionalidade**$+ \beta_7 \times$ **Nacionalidade**$+ \beta_8 \times$ **Transmissão**.

Tabela 5.8: Modelo de Regressão Logística inicial (Modelo Completo)

Variáveis explicativas	Estimativas dos coeficientes	*Odds ratio*s (OR)	Intervalo de confiança (95% OR)	Desvio padrão	Teste Wald	Valor-p
Constante	7,074	0,01	(0,04 ; 21,96)	2,040	3,466	0,005
Idade	-0,088	0,09	(0,09 ; 0.88)	0,017	-5,011	$< 0{,}001$
Sexo						
Masculino	-0,587	0,04	(0,15 ; 1,20)	0,524	-1,118	0,264
Nacionalidade						
América	16,358	1,27	(0,00 ; 0,00)	3956	0,004	0,996
Inglesa	-17,377	0,00	(0,00 ; 1,24)	2413	-0,007	0,994
Chinesa	-1,875	0,01	(0,01 ; 9,95)	1,968	-0,952	0,341
Filipina	-0,473	0,06	(0,04 ; 1,81)	1,477	-0,321	0,748
Desconhecido	-16,477	0,00	(0,00 ; 2,66)	1329	-0,012	0,991
Transmissão						
Local	-1,823	0,01	(0,03 ; 0,05)	0,668	-2,727	0,006

Na Tabela 5.8 estão apresentados os valores estimados dos coeficientes, os desvios padrão, a estatística de teste *Wald* e os valores de prova. Pode-se verificar que as variáveis explicativas **Idade** e **Transmissão** são estatisticamente significativas, evidenciando que existe associação com a variável resposta.

Aplicando este método de seleção o modelo final

Modelo 1

Status ~ Bernoulli(P)

$\log\left(\frac{p}{1-p}\right) = \beta_0 +$ **Idade** $+ \beta_1 \times$ **Transmissão**

Seleção *Forward* usando a estatística AIC

Este método inicia com o modelo sem variáveis explicativas adicionando cada variável para averiguar como influência a variável resposta, baseando-se na estatística AIC. O modelo selecionado usando este método é

Modelo 2

Status ~ Bernoulli(P)

$\log\left(\frac{p}{1-p}\right)$ **=**β_0 **+** $\beta_1\times$ **Idade +** $\beta_2\times$ **Transmissão**

Seleção *Both* usando a estatística AIC

Este método é uma combinação dos dois métodos antecedentes e consiste na remoção e inclusão das variáveis baseando-se na estatística AIC. O modelo selecionado usando este critério é

Modelo 3

Status ~ Bernoulli(p)
$\log\left(\frac{p}{1-p}\right)$ **=**β_0 **+** $\beta_1\times$ **Idade +** $\beta_2\times$ **Transmissão**

Podemos verificar que o modelo final é o mesmo para todos os métodos de seleção das covariáveis. Para confirmar que o modelo selecionado se ajusta melhor, compara-se o valor do AIC e a função desvio do modelo final com o modelo nulo. A Tabela 5.7 apresenta os resultado da comparação.

Da Tabela 5.9 observa-se que o valor do AIC e do desvio do modelo final é menor, do que o modelo nulo, indicando um melhor ajustamento.

Tabela 5.9: Comparação entre os Modelos

	AIC	Desvio
Modelo Nulo	187,11	185,11
Modelo Final	132,05	126,05

Análise de resíduos

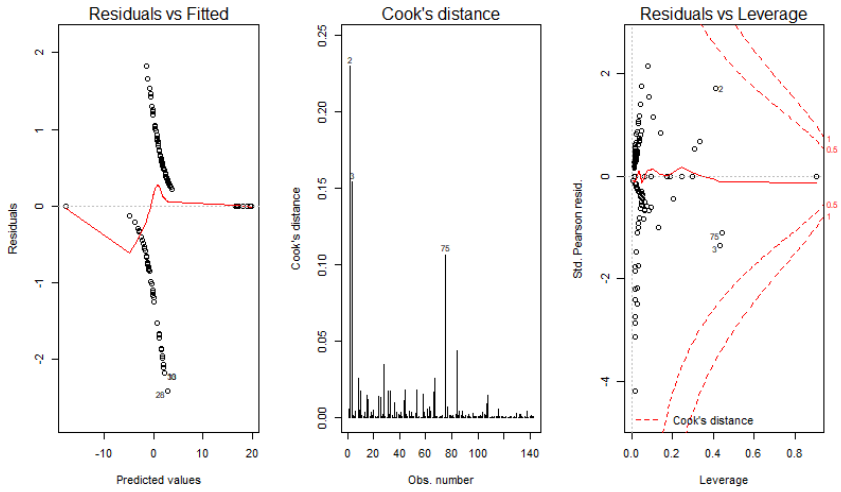

Figura 5.4: Gráficos da análise de resíduos do modelo de regressão Logística

Na Figura 5.4 apresenta-se o gráfico de resíduos do desvio onde mostra algumas observações estão fora do intervalo considerado [-2,2] portanto, há observações *outlier*.

