UNIVERSIDADE DE CABO VERDE
FACULDADE DE CIÊNCIAS E TECNOLOGIA
MESTRADO EM MATEMÁTICA E APLICAÇÕES

Aplicação da Regressão de Poisson e Logística para Dados de Internamento de Pacientes com COVID-19 no Hospital Universitário Agostinho Neto

Rosialy Monteiro Fonseca
Orientadores: Prof. Doutor José Moniz Fernandes | Profª Doutora Paula Amaral

Com o apoio da Fundação Calouste Gulbenkian

Praia, Abril de 2022

UNIVERSIDADE DE CABO VERDE
FACULDADE DE CIÊNCIAS E TECNOLOGIA
MESTRADO EM MATEMÁTICA E APLICAÇÕES

Aplicação da Regressão de Poisson e Logística para Dados de Internamento de Pacientes com COVID-19 no Hospital Universitário Agostinho Neto

Rosialy Monteiro Fonseca
Orientadores: Prof. Doutor José Moniz Fernandes | Profª Doutora Paula Amaral

Praia, Abril de 2022

It is not the knowlegde, but the act of learning,
not the posession but the act of getting there,
which grants the greatest enjoyment.
Carl Friedric Gauss

Dedicatória A minha amada mãe.

O júri :

Presidente _____

Arguente _____

Orientador *Prof. Doutor José Moniz Fernandes e Profª Doutora Paula Amaral*, Faculdade de Ciências e Tecnologia, Universidade de Cabo Verde, Cidade da Praia, Cabo Verde e Faculdade de Ciências e Tecnologia, Universidade Nova Lisboa, Cidade Lisboa, Portugal.

Agradecimentos

Sem o precioso apoio de várias pessoas a presente dissertação de mestrado não poderia chegar a bom parto. Gostaria de agradecer de uma forma geral a todas as pessoas que acompanharam e foram fundamentais para a realização de mais este sonho.

Especificamente, queria primeiramente agradecer aos meus orientadores pela orientação, motivação, competência, profissionalismo, dedicação, empenho e sentido prático com que sempre orientaram-me neste trabalho. Proporcionaram as condições ideais de pesquisa, bem como também ajudaram a ter a persistência necessária para a conclusão do trabalho.

Agradeço igualmente a todos os meus professores do mestrado, bem como aos meus colegas, cujo apoio e a amizade foi importantíssima para atingir os objetivos, fica aqui minha imensa gratidão.

Por último, quero agradecer à minha família e amigos pelo apoio incondicional, fizeram sempre questão que continuassem os meus estudos, tivesse uma boa formação, instigando a ser sempre a melhor versão de mim. Em especial aos meus pais que sempre foram referências na minha vida, sempre acreditarem em mim e por terem abdicado de suas vidas em prol das realizações e da felicidade de seus filhos. Aos meus irmãos, pelo amor, pela irmandade, preocupação, carinho e incentivo.

Resumo

Os estudos apresentados nesta dissertação resultam da necessidade de encontrar respostas para os diferentes desafios identificados no internamento da COVID-19 do Hospital Univérsitario Agostinho Neto (HUAN), com foco na identificação dos fatores associados à morte e o número de dias de internamento dos pacientes internados com COVID-19. Foram abordadas diferentes análises estatísticas para desenvolver modelos explicativos e de análise, com a utilização do Software R Studio. Para resumir as principais características da base de dados realizou-se uma análise exploratória dos dados. Foram explorados modelos de Regressão de Poisson que permitiram modelar os dias de internamento, obtendo as variáveis com maior relevância estatísca no modelo. A distribuição de Poisson é uma referência para os modelos de regressão para dados de contagem, no entanto, a restrição de equidispersão, isto é, o valor esperado e a variância condicional são iguais, não representa com precisão os dados reais. Para o estudo dos óbitos, foi utilizada uma análise de Regressão Logística, que é uma técnica para mineração de dados de resposta binária, que possibilita a estimação da probabilidade de ocorrência de eventos. Este é um tipo de modelo linear generalizado que utiliza a função de ligação logit. Foram discutidos métodos da avaliação do modelo ajustado, bem como a averiguação das suposições e avaliação dos fatores que contribuíram para a morte dos pacientes. No sentido de escolher o melhor modelo realizou-se a comparação através do teste de verossimilhança e do AIC.

Palavras-chave: COVID-19, Dias de Internamento, Óbitos, Modelo, Regressão de Poisson, Regressão Logística.

Abstract

The study presented in this dissertation results from the need to find answers to the different challenges identified in the COVID-19 hospitalization at Hospital Univérsitario Agostinho Neto (HUAN), with a focus on identifying the factors associated with death and the number of days of hospitalization of COVID-19 inpatients. Different statistical analyses were addressed to develop predictive models, with the use of R Studio Software. To summarize the main characteristics of the database an exploratory data analysis was performed. Poisson Regression models were explored that allowed modeling the days of hospitalization, obtaining the variables that most demonstrated a high statistical relevance. The Poisson distribution is a reference for regression models for counting, however, the equidispersion constraint, that is, the expected value and the conditional variance are equal, does not accurately represent the actual data. For the study of deaths, a Logistic Regression analysis was conducted, which is a technique for mining binary response data that enables estimation of the probability of occurrence of events. It is a type of generalized linear model that uses the logit binding function. Methods for evaluating the adjusted model were discussed, as well as the investigation of the assumptions and evaluation of the factors that contributed to the death of patients. In order to choose the best model was performed the comparison through the likelihood ratio test and the AIC.

Keyswords: COVID-19,Days of Hospitalization, Deaths, Model, Poisson Regression, Logistic Regression

Conteúdo

Lista de Figuras iii

Lista de Tabelas iv

Lista de Símbolos e Abreviaturas v

1 Introdução 1

Introdução 1
1.1 Contexto e Motivação . . . 2
1.2 Objetivos . . . 3
1.3 Estrutura do trabalho . . . 3

2 Análise de Regressão 5
2.1 Análise de Regressão Linear . . . 5
2.1.1 Análise de Regressão Linear Simples . . . 5
2.1.2 Análise de Regressão Linear Múltipla . . . 10
2.2 Modelo de Regressão para Dados de Contagem . . . 16
2.2.1 Modelo de Regressão de Poisson . . . 16
2.2.2 Distribuição de Poisson Generalizada . . . 19
2.3 Análise de Regressão Logística . . . 22
2.3.1 Regressão Logística Simples . . . 23
2.3.2 Regressão Logística Múltipla . . . 26

3 Análise Exploratoria dos Dados 31
3.1 Caracterização do Estudo . . . 31
3.2 Manipulação dos Dados . . . 31
3.3 Variáveis Analisadas . . . 32
3.4 Análise Exploratória . . . 32

4 Modelos de Regressão: Aplicação e Interpretação dos Resultados 36
4.1 Análise de Dias de Internamento . . . 36
4.2 Análise de Óbitos . . . 41

5 Conclusão **50**

Conclusão **50**
5.1 Trabalhos futuros 51

Bibliografia **52**

Lista de Figuras

2.1 Ilustração da RLS 6
2.2 Ilustração dos resíduos 7
2.3 Relação entre o valor esperado e a varânça para diferentes valores de α . . 21
2.4 Relação entre o valor esperado e a varânça para diferentes valores de α . . 22

3.1 Classe Etária por Género 33
3.2 Dias por Classe Etária e Género 33

4.1 Análise de Resíduos - Dias de Internamento por Classe Etária 37
4.2 Análise de Resíduos - Dias de Internamento por Género 38
4.3 Análise de Resíduos - Moelodo de Regressão de Poisson 41
4.4 Resíduos de Pearson e Deviance - Model 1 44
4.5 Pontos Influentes - Model 1 45
4.6 Pontos Influentes - Model 1 45
4.7 Efeito:ARDS - Model 1 46
4.8 Efeito:Alcoolismo - Model 1 46
4.9 Efeito:Diabete - Model 1 47
4.10 Efeito:Ventilação - Model 1 47
4.11 Efeito - Model 1 48
4.12 Efeito:IRA - Model 1 48
4.13 Curva ROC - Model 1 49

Lista de Tabelas

2.1 Análise de variância (ANOVA) sobre o modelo de RLM 13

3.1 Óbitos por Género Classe e Etária 33
3.2 Doenças Associadas por Genero Classe e Etária 34
3.3 Oxigênio e Ventilação por Genero Classe e Etária 35
3.4 Correlação entre as variáveis Dias e Doenças Associadas 35
3.5 Correlação entre as variáveis Óbitos e Doenças Associadas 35

4.1 Modelo de Regressão de Poisson - Dias de Internamento por Classe Etária 36
4.2 Modelo de Regressão de Poisson - Dias de Internamento por Género 37
4.3 Modelo de Regressão de Poisson - Dias de Internamento 39
4.4 IC(95%) - Modelo de Regressão de Poisson 39
4.5 ANOVA - Modelo de Regressão de Poisson 40
4.6 Modelo de Regressão Logística - Model 1 42
4.7 Modelo de Regressão Logística - Model 2 42
4.8 Teste de Verosimilhança - Model 1 e Model 2 42
4.9 AIC - Model 1 e Model 2 43
4.10 IC (95%) - Model 1 43
4.11 ANOVA - Model 1 43
4.12 OR - Model 1 44
4.13 Multicolinearidade - Model 1 46

Lista de Símbolos e Abreviaturas

AIC	Critério de Informação de Akaike
OR	Odds Ratio
ROC	Receiver Operating Characteristic Curve
DW	Durbin Watson
HUAN	Hospital Universitário Agostinho Neto
IC	Intervalo de Confiança
MMQ	Método dos Mínimos Quadrados
RLM	Regressão Linear Multipla
RLS	Regressão Linear Simples
ROC	Receiver Operating Characteristic
SQR	Soma dos Quadrados Explicativas
SQR	Soma dos Quadrados Residuos
SQT	Soma dos Quadrados Totais

Capítulo 1

Introdução

A Organização Mundial de Saúde (OMS) foi alertada na data de 31 de dezembro de 2019 acerca de casos de pneumonia advindos de uma nova cepa de coronavírus que não havia sido anteriormente identificada em seres humanos. Tais casos foram notificados na cidade de Wuhan, China, a 7 de janeiro de 2020 e as autoridades chinesas confirmaram que haviam identificado um tipo de coronavírus que não se enquadra nos outros já conhecidos. Em 11 de março de 2020, a OMS caracteriza a COVID-19 como uma pandemia, pela existência de surtos dessa doença em diversos países e regiões do mundo [13].

A COVID-19 é uma doença infecciosa causada pelo coronavírus SARS-CoV-2 [6] e tem como principais sintomas febre, cansaço e tosse seca. Outros sintomas menos comuns e que podem afetar alguns pacientes são: perda de paladar ou olfato, congestão nasal, conjuntivite, dor de garganta, dor de cabeça, dores nos músculos ou juntas, diferentes tipos de erupção cutânea, náusea ou vômito, diarreia, calafrios ou tonturas. Em 26 de novembro de 2021, a OMS designou a variante da COVID-19 *B*.1.1.529 como uma variante de preocupação denominada Ômicron, cuja principal característica é ser altamente transmissível, quando comparada a variantes de preocupação já conhecidas do SARS-CoV2 [25]. Essa variante apresenta um grande número de mutações.

As outras variantes que ainda estão em circulação são: Alfa, Beta, Gama e Delta. Dessa forma, quanto mais o vírus da COVID-19 circular, através da movimentação das pessoas, mais oportunidades terá de sofrer mutações. Portanto, para reduzir o risco de exposição ao vírus é necessário se vacinar contra a COVID-19, fazer o uso de máscaras, manter a higiene das mãos, deixar os ambientes bem ventilados sempre que possível, evitar aglomerações e reduzir ao máximo o contato com pessoas.

Desde o início do surto da COVID-19, as autoridades em Cabo Verde vêm adotando medidas no sentido de prevenir e evitar a propagação da infeção da SARS-CoV-2 no país. No seguimento, o Governo elaborou um plano nacional de contingência (10 de março de 2020), com fortes medidas de proteção e restrições, seguindo as diretrizes emanadas da OMS.

Cabo Verde, registou o primeiro caso importado em 19 de março de 2020, na ilha da Boa Vista. Com o surgimento desse paciente infetado pelo novo coronavírus, intensificaram-

se as medidas restritivas e de saúde pública. Em 25 de março de 2020 foi confirmado o primeiro caso positivo de COVID-19 na cidade da Praia, o quarto do país e o primeiro cidadão nacional a infetar-se pelo novo coronavírus. Perante o sucedido, as autoridades locais redobraram as medidas restritivas, levando a declaração do Estado de emergência a nível nacional (Decreto Presidencial nº 06/2020), por um período de 20 dias. Com o aumento de casos positivos da COVID-19, e com três Ilhas afetadas (Boa Vista, Santiago e São Vicente), foi prorrogado o estado de emergência (primeira prorrogação) [12].

Até ao dia 16 de maio de 2020, tinham sido notificados no país, 328 casos acumulados de COVID-19, 3 óbitos e 84 recuperados. A Cidade da Praia tornou-se o epicentro da doença com 262 casos, 30 recuperados e 2 óbitos [12].

Em Cabo Verde, o Instituto Nacional de Saúde Pública (INSP) é a entidade responsável pelas atividades de comunicação de risco, informação, comunicação e sensibilização da população sobre as medidas preventivas sobre a COVID-19.

1.1 Contexto e Motivação

Com o crescente avanço das novas tecnologias de informação, a velocidade, capacidade dos sistemas e uma melhoria notável no sistema de gerenciamento de base de dados, tem possibilitado gerar base de dados a partir de grandes quantidades de dados, e os métodos tradicionais de análise de dados têm-se tornado insuficientes no sentido de obter informações úteis para a tomada da decisão, levando à consideração a precisão, variabilidade e certeza presente nos dados. O surgimento da Big-Data e os 5 Vs (Volume, a Velocidade, a Variedade, a Veracidade e o Valor) fez com que novas técnicas de análise de dados fossem criadas, necessidades de novas tecnologias para armazenamento, extração de conhecimento, representação, tratamento e análise dos dados, sobretudo de modo eficiente [14].

A mineração de dados é uma ferramenta de apoio para tomada de decisão, baseada em dados computacionais. Este processo pode empregar algoritmos de inteligência artificial, análise estatística, reconhecimento de padrões e outros [4].

Através dos recursos estatísticos oferecidos, dentro da abrangência da análise multivariada pode-se fazer o uso da análise de regressão para encontrar uma função/modelo para explicar o comportamento de um conjunto de dados. O método de regressão tem-se tornado um componente integral para qualquer análise de dados[9]. A análise de regressão define um conjunto vasto de técnicas estatísticas usadas para modelar relações entre variáveis e predizer o valor de uma ou mais variáveis dependentes, a partir de um conjunto de variáveis independentes [15].

Os modelos de regressão são vários e são definidos consoante o número e distribuição das variáveis explicativas, cujos efeitos na variável explicada ou de resposta se pretende estudar. Caso se pretenda estudar a relação de apenas uma variável explicativa com a variável resposta aplica-se a Regressão Linear Simples (RLS). Mas se pretende relacionar a variável resposta com mais do que uma variável explicativa, então aplica-se um modelo de Regressão Linear Múltipla (RLM). Caso a variável resposta seja uma variável categórica, utiliza-se o Modelo de Regressão Logística.

A RLM é uma ferramenta da estatística experimental que envolve três ou mais variáveis, das quais apenas uma é considerada como dependente das demais, chamadas de independentes, explanatórias, covariáveis ou regressoras [1].

Os modelos de regressão de dados de contagem são uma ferramenta muito utilizada para análise de dados, onde a variável dependente assume apenas valores inteiros não negativos observados um dado período de interesse. A distribuição de Poisson é conhecida por assumir que o valor esperado e a variância são iguais (equidispersão).

A Regressão Logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias [11]. É amplamente usada em vários campos, incluindo a aprendizagem automática (machine learning), com inúmeras aplicações em diversas áreas, incluindo medicina e as ciências sociais [5] .

A análise de dados categorizados é uma parte integrante da análise multivariada, que visa evidenciar e interpretar a informação relevante, contida em dados provenientes de contagem de eventos [18].

Neste trabalho serão aplicados os vários modelos referidos, em problemas de contexto real (internamento de COVID-19), através do software estatístico R Studio, que é um ambiente integrado de desenvolvimento para o R. É uma linguagem de programação multiparadigma orientada a objetos, programação funcional, dinâmica, voltada à manipulação, análise e visualização de dados.

1.2 Objetivos

O objetivo principal deste trabalho é desenvolver um modelo de regressão logística para o número de óbitos, identificando os fatores associados à morte dos pacientes internados com base nos dados de internamento de COVID-19 do Hospital Universitário Agostinho Neto (HUAN). Em detalhe, no presente trabalho, foram definidos os seguintes objetivos específicos:

- Realizar uma análise descritiva dos dados, sobre as caraterísticas dos pacientes internados;
- Estudar o número de dias de internamento, em função de alguns dados biométricos, realizando análises de regressão de Poisson;
- Desenvolver um modelo de regressão logística binária para o número de óbitos, obtendo-se os fatores associados.

1.3 Estrutura do trabalho

Além do capítulo introdutório, este trabalho está dividido em quatro capítulos, a saber:

- O Capitulo 2, Análise de Regressão: consiste na definição dos modelos de regressão para dados de contagem (Possison), dos modelos de regressão linear (simples e múltipla) e modelos de regressão logística (simples e múltipla) respetivamente, na sua forma analítica e matricial, bem como da estimação dos parâmetros, análise da variância, análise dos resíduos, intervalo de confiança, teste de significância, entre outros;

- O capítulo 3, Análise Exploratória: consiste na caraterização do estudo, manipulação dos dados, variáveis analisadas e uma análise exploratória dos dados;

- O capítulo 4, Modelos de Regressão - Aplicação e Interpretação dos Resultados: Neste capítulo desenvolvido uma análise dos dados, utilizado R, para a obtenção análises de regressões para o estudo de dias de internamento e dos óbitos;

- O capítulo 5, Conclusão: serão descritas as conclusões gerais desta dissertação quanto aos modelos analisados e avaliados após a sua aplicação, bem como a apresentação de sugestões para trabalhos futuros.

Capítulo 2

Análise de Regressão

Análise de regressão é um método estatístico utilizado para estudar a relação existente entre variáveis (duas ou mais variáveis) através da construção de uma equação/modelo. Um conjunto de variáveis são consideradas explicativas (independentes) de uma outra (explicada ou de resposta), a variável dependente, e pretende-se ajustar um modelo que visa encapsular as relações entre esses dois conjuntos de variáveis, partindo de um conjunto de observações. Esta técnica permite assim determinar com confiança quais são os fatores mais importantes, quais podem ser ignorados e como eles se influenciam mutuamente.

Os modelos de regressão devem ser utilizados com embasamento teórico subjacente, buscando estimar o melhor modelo e analisar os resultados obtidos por meio de testes estatísticos [8].