Interpretação do Modelo Final

Como se pode verificar na Tabela 5.10 onde estão apresentados os valores da razão de chances *(odds ratio)* e os respetivos intervalos de confiança de 95%, também podemos observar as estimativas dos coeficientes que representam o efeito que cada covariável pode causar na variável resposta, o desvio padrão, o teste de *Wald* e o valor de prova. Será feita a interpretação do modelo baseando-se em razão da chances de cada variável explicativa incluída no modelo.

Tabela 5.10: Razão de chances e intervalo de confiança

Variáveis explicativas	Estimativas dos coeficientes	*Odds ratio*s (OR)	Intervalo de confiança (95% OR)	Desvio padrão	Teste Wald	Valor-p
Constante	-5,738	0,01	(-8,024 ; -3,799)	1,051	-5,36	<0,001
Idade	0,086	1,09	(0,056 ; 0,119)	0,016	5,34	<0,001
Transmissão Local	1,632	5,12	(0,533 ; 2,818)	0,576	2,83	0,005

A chance de não recuperar aumenta em cerca de 9% com aumento de uma unidade na idade.

Quanto à variável transmissão, verifica-se que a chance de um paciente não recuperar é 5,12 vezes maior em transmissão local em relação a transmissão não local.

Avaliação preditiva do modelo Final

Tabela 5.11: Matriz de confusão do modelo Final

Valor estimado	Não Recuperado (1)	Recuperado (0)	Total
Recuperado (0)	17	33	50
Não Recuperado (1)	82	11	93
Total	99	44	143

Sensibilidade é a capacidade do modelo para identificar corretamente os pacientes que não recuperaram da doença.
Sensibilidade = $\frac{82}{82+17} = 0,82 \times 100 = 82\%$

Especificidade é a capacidade do modelo para identificar corretamente os pacientes que recuperaram.
Especificidade = $\frac{33}{33+11} = 0,75 \times 100 = 75\%$

Exatidão é a proporção de indivíduos corretamente classificados pelo modelo.
Exatidão = $\frac{82+33}{143} = 0,804 \times 100 = 80,4\%$

A Tabela 5.11 mostra que dos 82% dos pacientes que não recuperaram da doença foram identificados corretamente pelo modelo e 75% dos pacientes que recuperaram da doença foram identificados corretamente pelo modelo. Verifica-se ainda que a proporção corretamente classificados foi de 80,4%.

Curva ROC

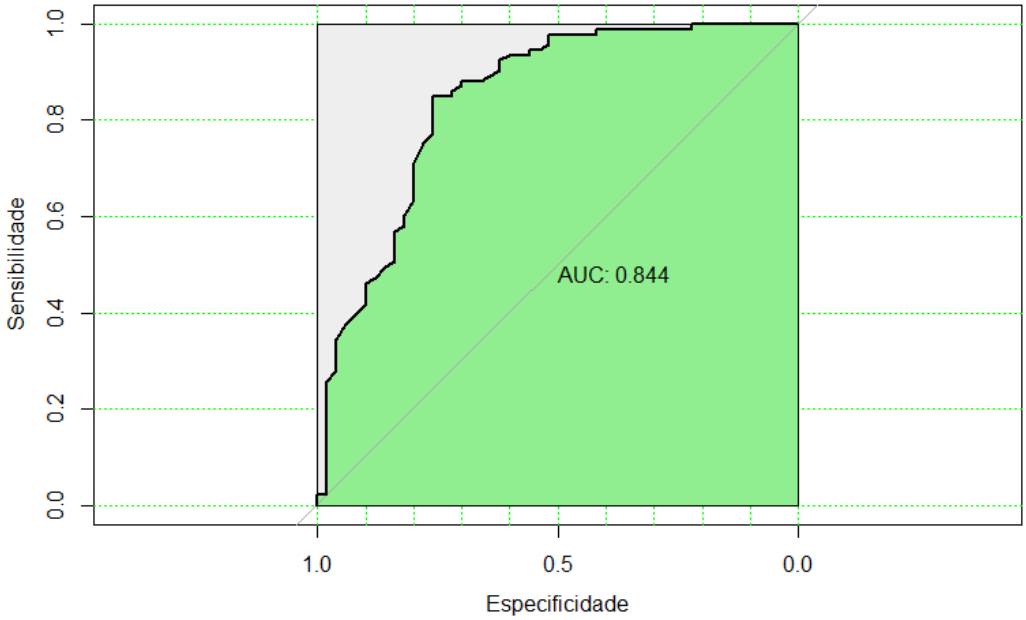

Figura 5.5: Gráfico da Curva ROC associada ao modelo de regressão Logística

Na Figura 5.5 como se pode observar a linha diagonal representada no gráfico indica uma classificação aleatória, quanto mais distante estiver a curva da diagonal principal melhor é o desempenho. Isto significa que a capacidade de predição do modelo pode ser considerada bom ou aceitável de acordo com os critério de avaliação proposto por *Hosmer e Lemeshow*.