2.1 Análise de Regressão Linear

2.1.1 Análise de Regressão Linear Simples

A Regressão Linear Simples (RLS) faz a modelação uma equação matemática linear que descreve o relacionamento entre duas variáveis [7].É uma técnica muito utilizada, o resultado é uma equação que pode ser utilizada para fazer projeções ou estimativas dos dados.

Define-se como a relação linear entre a variável dependente (Y) e uma variável independente (X).

Modelo Teórico

A equação representativa do modelo de RLS é dada por:

$$y_\imath = \beta_0 + \beta_1 x_\imath + \varepsilon_\imath, \quad \imath = 1, ..., n \tag{2.1}$$

onde:

- $y_\imath$ representa o valor da variável resposta ou dependente (Y)
- $x_\imath$ representa o valor da variável independente (X);
- $\varepsilon_\imath$ são variáveis aleatórias que correspondem ao erro (variável que permite explicar a variabilidade existente em Y e que não é explicada por X);
- β_0 e β_1 correspondem aos parâmetros do modelo.

O parâmetro β_0 representa o ponto em que a recta regressora corta o eixo dos yy quando quando $X = 0$, chamado de intercepto ou coeficiente linear. O parâmetro β_1 representa o declive da recta regressora, expressando a taxa de mudança em Y, ou seja, indica a mudança na média da distribuição de probabilidade de Y para um aumento de uma unidade na variável X.

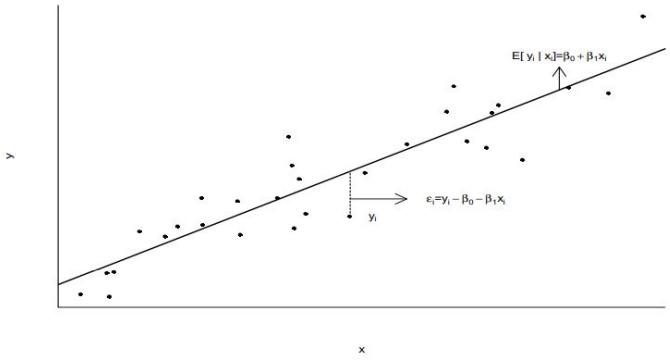

Fig. 2.1: Ilustração da RLS

A especificação do modelo de RLS implica que, condicional ao valor de X , a distribuição de Y tem as seguintes propriedades:

- A relação existente entre Y e X é linear;
- Os erros são independentes com média nula: $E(y_\imath) = E(\beta_0 + \beta_1 x_i + \varepsilon_\imath) = \beta_0 + \beta_1 x_i$
- A variância do erro é constante: $Var(y_\imath) = Var(\beta_0 + \beta_1 x_i + \varepsilon_\imath) + Var(\beta_0 + \beta_1 x_i) + Var(\varepsilon_\imath) = \sigma^2$
- Os erros $\varepsilon_\imath$, $\imath = 1, ..., n$ são normalmente distribuídos com $\epsilon_\imath = 0$.

Estimação dos Parâmetros: Método dos Mínimos Quadrados

O Método dos Mínimos Quadrados (MMQ) é uma abordagem padrão em análise de regressão para aproximar a solução de sistemas sobredeterminados (conjuntos de equações em que há mais equações do que incógnitas) minimizando a soma dos quadrados dos resíduos [15].

O primeiro passo na análise da regressão linear é obter as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ relativas aos parâmetros β_0 e β_1, minimizando os resíduos do modelo de regressão linear, calculados como a diferença entre os valores observados, $y_\imath$ e valores estimados, $\hat{y}_\imath$ fornecido por um modelo:

$$\varepsilon_\imath = y_\imath - \hat{y}_\imath, \imath = 1, ..., n \tag{2.2}$$

Os resíduos são representados pelas distâncias verticais entre os valores observados e os valores ajustados.

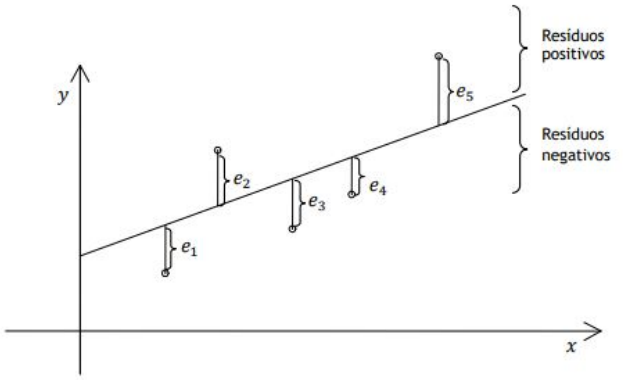

Fig. 2.2: Ilustração dos resíduos

O MMQ propõe então encontrar os valores β_0 e β_1 para os quais a Soma dos Quadrados dos Resíduos (SQE) é mínima. Tem-se então:

$$SQE = \sum_{\imath=1}^{n} e_\imath^2 = \sum_{\imath=1}^{n} (y_\imath - \hat{y}_\imath)^2 = \sum_{\imath=1}^{n} (y_\imath - (\beta_0 + \beta_1 x_\imath)^2 = \sum_{\imath=1}^{n} (y_\imath - \beta_0 - \beta_1 x_\imath)^2 \tag{2.3}$$

com $\sum_\imath^n e_\imath = 0$

Para determinar as estimativas, $\hat{\beta}_0$ e $\hat{\beta}_1$ é necessário obter as derivadas parciais:

$$\frac{\partial SQE}{\partial \beta_0} = -2 \sum_{\imath=1}^{n} (y_\imath - \beta_0 - \beta_1 x_\imath) \quad e \quad \frac{\partial SQE}{\partial \beta_1} = -2 \sum_{\imath=1}^{n} (y_\imath - \beta_0 - \beta_1 x_\imath) x_\imath \tag{2.4}$$

Para obter, $\hat{\beta}_0$ e $\hat{\beta}_1$ de forma a minimizar SNQ, realiza-se a substituição de β_0 e β_1 e igualam-se a zero as expressões das derivadas parciais, tal que :

$$-2 \sum_{\imath=1}^{n} (y_\imath - \hat{\beta}_0 - \hat{\beta}_1 x_\imath) = 0 \quad e \quad -2 \sum_{\imath=1}^{n} (y_\imath - \hat{\beta}_0 - \hat{\beta}_1 x_\imath) x_\imath = 0 \tag{2.5}$$

Obtendo assim, após simplificação, as seguintes equações:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad e \quad \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{2.6}$$

Obtém-se:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x_i} \tag{2.7}$$

Em que, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ são as médias das variáveis X e Y. Após substituição e simplificação, obtém-se:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.8}$$

Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, relativos aos parâmetros β_0 e β_1 assim determinados, são denominados estimadores dos mínimos quadrados, porque são a solução da reta ajustada pelo método dos mínimos quadrados dada pela expressão

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.9}$$

De seguida vamos apresentar algumas propriedades dos estimadores:

- Valor esperado de $\hat{\beta}_1$: $E[\hat{\beta}_1] = \beta_1$
- Variância de $\hat{\beta}_1$: $Var[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- Valor esperado de $\hat{\beta}_0$: $E[\hat{\beta}_0] = \beta_0$
- Variância de $\hat{\beta}_0$: $Var[\hat{\beta}_0] = \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})})$
- Covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$: $Cov[\hat{\beta}_0, \hat{\beta}_1] = -\sigma^2 \frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- Distribuição amostral de $\hat{\beta}_1$: $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})})$
- Distribuição amostral de $\hat{\beta}_0$: $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})}))$

Estimador de σ^2

A variância residual também deve ser estimada de forma a garantir análises precisas sobre as inferências feitas a partir do modelo linear. A estimativa da variância residual consiste no quociente entre a soma dos quadrados dos resíduos (ou erros), e os graus de liberdade, assumindo que os erros são independentes e com média zero. Assim, temos que:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2} \tag{2.10}$$

O número de graus de liberdade é determinado através da subtração do valor dos n casos e o número de parâmetros do modelo linear, ou seja $n-2$. A RLS requer apenas dois parâmetros. O estimador $\hat{\sigma}^2$ também é denominado Quadrado Médio dos Erros (QME):

$$\hat{\sigma}^2 = \frac{SQE}{n-2} = QME \tag{2.11}$$

Teste de Intervalos de Confiança para os Parametros do Modelo

Construiremos testes de hipóteses e intervalos de confiança para β_0 e β_1, considerando os pressupostos anteriormente referidos. Estes pressupostos levaram-nos a concluir que as observações $y_\iota \sim N(\beta_0 + \beta_1 x_\iota \sigma^2)$.

A realização de testes de hipóteses sobre o parâmetro β_1 é uma das formas de avaliar a capacidade da variável explicativa X como preditor da variável Y. A hipótese $\beta_1 = 0$ significa que não existe uma relação linear entre X e Y. Com as variâncias e distribuições anteriormente definidas, o teste de hipóteses consiste em testar:

$$H_0 : \beta_1 = 0$$

$$vs$$

$$H_1 : \beta_1 \neq 0$$

A estatística de teste poderá ser reescrita da seguinte forma:

$$T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t(n-2) \tag{2.12}$$

onde, $S_{\hat{\beta}_1} = \sqrt{\frac{QME}{\sum_{\iota=1}^{n}(x_\iota - \bar{x})^2}}$ e a estatística T tem uma distribuição t-student com $n-2$ graus de liberdade.

Suponhamos que pretendemos testar as hipóteses:

$$H_0 : \beta_1 = \beta_1^{\backprime}$$

$$vs$$

$$H_1 : \beta_1 \neq \beta_1^{\backprime}$$

A estatística de teste poderá ser escrita da seguinte forma:

$$T = \frac{\hat{\beta}_1 - \beta_1^{\backprime}}{S_{\hat{\beta}_1}} \sim t(n-2) \tag{2.13}$$

Logo, rejeita-se H_0, para um nível de significância α se $|T_{obs}| > t_{(1-\frac{\alpha}{2},n-2)}$.

O Intervalo de Confiança (IC) para $(1-\alpha)100\%$ é dado por:

$$[\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2},n-2)} S_{\hat{\beta}_1}]$$

.

Para o intercepto β_0, também são calculados IC. Assim, considerando as hipóteses:

$$H_0 : \beta_0 = 0$$
$$vs$$
$$H_1 : \beta_0 \neq 0$$

A estatística de teste poderá ser reescrita da seguinte forma:

$$T^0 = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} \sim t(n-2) \tag{2.14}$$

Consideremos também as hipóteses:

$$H_0 : \beta_0 = \beta_0^{\backprime}$$
$$vs$$
$$H_1 : \beta_0 \neq \beta_0^{\backprime}$$

A estatística de teste poderá ser escrita da seguinte forma:

$$T = \frac{\hat{\beta}_0 - \beta_0^{\backprime}}{S_{\hat{\beta}_0}} \sim t(n-2) \tag{2.15}$$

Logo, rejeita-se H_0, para um nível de significância α se $|T^0_{obs}| > t_{(1-\frac{\alpha}{2},n-2)}$.
O IC para $(1-\alpha)100\%$ é dado por:

$$[\hat{\beta}_0 \pm t_{(1-\frac{\alpha}{2},n-2)} S_{\hat{\beta}_0}]$$

.

2.1.2 Análise de Regressão Linear Múltipla

Anteriormente, sobre a RLS foram definidos os principais conceitos e técnicas para se analisar e utilizar a relação linear entre duas variáveis. Esta análise conduz a uma equação que pode ser utilizada para se estimarem valores de uma variável resposta de interesse Y dados valores de uma variável independente X associada. Por vezes, é necessário mais do que uma variável explicativa (regressora) para modelar a variável resposta de interesse.

A análise de Regressão Linear Múltipla (RLM) é uma metodologia estatística de previsão e estimação de valores de uma variável de resposta de interesse através de um conjunto de variáveis regressoras. Esta metodologia pode ser utilizada também para a avaliação dos efeitos das variáveis regressoras como previsoras das variáveis de resposta.

Os conceitos e técnicas para a realização e análise das relações lineares entre uma variável de resposta e várias variáveis regressoras são uma extensão natural do que foi apresentado na RLS.

Na RLM assume-se que existe uma relação linear entre uma variável y (variável dependente) e p variáveis independentes (preditoras), $x_1, ..., x_p$.

Modelo Teórico

O modelo da RLM é defenido da seguinte forma:

$$y_\imath = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_\imath, \quad \imath = 1, ..., n \tag{2.16}$$

onde:

- $y_\imath$ representa o valor da variável resposta ou dependente;
- $x_{i1}, ..., x_{ip}$, $\imath = 1, ..., n$ representa os valores da i-ésima observação das variáveis explicativas;
- β_0, β_1, β_2 ... β_p correspondem aos parâmetros ou coeficientes da regressão;
- $\varepsilon_\imath$ são variáveis aleatórias que correspondem ao erro.

As propriedades para o modelo de RLM, são análogicas às do modelo de RLS, tem-se:

- $E(\varepsilon_\imath) = 0, \imath = 1, ..., n$;
- Os erros são independentes;
- $V(\varepsilon_\imath) = \sigma^2$, $\imath = 1, ..., n$.

Destas propriedades, concluímos que $\varepsilon_\imath \sim N(0, \sigma^2)$, $\imath = 1, ..., n$; e consequentemente que y tem distribuição normal com varância σ^2.

Este modelo, devido às dificuldades de cálculo no manuseamento do elevado número de parâmetros, também é apresentado matricialmente por:

$$Y = X\beta + \varepsilon \tag{2.17}$$

sendo

$$AY = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 2 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ n & x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \dots & x_p \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

Sendo que,

- ε é um vector de dimensão $n \times 1$ cujas componentes são os erros aleatórios;
- Y é um vetor de dimensão $n \times 1$ cujas componentes correspondem às n respostas;
- X é uma matriz de dimensão $n \times (p + 1)$ denominada matriz do modelo;
- β é um vetor coluna cujos elementos são os coeficientes de regressão.

Estimação dos Parâmetros: Método dos Mínimos Quadrados

Usando o MMQ, pretendemos encontrar o vetor de estimadores $\hat{\beta}$, que minimiza

$$SQE = \sum_{i=1}^{n} e_i^2 = e`e = (Y - X\beta)` = Y`Y - Y`X\beta - \beta`X`Y + \beta`X`X\beta = Y`Y - 2\beta`X`Y + \beta`X`X\beta, \tag{2.18}$$

Uma vez que se tem $Y`X\beta = \beta`X`Y$ pois este produto é igual a um escalar.
Derivando $\hat{\beta}$ obtemos:

$$\frac{\partial SQE}{\partial \beta} = -2X`Y + 2X`X\beta \tag{2.19}$$

Igualando a derivada a zero e substituindo β por $\hat{\beta}$, obtemos:

$$-2X`Y + 2X`X\hat{\beta} = 0 \Leftrightarrow (X`X)\hat{\beta} = X`Y \Leftrightarrow \hat{\beta} = (X`X)^{-1}X`Y \tag{2.20}$$

Onde, X^{-1} representa a matriz inversa de X.
O modelo de regressão linear ajustado é dado po:

$$\hat{Y} = X\hat{\beta} \tag{2.21}$$

O vetor de resíduos, é dados por:

$$e = Y - \hat{Y} \tag{2.22}$$

O estimador de mínimos quadrados $\hat{\beta}$ é não enviesado e tem variância mínima entre todos os estimadores não enviesados que são combinações lineares dos y_i [29]. Serão algumas das propriedades dos estimadores:

- Valor esperado de $\hat{\beta}$: $E(\hat{\beta}) = \beta$
- Matriz de covariâncias de $\hat{\beta}$: $Cov(\hat{\beta}) = \sigma^2(X`X)^{-1}$

Estimador de σ^2

Consideremos a Soma do Quadrado dos Resíduos (SQE), defenida por:

$$SQE = Y`Y - 2\hat{\beta}`X`Y` + \hat{\beta}`X`Y = Y`Y - \hat{\beta}`X`Y = Y`Y - Y`X(X`X)^{-1}X`Y = Y`(l_n - X(X`X)^{-1}X`)Y \tag{2.23}$$

Se $Y \sim N(\mu; \Sigma)$ então, $Y`AY$ segue uma distribuição qui-quadrado não central com g graus de liberdade
Como assumimos que o vetor dos erros $\varepsilon \sim N(0; \sigma^2 l_n)$, segue que $Y \sim N(X\beta; \sigma^2 l_n)$.
Desta forma, obtemos que:

$$\frac{SQE}{\sigma^2} = \frac{Y`}{\sigma^2}[l_n - X(X`X)^{-1}X`]Y \sim x^2_{r;\delta} \tag{2.24}$$

com $r = cor(l_n - X(X`X)^{-1}X`)$ e

$$\delta = \frac{1}{2}\frac{\beta`X`(l_n - X(X`X)^{-1}X`)X\beta}{\sigma^2} = 0 \tag{2.25}$$

e $r = n - (p+1)$ então $\frac{SQE}{\sigma^2}$ segue uma distribuição qui-quadrado com $n - (p+1)$.
Portanto, um estimador não viciado para σ^2 é dado por:

$$\hat{\sigma}^2 = QME = \frac{SQE}{n-p-1} \tag{2.26}$$

Análise da Variância

A análise de variância, baseia-se na decomposição da Soma dos Quadrados Total (SQT), que corresponde à variação da variável resposta, na Soma dos Quadrados Explicada (SQR), que corresponde à variação da variável resposta que é explicada pelo modelo e na Soma dos Quadrados dos Resíduos (SQE), que corresponde à variação da variável resposta que não é explicada pelo modelo).

Podemos escrever:

$$SQT = SQR + SQE \Leftrightarrow \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.27}$$

Na RLM, as hipóteses a testar serão:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$vs$$

$$H_1 : \exists_j \beta_j \neq 0, j = 1, ..., p$$

Para testar a hipótese, H_0 utiliza-se a estatística de teste:

$$F = \frac{\frac{SQR}{p}}{\frac{SQE}{n-p-1}} = \frac{QMR}{QME} \sim F_{p,n-p-1} \tag{2.28}$$

Portanto, se $F_{obs} > F_{1-\alpha;p,n-p-1}$ rejeita-se a hipótese H_0, com F_{obs}, o valor observado da estatística F e $F_{1-\alpha;p,n-p-1}$ o quantil $1-\alpha$ de distribuição F.Ao rejeitar H_0 concluimos que pelo menos uma das variáveis explicativas contribui significativamente para o modelo.

Tab. 2.1: Análise de variância (ANOVA) sobre o modelo de RLM

Variação	**Soma Quadrados**	**Graus Liberdade**	**Quadrados Médios**	F
Regressão	SQR	p	QMR=SQR/p	QMR/QME
Resíduo	SQE	n-p-1	QME=SQE/p	
Total	SQT	n-1		

Tal como na RLS, a proporção de variabilidade em Y explicada pelos termos regressores é dada pelo quociente.