5.2 Modelo de Regressão de Poisson

Os dados utilizados referem-se a casos de COVID-19 de 130 países a nível mundial. A amostra foi obtida a 3 de agosto do ano 2020 e contém 130 observações. Estes dados foram preenchidos pela Organização Mundial da Saúde (OMS) e os mesmos estão disponível no site https://www.worldometers.info/coronavirus/
e https://www.gapminder.org/data/. Os dados foram copiados e colados num ficheiro de texto com a denominação Covid2021.txt.

Na Tabela 5.12 são apresentadas as variáveis de estudo.

Tabela 5.12: Variáveis em Estudo

Variável	Descrição
Recuperado	Número total de pessoas recuperadas da COVID-19 no país;
Casos	Número total de casos contaminados por COVID-19 no país;
Mortes	Número total de pessoas que morreram por COVID-19 no país;
Testes	Número total de Testes efetuados no país;
Ativos	Número de casos Ativo com COVID-19 até ao momento no país;
Crítico	Total de pessoa em estado Crítico com COVID-19 até ao momento no país;
População	Total da população no país;
Expectativa	Esperança de vida dos pacientes com COVID-19 no país;
IDS	Índice de desigualdade social no país;
GTSP	Gasto total na saúde por pessoa no país.

A variável **Recuperado** é a variável dependente. Na Tabela 5.13 estão apresentados o valor mínimo, a mediana, a média, o valor máximo, e o desvio padrão de cada variável, assim como o coeficiente de variação das variáveis em estudo. O número médio das pessoas que recuperaram nos 130 países do estudo foi de 46845. Observa-se que o número máximo dos países ja recuperaram 1165442 pessoas mas em outros países apenas recuperaram 18 pessoas.

Tabela 5.13: Tabela das medidas de tendência central e dispersão das variáveis

Variáveis	Média	Mediana	Desvio padrão	Coeficiente de Variação	Mínimo	Máximo
Recuperado	46845	6354	130687,62	278,9813	18	1165442
Casos	66649	9804	188647,5	283,0446	27	1780268
Mortes	2447	169	6995	286	1,0	47472,0
Testes	1774706	216588	8470158	477,271	4,0	90410000
Ativos	17334	2264	57145,45	329,6711	0,0	577136
População	4,643e+07	1,003e+07	175956291	378,9742	9,800e+04	1,439e+09
Expectativa	74,13	75,15	6,496851	8,764587	56,60	85,30
IDS	63,88	38,30	284,3707	445,1533	25,00	328,00
GTSP	1072,3	469,5	1461,203	136,2691	16,0	8360,0

O número total de pessoas que morreram por COVID-19 difere nos 130 países em estudo; num dos país ocorreu apenas uma morte enquanto que em outro país ocorreram 47472 mortes.

O índice de desigualdade social no país difere nos 130 países em estudo; num dos país o número de desigualdade foi muito baixa com apenas 25 casos enquanto que em outro país foi de 328.

O gasto total na saúde por pessoa no país com COVID-19 difere nos 130 países em estudo; num dos país gastou-se apenas 16 Dólares enquanto que em outro país gastou-se 8360 Dólares por pessoa.