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \tag{2.29}$$

Quanto mais próximo R^2 estiver de 1, maior é a explicação da variável de resposta pelo modelo ajustado. Dado que o valor de R^2 aumenta à medida que são adicionados termos ao modelo, é prudente utilizar um valor ajustado de R^2 [17].

Análise dos Resíduos

Como já foi visto anteriormente (RLS) os resíduos são dados pela diferença entre os valores da variável resposta observada e a variável resposta estimada.

Testes de Ajustamento à Normalidade: O teste de Kolmogorov-Smirnov (K-S) é um teste de aderência que verifica o grau de concordância entre distribuições num conjunto de valores, com o objetivo de identificar se os dados seguem uma distribuição normal. Este teste utiliza a distribuição de frequência acumulada que ocorreria dada a descrição teórica, e compara esta com a distribuição de frequência acumulada observada.

O teste de K-S é utilizado para testar as hipóteses:

$$H_0 : \text{A distribuição é normal}$$
$$vs$$
$$H_1 : \text{A distribuição não é normal}$$

A estatística do teste espera que quando é verdadeira, as diferenças entre a proporção de casos esperados e a distribuição de frequências sejam pequenas e estejam dentro do limite dos erros aleatórios.

O Teste de Shapiro-Wilk (s-w) é determinar uma variável estatística (W) calculada sobre os valores amostrais ordenados e elevados ao quadrado, com o objetivo de aferir se uma amostra aleatória x_i que provém de uma distribuição normal. Este método tem sido profusamente adotado nos testes de normalidade utilizando-se preferencialmente em amostras de dimensão pequenas $n < 30$.

As hipóteses a serem testadas são as mesmas de K-S, e a variável W é calculada da seguinte forma:

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.30}$$

Em que, a_i são constantes geradas a partir da média, variância e covariância de n ordens [15]

Diagnóstico de Homocedasticidade (variância constante): A variância ser constante equivale a supor que não existem observações incluídas na variável residual cuja influência

seja mais intensa na variável dependente. Uma das técnicas usadas para verificar a suposição de que os resíduos são homocedásticos, é a análise do gráfico dos resíduos versus valores ajustados.

A homocedasticidade não se mantém quando a variância dos fatores não-observáveis muda ao longo de diferentes segmentos da população

Diagnóstico de Independência: Um dos testes mais conhecidos para verificação de autocorrelação ou independência dos resíduos é a estatística de Durbin Watson (DW), que envolve o cálculo de um teste estatístico baseado nos resíduos do método de regressão de mínimos quadrados.

O teste de DW testa as hipóteses:

$$H_0 : \text{Não existe autocorrelação dos resíduos}$$
$$vs$$
$$H_1 : \text{Existe autocorrelação positiva dos resíduos}$$

A estatística de teste é dada por:

$$DW = \frac{\sum_{\imath=2}^{n}(e_\imath - e_{\imath-1}^2)}{\sum_{\imath=1}^{n} e_\imath^2} \tag{2.31}$$

E toma os valores $0 \leq DW \leq 4$. Esta estatística mede a correlação entre cada resíduo e o resíduo correspondente à observação imediatamente anterior.

Diagnóstico de Outliers e Observações Influentes: Outliers são observações extremas que se encontram de tal forma afastadas da maioria dos dados que surgem dúvidas sobre se elas poderão ou não ter sido geradas pelo modelo proposto para explicar essa maioria dos dados [22].

Se um Outlier for influente vai interferir sobre a função de regressão ajustada, o que significa que a inclusão ou não desse ponto modifica substancialmente os valores ajustados. Assim, um ponto é influente se a sua exclusão na regressão ajustada provoca uma mudança substancial nos valores ajustados

Uma vez que uma observação pode ser considerada um Outlier e pode ou não ser uma observação influente é importante identificar quais as observações influentes.

As observações influentes são aquelas que individualmente ou em conjunto com as outras observações demonstram ter mais impacto do que as restantes no cálculo dos estimadores [24].

Colinearidade e Multicolinearidade: O termo colinearidade é utilizado para expressar a existência de correlação elevada entre duas variáveis independentes, enquanto o termo multicolinearidade é utilizado quando se trata de mais do que duas variáveis independentes fortemente correlacionadas. No entanto, existem autores que definem colinearidade como a existência de relação linear entre duas variáveis independentes e multicolinearidade como a existência de relação linear entre uma das variáveis independentes e as restantes.

Se considerarmos duas quaisquer variáveis independentes, X_1 e X_2 entre as quais existe uma elevada correlação, a proporção da variação total da variável dependente, explicada por X_1 é idêntica à proporção da variação total da variável dependente, explicada por X_1.

A colinearidade poderá ser diagnosticada:

- Verificando se a matriz de correlações das variáveis independentes demonstra correlações elevadas. Caso a correlação de duas variáveis seja muito próxima de 1, indica um problema;
- Verificando se, ao se realizar a regressão de X_i em função das outras variáveis independentes, o valor de $R \simeq 1$.

Um indicador usado com frequência para detectar a multicolinearidade é o Variance Inflation Factor (VIF).

2.2 Modelo de Regressão para Dados de Contagem

Os modelos de regressão para dados de contagem (dentro dos Modelos Lineares Generalizados) tem como foco estudar o comportamento de uma variável dependente, que assume apenas valores inteiros não negativos, com base no comportamento de variáveis explicativas. São amplamente utilizados nas mais diversas áreas de estudo para modelação de diversos fenômenos [19].

De entre as famílias dos Modelos Lineares Generalizados,o modelo de Poisson, é muito utilizado, visto que é um modelo mais simples [26]. A adequação do seu uso depende da correta especificação do modelo, assim como as conclusões tiradas. A hipótese de equilíbrio constitui um pré-requisito muito forte e rígido.

O modelo de regressão de Poisson tem por característica a análise de dados contados na forma de proporções ou razões de contagem [16].

A variável resposta de uma regressão de Poisson deve seguir uma distribuição de Poisson e os dados devem possuir igual dispersão, ou seja, a média da variável resposta deve ser igual à variância.

Neste capítulo iremos abordar o modelo simples de Poisson e o modelo de regressão de Poisson Generalizado.

2.2.1 Modelo de Regressão de Poisson

A modelização de um conjunto de observações de uma dada variável de contagem começa, normalmente, pela estimação do modelo de Poisson. As razões fundamentais para esta abordagem são: a distribuição de Poisson adequa-se totalmente à principal característica dos dados de contagem, o facto destes assumir apenas valores inteiros não negativos; após a sua estimação, é possível inferir relativamente à probabilidade de ocorrência futura de um dado acontecimento, admitindo que o modelo está bem especificado; o modelo de Poisson possui uma estrutura simples, podendo ser facilmente estimado [23].

Especificação

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

$$f(y_\imath, \mu_\imath) = P(Y_\imath = y_\imath) = \frac{exp(-\mu)\mu_\imath^{y_\imath}}{y_\imath!}, y_\imath = 0, 1, 2, 3, ... \tag{2.32}$$

sendo que, $\mu_\imath > 0$ representa o número médio de ocorrências de um dado acontecimento.
O valor esperado e a variância da distribuição de Poisson são iguais e são dados por:

$$E(Y) = \sum_{y=0}^{\infty} yP(Y = y) \tag{2.33}$$

Sendo,

$$E(Y) = \sum_{y=0}^{\infty} y\frac{exp(-\mu)\mu^y}{y!} = exp(-\mu)\sum_{y=0}^{\infty}\frac{\mu^y}{(y-1)!} \tag{2.34}$$

Subtituindo $\jmath = y - 1$, tem-se:

$$E(Y) = exp(-\mu)\sum_{y=\jmath+1}^{\infty}\frac{\mu^{\jmath+1}}{\jmath!} = \mu exp(-\mu)\sum_{\jmath=0}^{\infty}\frac{\mu^\jmath}{\jmath!} = \mu exp(-\mu)\mu exp(\mu) = \mu \tag{2.35}$$

e, de forma análogica , tem-se que:

$$Var(Y) = E(Y^2) - [E(Y)]^2 = E(Y) = \mu \tag{2.36}$$

Sejam $Y_1, Y_2, ..., Y_n$ uma amostra aleatória de uma variável aleatória Y que representa o número de ocorrências de um acontecimento raro num determinado período de tempo ou espaço. Dado um vetor de variáveis explicativas $X = (X_1, X_2, ..., X_P)$ e uma observação $X_\imath^T = (1, x_{\imath 1}, x_{\imath 2}, ..., x_{\imath p})$ do indivíduo $\imath$, assume-se que, $Y|X = x_\imath \sim P(\mu(x_\imath))$, onde, $\mu_\imath = \mu(x_\imath)$ e representa o número médio de ocorrências de um determinado acontecimento dada a observação $x_\imath$.Tem-se que:

$$\mu_\imath = E(Y|X = x_\imath) = Var(Y|X = x_\imath) \tag{2.37}$$

Modelando o valor esperado de $Y|X = x_\imath$ como combinação linear das variáveis explicativas, escreve-se um modelo linear simples na forma

$$\mu(X_\imath) = x_\imath^T\beta = \beta_0 + \beta_{1x_{\imath 1}} + ... + \beta_{px_{\imath p}} \tag{2.38}$$

O valor médio da distribuição de Poisson só pode tomar valores não negativos, sendo que este modelo é desajustado uma vez que o preditor linear, do lado direito, pode assumir qualquer valor real. Para solucionar este problema, passa por considerar a transformação logarítmica como uma função de ligação do modelo linear generalizado

$$log(\mu(X_\iota)) = \beta_0 + \beta_{1x_{\iota 1}} + ... + \beta_{px_{\iota p}} \tag{2.39}$$

O modelo de regressão de Poisson é definido por:

$$Y|X = x_\iota \sim P(\mu(x_\iota)) \tag{2.40}$$

e

$$log(\mu(X_\iota)) = \beta_0 + \beta_{1x_{\iota 1}} + ... + \beta_{px_{\iota p}} \tag{2.41}$$

Estimação

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

$$L(\beta) = \prod_{\iota=1}^{n} \frac{exp(-\mu_\iota)\mu_\iota^{y_\iota}}{y_\jmath!} \tag{2.42}$$

e a função de log-verossimilhança é dado por

$$l(\beta) = \sum_{\iota=1}^{n} [-\mu_\iota + y_\iota ln(\mu_\iota) - ln(y_\iota!)] \tag{2.43}$$

Sendo $\mu_\iota = exp(x_\iota^T \beta)$, temos que:

$$l(\beta) = \sum_{\iota=1}^{n} [-exp(x_\iota^T \beta) + y_\iota ln(exp(x_\iota^T \beta)) - ln(y_\iota!)] \tag{2.44}$$

Maximizando a função de log-verossimilhança $l(\beta)$, temos que:

$$\frac{l(\beta)}{l\beta_\jmath} = \sum_{\iota=1}^{n} [y_\iota - exp(x_\iota^T \beta)] \tag{2.45}$$

com $x_\iota = 0$ e $\jmath = 0, 1, 2, ..., p$.
A função desvio no modelo de regressão de Poisson é definida pela seguinte forma:

$$D(Y : \hat{\mu}) = 2 \sum_{\iota=1}^{n} \left(y_\iota ln(\frac{y_\iota}{\hat{\mu}_\iota}) \right) \tag{2.46}$$

A estatística de Pearson generalizada é definida por:

$$\chi = \sum_{\iota=1}^{n} \frac{(y_\iota - \hat{\mu}_\iota)^2}{\hat{\mu}_\iota} \tag{2.47}$$

2.2.2 Distribuição de Poisson Generalizada

Especificação

Uma variável aleatória Y que assume valores inteiros não negativos que segue uma Distribuição de Poisson Generalizada (DPG) com parâmetros θ e γ quando $\theta_i > 0$ e $0 \leq \gamma < 1$, a função de probabilidade é dada por [21]:

$$P(Y_i = y_i|\theta_i, \gamma) = \theta_i(\theta_i + \gamma y_i)^{y_i-1}\frac{exp(-(\theta_i + \gamma y_i))}{y_i!}, y_i = 0, 1, 2, 3, ... \tag{2.48}$$

Quando for 0, $y_i > m$ para $\gamma < 0$

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

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

$$P(Y_i = y_i|\theta_i, \gamma) = \frac{P(Y_i = y_i|\theta_i, \gamma)}{\sum_{y=0}^{m} f(Y_i = y_i|\theta_i, \gamma)} = \frac{P(Y_i = y_i|\theta_i, \gamma)}{F_m} \tag{2.49}$$

onde, $\theta < 0$ e $F_m = \sum_{y=0}^{m} f(Y_i = y_i|\theta_i, \gamma)$.

A soma das probabilidades passou a ser igual a 1, entretanto, o modelo é inadequado, visto que, emite valores negativos para θ_i.

Estimação

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

$$E[Y_i] = \frac{\theta_i}{1 - \gamma} \tag{2.50}$$

e

$$Var[Y_i] = \frac{\theta_i}{(1 - \gamma)^3} \tag{2.51}$$

Comparando-se estes momentos com os momentos amostrais, tem-se que os estimadores para θ e γ são obtidos como soluções das equações:

$$\frac{\tilde{\theta}_i}{1 - \tilde{\gamma}} = \frac{1}{n}\sum_{i=0}^{n} Y_i = \overline{Y} \tag{2.52}$$

e

$$\frac{\tilde{\theta}_i}{(1-\tilde{\gamma})^3} + \frac{(\tilde{\theta}_i)^2}{(1-\tilde{\gamma})^2} = \frac{1}{n}\sum_{i=0}^{n} Y_i^2 \tag{2.53}$$

que fornece os estimadores

$$\tilde{\theta}_i = \sqrt{\frac{\overline{Y}^3}{S^2}} \tag{2.54}$$

e

$$\tilde{\gamma} = 1 - \sqrt{\frac{\overline{Y}}{S^2}} \tag{2.55}$$

com $S^2 = \frac{1}{n}[\sum_{i=0}^{n} Y_i^2 - n\overline{Y}]$

Distribuição de Poisson Generalizada I

As parametrizações para a DGP, são definidos de forma diferente para vários autores. Para Wang e Famoye (1997) propuseram as seguintes parametrizações:

$$\theta_i = \frac{\mu_i}{1+\alpha\mu_i} \tag{2.56}$$

e

$$\gamma = \alpha\frac{\mu_i}{1+\alpha\mu_i} \tag{2.57}$$

A DPG I com parâmetros μ e α é definida da seguinte forma:

$$P(Y_i = y_i|\mu_i, \alpha) = \frac{(\mu_i)^{y_i}(1+\alpha y_i)^{y_i-1} exp\left(\frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right)}{(1+\alpha\mu_i)^{y_i} y_i!} \tag{2.58}$$

Quando for 0, $y_i > m$ para $\alpha < 0$. O valor esperado $E[Y_i] = \mu_i$ e a variança $Var[Y_i] = \mu_i(1+\alpha y_i)^2$.

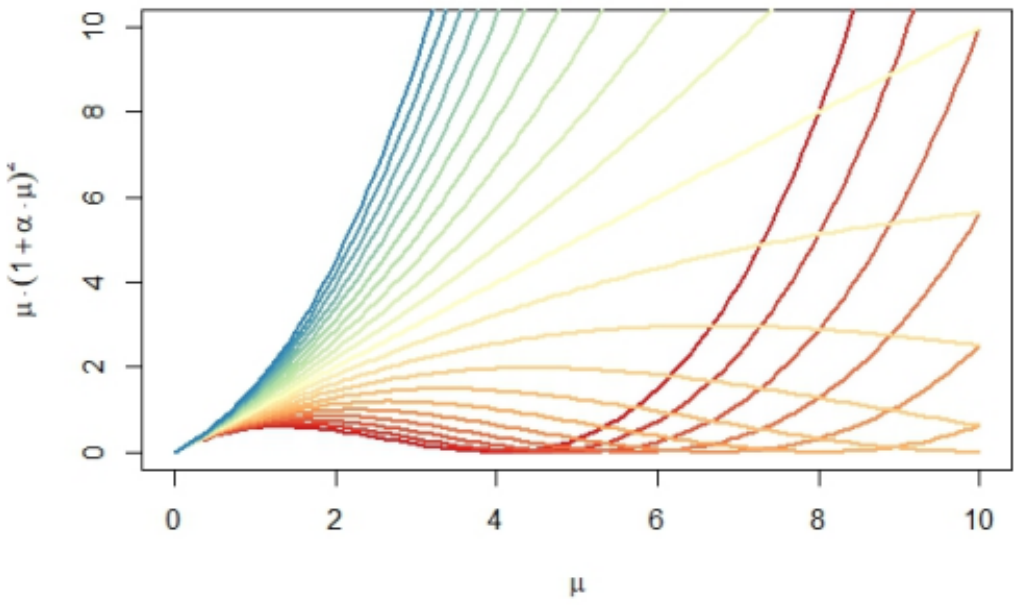

Fig. 2.3: Relação entre o valor esperado e a varânça para diferentes valores de α
DPG I

Distribuição de Poisson Generalizada II

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

$$\theta_i = \frac{\mu_i}{\alpha} \tag{2.59}$$

e

$$\gamma = \frac{\alpha - 1}{\alpha} \tag{2.60}$$

A DPG II é definida da seguinte forma:

$$P(Y_i = y_i|\mu_i, \alpha) = \mu_i(\mu_i + (\alpha - 1)y_i)^{y_i - 1}\frac{exp(-\alpha^{-1}(\mu_i + (\alpha - 1)y_i))}{y_i!} \tag{2.61}$$

Quando for 0, $y_i > m$ para $\alpha < 1$. O valor esperado $E[Y_i] = \mu_i$ e a variança $Var[Y_i] = \alpha^2\mu_i$.