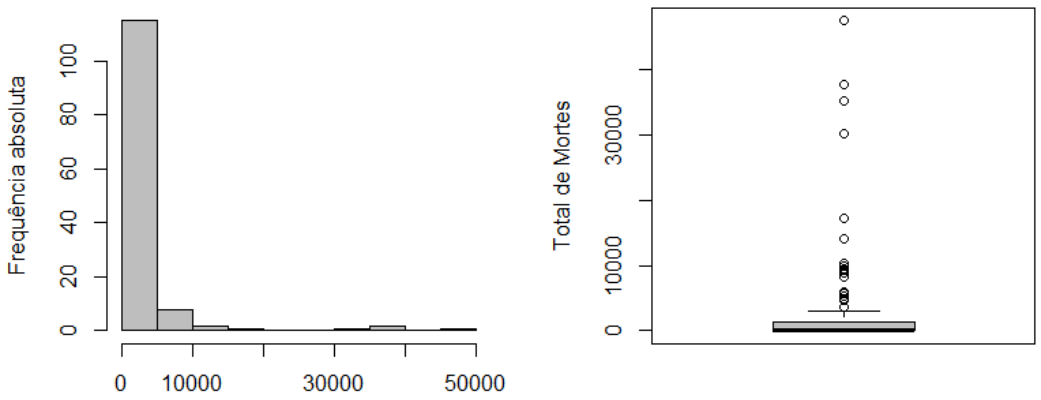

Figura 5.6: Histograma e caixa com Bigodes para a variável **Mortes**

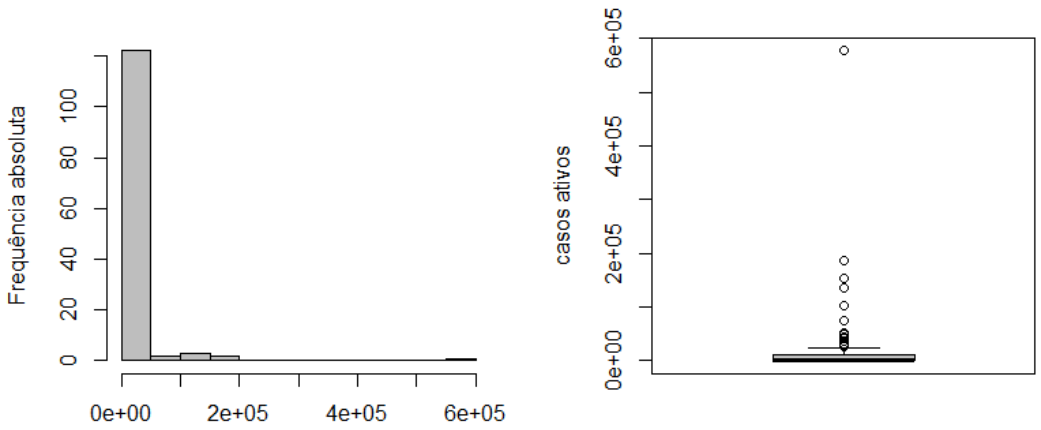

Figura 5.7: Histograma e caixa com Bigodes para a variável **Ativos**

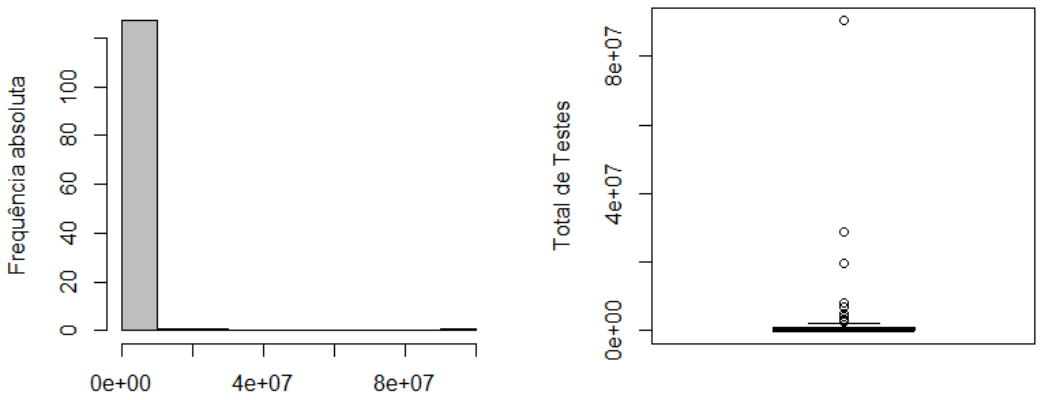

Figura 5.8: Histograma e caixa com Bigodes para a variável **Testes**

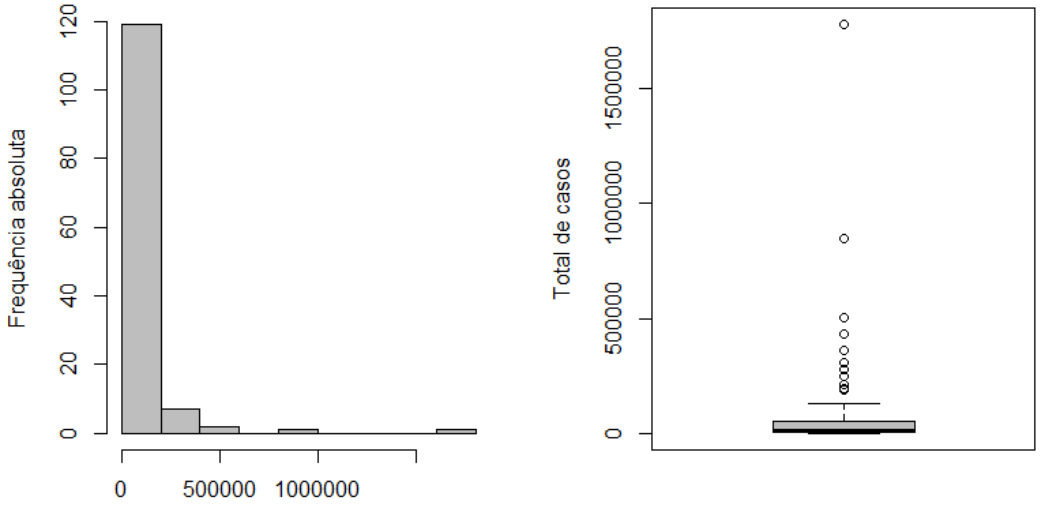

Figura 5.9: Histograma e caixa com Bigodes para a variável **Casos**

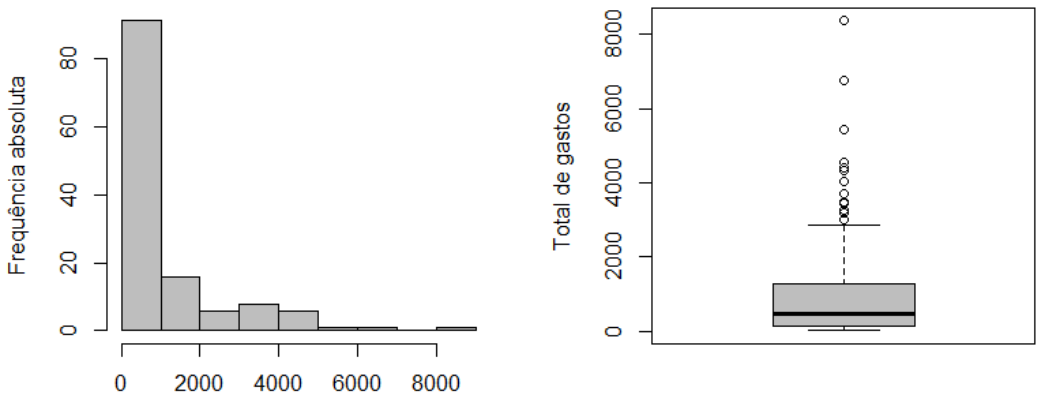

Figura 5.10: Histograma e caixa com Bigodes para a variável **GTSP** (Total de Gastos na saúde por pessoa)

Da Figura 5.6 até a Figura 5.10, apresentam-se os Histograma e as caixas de bigode das variáveis explicativas dos modelos.

Os gráficos sugerem que estas variáveis têm distribuições enviesadas à direita, ou enviesamento positivo, uma vez que se apresentam concentradas no lado esquerdo com uma larga cauda para a direita. Além disso pode-se observar a presença de observações outliers.

5.2.1 Associação entre as variavéis

Nesta secção usaremos o coeficiente de correlação de *Spearman* para estudar a existência da correlação entre a variável dependente com cada uma das variáveis independentes.

Na Tabela 5.14 mostra a correlação de Spearman entre a variáveis. Nestes resultados, a correlação de Spearman entre a variável resposta Recuperado e as variáveis explicativas Mortes, Casos, Testes, População, Ativos, Expectativa e GTSP para o nível de significância de 5% indica que existe uma associação positiva entre as variáveis. Por exemplo, quando a expectativa de vida do pais aumenta, o número de recuperados aumenta.

Tabela 5.14: Tabela de correlação de *Spearman* entre as variáveis