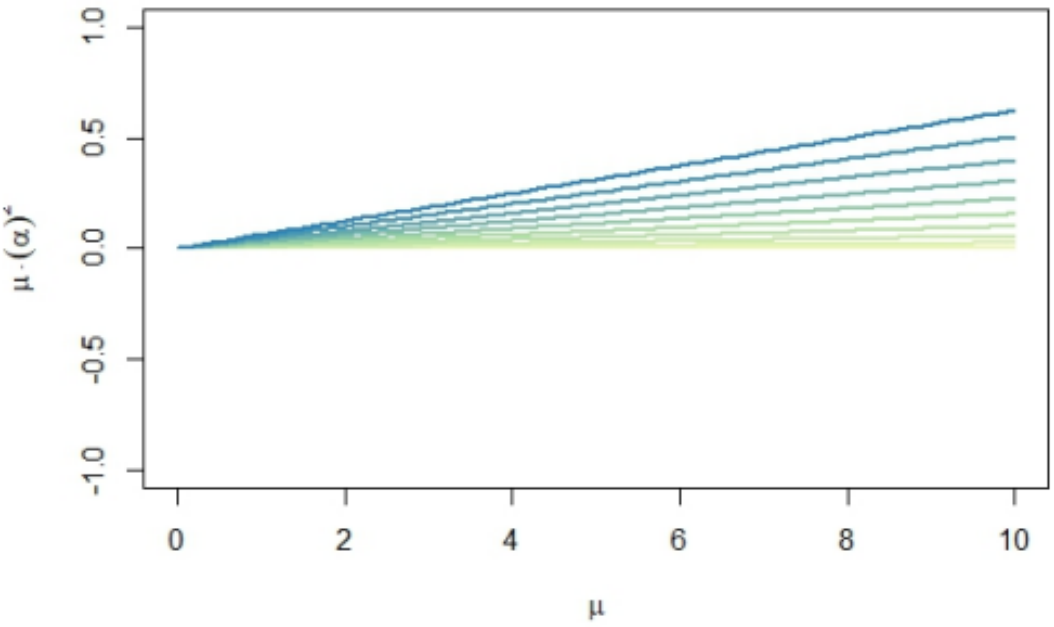

Fig. 2.4: Relação entre o valor esperado e a varânça para diferentes valores de α na DPG II

2.3 Análise de Regressão Logística

O modelo de regressão logística é um tipo de modelo linear generalizado que utiliza a função de ligação logit para a resposta que segue uma distribuição binomial (binária) [27].

É uma técnica estatística utilizada para descrever o comportamento entre uma variável dependente binária (0 ou 1) e variáveis independentes contínuas e/ou binárias. O fato da variável dependente ser binária (0 ou 1) possibilita associações de classificação dos fenômenos e interpretações em termos de probabilidade de chance do fenômeno investigado ocorrer ou não ocorrer [5].

De uma forma geral, existem 3 procedimentos distintos para manipulação de dados, na regressão logística: binária, ordinal e nominal. A escolha do método depende do número de categorias e das características da variável resposta. Uma variável binária é aquela que aceita dois níveis de resposta, uma variável ordinal segue uma classificação ou uma ordenação natural e a variável nominal pode ter mais de 3 níveis e não considera nenhuma ordenação [20]

Uma análise de regressão logística binária pode ser simples ou múltipla. No primeiro caso o modelo é constituído por apenas uma variável independente, enquanto no segundo caso o modelo é uma generalização do primeiro, o qual é constituído por mais de uma covariável.

2.3.1 Regressão Logística Simples

Seja x_ι a variável explicativa e y_ι o número de ocorrências de um evento, em que $\iota = 1, ..., n$ representa o número de observações. Assume-se ainda, que a variável resposta tem distribuição binomial ($Y_\iota \sim B(\pi)$) com $\pi = E(Y_\iota)$

$$P[Y_\iota = y_\iota] = \pi_\iota^{y_\iota}(1 - \pi_\iota)^{1-y_\iota} \tag{2.62}$$

Por forma de adequar a resposta média ao modelo linear é utilizada a função de ligação

$$\pi(x_\iota) = \frac{e^{\beta_0+\beta_1 x_\iota}}{1 + e^{\beta_0+\beta_1 x_\iota}}, \quad \iota = 1, ..., n \tag{2.63}$$

Para facilitar a obtenção dos parâmetros β_0, β_1 pode-se linearizar o modelo logístico, aplicando-se o logaritmo natural da razão e o resultado desta transformação são números reais. Em síntese a seguinte transformação, chamada de transformação logit da probabilidade, que pode ser escrita por:

$$g(x) = ln\left(\frac{\pi_\iota}{1 - \pi_\iota}\right) = \beta_0 + \beta_\iota x_\iota \tag{2.64}$$

Esta transformação é importante pois permite à função de ligação g(x) manter propriedades associadas ao modelo de regressão linear [10]

Estimação dos Parâmetros do Modelo

Para estimar os parâmetros β_0 e β_1, é utilizado o método da máxima verossimilhança, que consiste em determinar os valores dos parâmetros que maximizem a probabilidade de obter o conjunto de valores observado.

A função de verossimilhança tem a seguinte forma:

$$P[Y_i = y_\iota, ..., y_n|\beta_0, \beta_1] = \prod_{\iota=1}^{n} \pi_i^{y_\iota}(1 - \pi_\iota)^{1-y_\iota} = \prod_{\iota=1}^{n}\left(\frac{\pi_\iota}{1 - \pi_\iota}\right)(1 - \pi_\iota)^{1-y_\iota} \tag{2.65}$$

Aplicando logaritmo dos dois lados, a expressão fica

$$lnL(\beta_0, \beta_1|(x_\iota, y_\iota)) = \sum_{\iota=1}^{n} y_\iota(\beta_0 + \beta_1 x_\iota) - \sum_{\iota=1}^{n} ln(1 + e^{\beta_0+\beta_1 x_\iota}) \tag{2.66}$$

Os estimadores de máxima verossimilhança para os parâmetros β_0 e β_1 são os valores $\hat{\beta_0}$ e $\hat{\beta_1}$ que maximizam o logaritmo da função de verossimilhança.

Para maximizar a função de verossimilhança é necessário derivar a expressão em relação aos parâmetros do modelo, da seguinte forma.

$$\frac{\partial}{\partial\beta_0} lnL(\beta_0, \beta_1|(x_\iota, y_\iota)) = \sum_{\iota=1}^{n} y_\iota - \sum_{\iota=1}^{n} \frac{e^{\beta_0+\beta_1 x_\iota}}{1 + e^{\beta_0+\beta_1 x_\iota}} \tag{2.67}$$

$$\frac{\partial}{\partial\beta_1} lnL(\beta_0,\beta_1|(x_\imath,y_\imath)) = \sum_{\imath=1}^{n} y_\imath x_\imath - \sum_{\imath=1}^{n} x_\imath \frac{e^{\beta_0+\beta_1 x_\imath}}{1+e^{\beta_0+\beta_1 x_\imath}} \tag{2.68}$$

O método da máxima verossimilhança, é definido como sendo os valores dos parâmetros que geram com maior frequência a amostra observada[3].

Ao igualar a zero e substituindo β_0 e β_1 por $\hat{\beta}_0$ e $\hat{\beta}_1$ obtém-se:

$$\sum_{\imath=1}^{n} y_\imath - \sum_{\imath=1}^{n} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}} = 0 \tag{2.69}$$

$$\sum_{\imath=1}^{n} y_\imath x_\imath - \sum_{\imath=1}^{n} x_\imath \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}} = 0 \tag{2.70}$$

Os resultados desta aplicação são incluídos na matriz denominada de Informação de Fisher. A matriz de informação de Fisher, para o modelo logístico com uma variável, tem a seguinte forma:

$$I(\hat{\beta}) = \begin{bmatrix} \sum_{\imath=1}^{n} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}}{(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath})^2} & \sum_{\imath=1}^{n} x_\imath \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}}{(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath})^2} \\ \sum_{\imath=1}^{n} x_\imath \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}}{(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath})^2} & \sum_{\imath=1}^{n} x_\imath^2 \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}}{(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath})^2} \end{bmatrix} \tag{2.71}$$

Após obter as estimativas dos parâmetros do modelo é possível calcular as probabilidades estimadas

$$\pi_\imath = \sum_{\imath=1}^{n} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_\imath}} \tag{2.72}$$

A interpretação dos parâmetros da regressão logística é semelhante aos da regressão linear, usando a função Odds Ratio - OR (Razão de chances).

Seja $g(x) = \frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x_\imath}$, ao tomarmos dois valores distintos da variável explicativa $x_\jmath$ e $x_{\jmath+1}$, obtemos:

$$OR = \frac{g(x_{j+1})}{g(x_j)} = \frac{e^{\beta_0+\beta_1 x_{j+1}}}{1+e^{\beta_0+\beta_1 x_j}} \tag{2.73}$$

Temos que,

$$ln(OR) = ln\left[\frac{g(x_{j+1})}{g(x_j)}\right] = ln[g(x_{j+1})] - ln[g(x_j)] = \beta_1(x_{j+1}-x_j) \tag{2.74}$$

Quando a diferença entre as variáveis explicativas é de uma unidade, temos:

$$ln(OR) = ln(e^{\beta_1}) = \beta_1 \tag{2.75}$$

A probabilidade do resultado ocorrer entre os indivíduos x_{j+1} em relação aos indivíduos x_j. Efetuando algumas análises, temos:

$$\beta_1 > 0 \Rightarrow OR > 1 \Rightarrow \pi(x_{j+1}) > \pi(x_j)$$
$$\beta_1 < 0 \Rightarrow OR < 1 \Rightarrow \pi(x_{j+1}) < \pi(x_j)$$

Estimativa do Desvio Padrão

No modelo de regressão logístico o desvio padrão dos estimadores é obtido invertendo a matriz de informação de Fisher. Invertendo a matriz obtemos as variâncias e covariâncias dos estimadores $\hat{\beta} = (\beta_0, \beta_1)$ isto é, calculando $I^{-1}(\hat{\beta})$.

O j-ésimo elemento da diagonal principal da matriz $I^{-1}(\hat{\beta})$ é a variância do estimador $\hat{\beta}_j$ denominada de $(\sigma^2)\hat{\beta}_j$. Os restantes elementos da matriz I^{-1} são as covariâncias entre $(\hat{\beta}_j, \hat{\beta}_u)$ com $j \neq u$.

Sendo assim, o desvio padrão é definido da seguinte forma:

$$DP(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2(\hat{\beta}_j)} \tag{2.76}$$

Intervalos de Confiança

A elaboração das estimativas do Intervalo de Confiança (IC) para os parâmetros têm por base a mesma teoria estatística que é utilizada para os testes de significância do modelo. Em particular, os intervalos de confiança para a inclinação e intercepto são baseados nos respetivos testes de Wald.