Variáveis								
Casos	**1**							
Mortes	**0,92**	**1**						
Recuperado	**0,97**	**0,91**	**1**					
Ativos	0,89	0,83	0,82	**1**				
Testes	0,74	0,67	0,78	0,58	**1**			
População	0,57	0,58	0,56	0,53	0,57	**1**		
Expectativa	0,27	0,26	0,32	0,10	0,49	-0,11	**1**	
IDS	-0,15	-0,24	-0,21	-0,01	-0,31	-0,04	-0,37	**1**
GTSP	0,16	0,10	0,19	0,13	0,15	0,06	0,09	-0,08 **1**

Dado que se tratam de dados de contagem recorre-se o Modelo de Regressão de Poisson no ajustamento do modelo.

Modelo Inicial

Recuperados $\sim$ P(μ)

$\log(\frac{\mu}{casos}) = \beta_0 + \beta_1 \times$ **Ativos** $+ \beta_2 \times$ **Mortes** $+ \beta_3 \times$ **Testes** $+ \beta_4 \times$ **IDS** $+ \beta_5 \times$ **população** $+ \beta_6 \times$ **Expectativa**$+ \beta_7 \times$ **GTSP**

Tabela 5.15: Modelo de Regressão de Poisson

Variáveis explicativas	Estimativas dos coeficientes	Desvio padrão	Teste Wald	Valor-p
Constante	-0,001	0,009	-46,79	0,001
Ativos	-0,006	0,005	-108,7	0,001
Mortes	0,007	0,001	10,34	0,001
Testes	0,001	0,001	99,62	0,001
IDS	-0,002	0,001	-6,971	0,001
Expectativa	0,003	0,001	9,799	0,001
GTSP	0,001	0,003	62,33	0,001

Na Tabela 5.15 estão apresentados as estimativas dos coeficientes do modelo, os valores da estatística de Wald e os respetivos valores de prova indicam que todos os coeficientes associados a cada variável explicativa são estatisticamente significativos.

O parâmetro de dispersão $\hat{\phi} = 239$ o que evidencia dispersão dos dados.

Analisando da Figura 5.11 referente ao gráfico Normal de Probabilidade do Modelo de Regressão de Poisson ajustado verifica-se que o modelo não traduz um bom ajustamento aos dados.

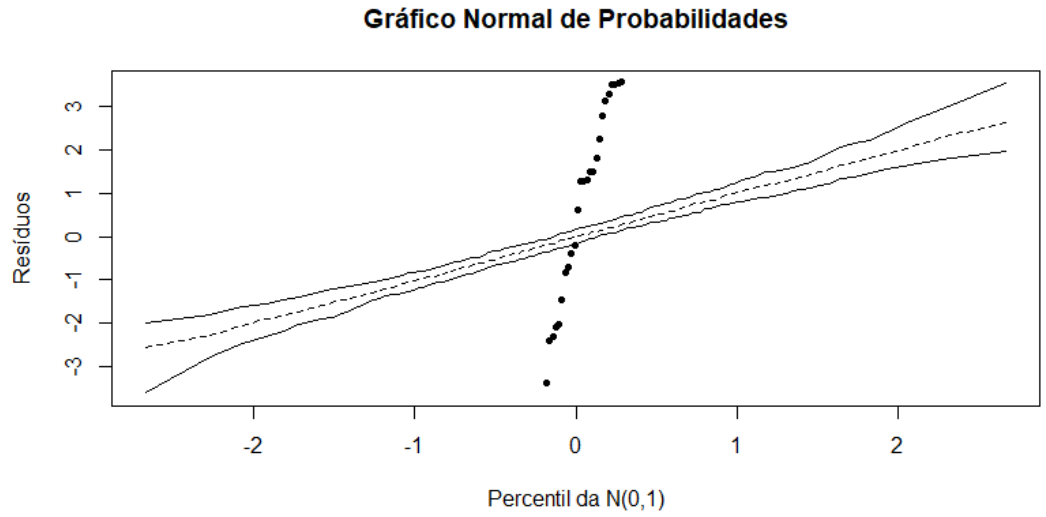

Figura 5.11: Gráfico normal de probabilidade referente ao modelo de poisson

Mediante este problema vai-se ajustar aos dados um Modelo de Regressão Binomial Negativo.

Regressão Binomial Negativa

O ajustamento aos dados foi realizado de forma análoga ao modelo Regressão de Poisson. Começou-se por se ajustar o modelo inicial, com todas as variáveis. Na Tabela 5.13 apresentam-se as estimativas dos coeficientes, desvio padrão, teste Wald e os respetivos valor de prova.

Modelo inicial

$\log(\frac{\mu}{casos}) = \beta_0 + \beta_1 \times$ **Ativos** $+ \beta_2 \times$ **Mortes** $+ \beta_3 \times$ **Testes** $+ \beta_4 \times$ **IDS** $+ \beta_5 \times$ **População** $+ \beta_6 \times$ **Expectativa** $+ \beta_7 \times$ **GTSP**

Tabela 5.16: Modelo de Regressão Binomial Negativa modelo inicial

Variáveis explicativas	Estimativas dos coeficientes	Desvio padrão	Teste Wald	Valor-p
Constante	-1,706	0,001	-4,148	0,003
Ativos	-0,001	0,006	-0,943	0,345
Mortes	-0,004	0,006	-0,067	0,946
Testes	0,001	0,009	0,054	0,957
IDS	0,005	0,004	-6,971	0,962
População	0,003	0,001	0,625	0,532
Expectativa	0,002	0,003	3,078	0,002
GTSP	0,005	0,002	1,281	0,201

Para o nível de significância de 5% , os valores da estatística de teste e os respetivos valores de prova, levam a concluir que apenas a variável **Expectativa** é significativa.

Modelo de seleção

Na seleção das variáveis iniciou-se com o método *Backward* usando a estatística AIC. Neste método começa-se com modelo completo e vai-se retirando cada variável baseando-se na estatística AIC, isto é, constrói-se um novo modelo retirando cada variável explicativa e escolhe-se o modelo com o menor valor de AIC. O processo termina quando ao retirar-se uma variável, o valor do AIC aumenta.

Seleção *Backward* usando a estatística AIC
Modelo 1:
$\log(\frac{\mu}{casos}) = \beta_0 + \beta_1 \times$ **Expectativa**.

Seleção *Forward* usando a estatística AIC
Modelo 2:
$\log(\frac{\mu}{casos}) = \beta_0 + \beta_1 \times$ **Expectativa** .

Seleção *Both* usando a estatística AIC
Modelo 3:
$\log(\frac{\mu}{casos}) = \beta_0 + \beta_1 \times$ **Expectativa**.

Ao comparar os resultados dos métodos de seleção verifica-se que o modelo final é

o mesmo.

Tabela 5.17: Comparação entre os Modelos

	AIC	Desvio
Modelo inicial	289033	287650
Modelo Final	2400.4	133.72

Na Tabela 5.17 compara-se o modelo inicial com o modelo final e verifica-se que o modelo final teve um menor valor de AIC e um menor valor da função Desvio em relação ao modelo inicial.

As estimativas dos coeficientes do modelo, do desvio padrão, da estatística de teste de Wald e respetivos valores provas são apresentados na tabela 5.18.

Tabela 5.18: Modelo Final de Regressão Binomial Negativa

Variáveis explicativas	Estimativas dos coeficientes	Desvio padrão	Teste Wald	Valor-p
Constante	-1,694	0,405	-4,178	0,001
Expectativa	0,017	0,005	3,189	0,002

Na Figura 5.12 está representado o gráfico Normal de probabilidade para o modelo final de regressão Binomial Negativa ajustado. Analisando a Figura é evidente que o modelo Binomial Negativa é mais adequado para explicar a variabilidade dos dados do que o modelo de Regressão de Poisson. No entanto, a qualidade de ajustamento aos dados deve ser melhorado.

Várias tentativas foram realizadas para obter um melhor ajustamento aos dados mas não se obteve resultados positivos.

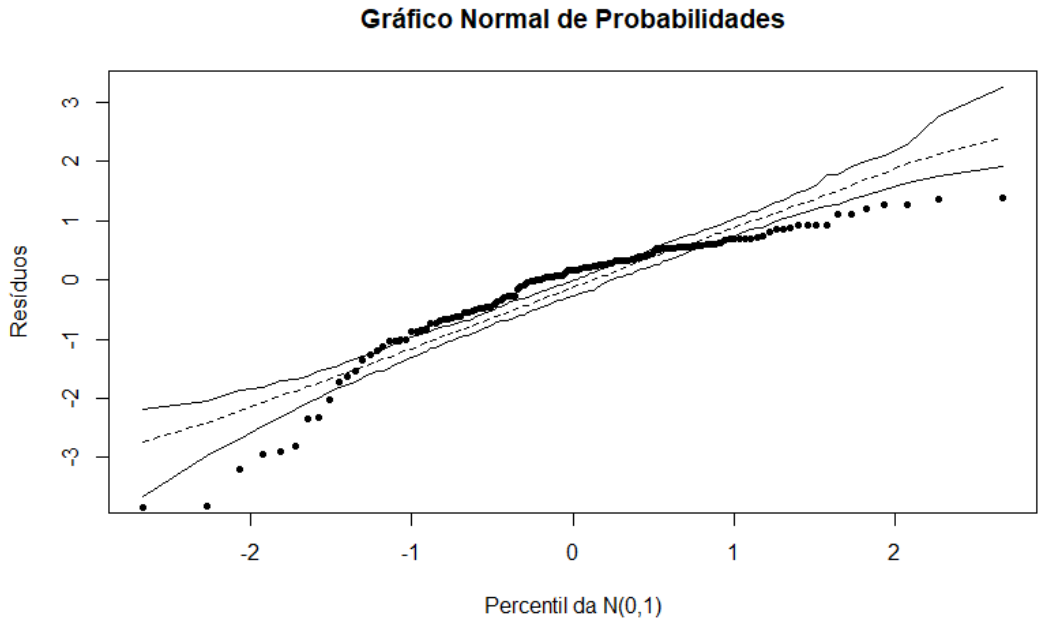

Figura 5.12: Gráfico Normal de probabilidade do modelo Binomial Negativa

Na Figura 5.13 está apresentada os gráficos da análise de resíduos. Estes gráficos permitem verificar a qualidade do modelo ajustado com a Regressão Binomial Negativa. Da observação destes gráficos, identificam-se observações outliers. No entanto, o modelo não sofre alteração significativa quando se eliminam essas observações.

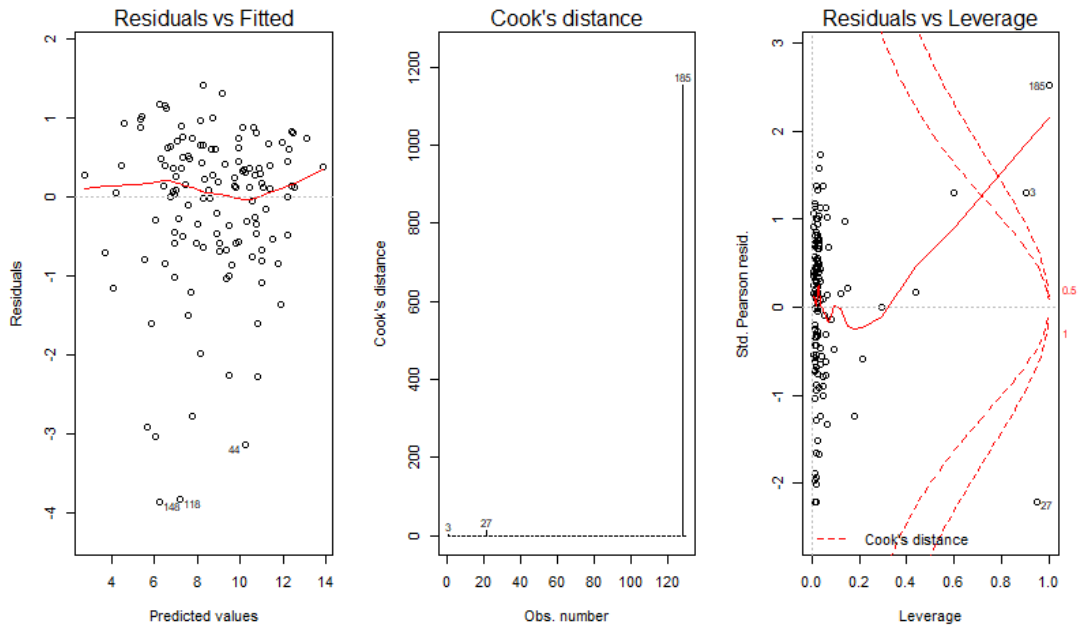

Figura 5.13: Gráfico da análise de resíduos do modelo de regressão Binomial Negativa

Interpretação do modelo ajustado da Regressão Binomial Negativa

Para o aumento de uma unidade na esperança de vida do país, a taxa de doentes recuperados da COVID-19 aumenta 1,017% como é visto na Tabela 5.18.

Para comparar a qualidade de ajustamento dos dois modelo, o modelo de Poisson e o modelo de Binomial Negativa, efetua-se o teste *Vuong*.

Tabela 5.19: Teste Vuong entre os modelos de Regressão

Teste Vuong	Binomial Negativa e Poisson
Estatística	286626,128
Valor de prova	<0,0001
Ajustamento preferível	Binomial Negativa

Os resultados apresentados na Tabela 5.19 indicam que o modelo Binomial Negativa é preferível ao modelo de Poisson.

Capítulo 6

Conclusão

No trabalho apresentado foram estudados os Modelos Lineares Generalizados em particular o Modelo de Regressão Logística, o Modelo de Regressão de Poisson e o Modelo de Regressão Binomial Negativa. Inicialmente desenvolveram-se modelos estatísticos para identificar os principais fatores associados à recuperação dos doentes com a SARS-CoV-2, utilizando o Modelo de Regressão Logística.

Estes dados referem-se ao número de doentes contaminados com o COVID-19 nas Filipinas no mês de fevereiro de 2020.

Uma análise exploratória dos dados foi realizada, pois ela pode sugerir se existe uma associação entre a variável resposta e as variáveis explicativas. No modelo inicialmente ajustado foi utilizado a distribuição binomial onde foram incluídas as variáveis explicativas **Idade**, **Sexo**, **Nacionalidade**, **Transmissão** e a variável resposta *Status* (estado do doente) para saber se o doente recuperou ou não recuperou da doença da COVID-19. Através deste notou-se que existe uma associação significativa entre a variável resposta ***(Status)*** com a variável **Idade** e a variável **Transmissão** e não existe associação significativa com as variáveis **Nacionalidade** e **Sexo**, ao nível de significância de 5%.

Concluíu-se que a chance de não recuperar aumenta com a idade em cerca de 9% com aumento de uma unidade na idade. Quanto à variável transmissão, verifica-se que a chance de um paciente não recuperar é 5,12 vezes maior em transmissão local em relação a transmissão exportada.

Numa segunda abordagem, desenvolveram-se modelos estatísticos para identificar que influencia o número de doentes recuperados da COVID-19.

Ajustou-se o modelo de regressão de Poisson onde se verificou que existem sobredispersão, razão pela qual foi utilizada o modelo de Regressão Binomial Negativa e verificou-se que este modelo era preferível em comparação ao modelo anterior. As variáveis explicativas utilizadas foram **Casos**, **Mortes**, **Testes**, **Ativos**, **Crítico**, **População** e **Expectativa** e a variável resposta é **Total de Recuperado**. Os resultados da análise mostraram que quanto maior a expectativa de vida no país maior o número total de doentes recuperados.

Para trabalho futuro sugerimos a aplicação de outras metodologias para modelar dados de contagem. Outrossim, seria interessante continuar este trabalho tendo já acesso a dados mais completos e atuais sobre a doença COVID-19.

Bibliografia

Agresti, A. (2013). *Categorical Data Analysis. 3rd ed. New York*. John Wiley Sons, Inc. New York.

Cadima, J. (2018). Modelos matemáticos e aplicações modelos lineares generalizados.

Cordeiro, Gauss Moutinho e Neto, E. d. A. (2004). Modelos paramétricos. *Pernambuco: UFRPE*.

Coxe, S., W. S. e. A. L. (2009). The analysis of count data: A gentle introduction to poisson regression e its alternatives. *Journal of personality assessment*, 91:121–136.

Gauss Moutinho Cordeiro, G. e. C. G. D. (2013). Modelos lineares generalizados e extensões. *Pernambuco: UFRPE*.

Hein, S. e Weiskittel, A. R. (2010). Cutpoint analysis for models with binary outcomes: a case study on branch mortality. *Eur J of Forest Res*, 129:585–590.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press, 2nd edition.

Hosmer, D e Lemeshow, S. (1989). *Applied Logistic Regression* wiley & sons. *New York*.

Mair, P, R. S. P. e. B. P. M. (2008). Quality of fit using logistic regression approaches.

Nelder, J.A e Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*.

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.

Santos, J. (2013). *Modelos para dados de contagem com excesso de zeros. Tese de Mestrado, Universidade do Minho*. PhD thesis.

Turkman, Maria A Amaral e Silva, G. L. (2000). Modelos lineares generalizados: da teoria á prática. *Sociedade Portuguesa de Estatística, Lisboa*.