O IC ao nível de confiança $100(1-\alpha)\%$ para o parâmetro β_1 e para o intercepto β_0 , são:

$$IC_{\beta_1,1-\alpha} = [\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} DP(\hat{\beta}_1)] \tag{2.77}$$
$$IC_{\beta_0,1-\alpha} = [\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} DP(\hat{\beta}_0)] \tag{2.78}$$

Em que, $z_{1-\frac{\alpha}{2}}$ é o valor crítico da distribuição normal padrão correspondente $100(1-\frac{\alpha}{2})\%$.

A parte linear do modelo de regressão logística é denominada de logit, assim o estimador para o mesmo é dado pela seguinte expressão:

$$IC_{\hat{g}(x),1-\alpha} = [\hat{g}(x) \pm z_{1-\frac{\alpha}{2}} DP(\hat{g}(x))] \tag{2.79}$$

Sendo que, $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ é o estimador para logit e $DP(\hat{g}(x))$ é a raiz quadrada de $V\hat{a}r[\hat{g}(x)] = V\hat{a}r[\hat{\beta}_0] + x^2 V\hat{a}r[\hat{\beta}_1] + 2xC\hat{o}v(\hat{\beta}_0\hat{\beta}_1)$.

O estimador de logit e seu intervalo de confiança fornece o estimador dos valores ajustados. O intervalo de confiança dos valores ajustados é dado por:

$$IC_{\pi,1-\alpha} = \left[\frac{e^{\hat{g}(x) \pm z_{(1-\frac{\alpha}{2})} DP(\hat{g}(x))}}{1 + e^{\hat{g}(x) \pm z_{(1-\frac{\alpha}{2})} DP(\hat{g}(x))}} \right] \tag{2.80}$$

Pode-se também calcular o intervalo de confiança para OR, dados os limites inferior (β_I) e superior (β_S), pela seguinte expressão:

$$IC_{OR,1-\alpha} = \left[e^{\beta_I}, e^{\beta_S}\right] \tag{2.81}$$

2.3.2 Regressão Logística Múltipla

Tal como no modelo de regressão linear, podemos ajustar um modelo logístico para a variável resposta com mais de uma variável explicativa (covariáveis), que é denominado modelo de Regressão Logística Múltipla.

Seja um conjunto de p covariáveis $x_{1,\ldots,x_p}$, onde a probabilidade de ocorrência de um certo é representada pela definição $P = Y = \frac{1}{X} = \pi(X)$ e e a função de ligação ou logit é dado, respectivamente, por:

$$E[Y] = \pi(X) = \frac{e^{g(X)}}{1 + e^{g(X)}} \tag{2.82}$$

e

$$g(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{2.83}$$

Algumas variáveis independentes poderão ser discretas nominais e/ou ordinais, (tal como género, grupo etário, etc) e, nesses casos, estas variáveis não devem ser incluídas no modelo como se fossem escalões ou valores numéricos. Então, para contornar esta situação, devem ser utilizadas variáveis denominadas dummy. Estas variáveis representam a presença ou ausência de determinada característica.

Estimativa dos Parâmetros do Modelo

Será utilizado o método da máxima verossimilhança para obter as estimativas dos componentes do vetor, assim:

$$L(\beta_0, \beta_1, \ldots, \beta_p | (x_\imath, m_\imath, y_\imath)) = \sum_{\imath=1}^{n} \left[y_\imath g(X) - ln(1 + e^{g(X)})\right] \tag{2.84}$$

Onde $g(X)$ é a função de ligação e $(x_\imath), y_\imath$ são os dados observados.

Derivando a equação da máxima verossimilhança, igualando a zero e substituindo pelos estimadores dos parâmetros obtêm-se as seguintes equações

$$\sum_{\imath=1}^{n} y_\imath \left(1 + e^{g(X)}\right) - \sum_{\imath=1}^{n} e^{g(X)} = 0$$

$$\sum_{\imath=1}^{n} y_\imath x_\imath \left(1 + e^{g(X)}\right) - \sum_{\imath=1}^{n} x_\imath e^{g(X)} = 0$$

Aplicando o método iterativo dos mínimos quadrados ponderados é possível encontrar as raízes dessas equações que na verdade são as estimativas para os parâmetros do modelo de regressão logística multivariada.

Com isso, pode-se estimar as variâncias e covariâncias dos coeficientes pela estimação de máxima verossimilhança. Os estimadores são obtidos a partir da matriz de covariância de segundas derivadas parciais da função log de verossimilhança.

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \tag{2.85}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_i} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \tag{2.86}$$

Onde $j, i = 0, 1, ..., p$.

Seja a matriz $(p + 1)x(p + 1)$ que contém os termos negativos das derivadas parciais e denotada por $I(\beta)$, denominada de Matriz de Informação de Fisher. As variâncias e covariâncias dos coeficientes estimados são obtidas a partir da inversão da matriz $I(\beta)$.

A Matriz de Informação de Fisher estimada pode ser obtida por:

$$\begin{bmatrix} -\frac{\partial^2 l(\beta;x)}{\partial \beta_1^2} & \cdots & -\frac{\partial^2 l(\beta;x)}{\partial \beta_1 \beta_k} \\ \vdots & \ddots & \vdots \\ -\frac{\partial^2 l(\beta;x)}{\partial \beta_k \beta_1} & \cdots & -\frac{\partial^2 l(\beta;x)}{\partial \beta_k^2} \end{bmatrix}$$

O estimador da máxima verossimilhança $\beta_1, \beta_2, ..., \beta_p$ são obtidas como soluções da equação:

$$\frac{\partial l(\beta;x)}{\partial \beta_i} \tag{2.87}$$

Testes de significância

Após o ajuste do modelo ao conjunto de dados, segue o teste de significância das variáveis incluídas no modelo. A significância estatística dos resultados obtidos na análise de regressão deve ser estabelecida antes do uso desses resultados numa previsão. O propósito dos testes de significância estatística é determinar a confiança que pode ser depositada nos resultados da regressão e a sua aplicabilidade na população de valores possíveis.

O teste da razão de verossimilhança para a significância dos p coeficientes das variáveis independentes do modelo é realizado da mesma forma que no modelo de regressão logística simples. A estatística teste G é dada por:

$$D = -2ln(L_s) + 2ln(L_c) \tag{2.88}$$

Em que, L_s é a verossimilhança do modelo sem a covariável e L_c é a verossimilhança do modelo com a covariável.

No caso da regressão múltipla, temos o interesse em saber se pelo menos uma variável é significativa para o modelo. Sob a hipótese nula, os p coeficientes são iguais a zero , assim, a estatística G tem distribuição Qui-Quadrado com p graus de liberdade, nesse caso L_c é a verossimilhança do modelo com as p variáveis explicativas e L_s é a verossimilhança do modelo apenas com o intercepto.

O teste de Wald tem como objetivo testar a significância de cada coeficiente dentro do modelo obtido, ou seja se o coeficiente é diferente de zero. Deste modo, o teste de Wald averigua se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente. Assim, pretende-se testar:

$$H_0 : \beta_J = 0$$
$$vs$$
$$H_1 : \beta_J \neq 0$$

A estatística de teste é dada por:

$$W_J = \frac{\hat{\beta}_J}{\hat{\sigma}(\hat{\beta}_J)} \tag{2.89}$$

O teste de Wald também pode ser obtido pela multiplicação dos seguintes vetores:

$$W = \hat{\beta^T}(X^T V X)\hat{\beta} \tag{2.90}$$

Com distribuição Qui-quadrado e $p+1$ grau de liberdade sob a hipótese que cada um dos $p+1$ coeficientes é igual a zero.

Intervalos de Confiança

O IC para um parâmetro β_J é baseado no teste de Wald. O IC de $1000(1-\alpha)$ para o parâmetro β_J, é dada por:

$$IC_{(\beta_J,1-\alpha)} = \left[\hat{\beta}_J \pm z_{1-\alpha}\sigma(\hat{\beta}_J)\right] \tag{2.91}$$

O IC para o logit é:

$$IC_{(\hat{g}(x),1-\alpha)} = \left[\hat{g}(x) \pm z_{1-\frac{\alpha}{2}}\sigma(\hat{g}(x))\right] \tag{2.92}$$

O estimador dos valores ajustados é calculado a partir do estimador do logit e o intervalo de confiança intervalo de confiança dos valores ajustados é dado por:

$$IC_{(\pi,1-\alpha)} = \left[\frac{e^{\hat{g}(x)\pm z_{1-\frac{\alpha}{2}}\sigma(\hat{g}(x))}}{1+e^{\hat{g}(x)\pm z_{1-\frac{\alpha}{2}}\sigma(\hat{g}(x))}}\right] \tag{2.93}$$

o IC para a OR é:

$$IC_{(OR,1-\alpha)} = \left[e^{\beta_I}, e^{\beta_S}\right] \tag{2.94}$$

Análise do modelo

Existem dois tipos de resíduos possíveis que poderão ser utilizados para avaliar a qualidade do ajustamento: os resíduos de Pearson e os resíduos da Deviance.

O resíduo de Pearson para o j-ésimo elemento é definido por:

$$r(y_i, \hat{\pi}_j) = r_j = \frac{y_i - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}}, \quad j = 1, 2, ..., n \tag{2.95}$$

A estatística de teste baseada nos resíduos de Pearson é designada por estatística de Qui-Quadrado de Pearson e é calculada da seguinte forma:

$$x^2 = \sum_{j=1}^{n} r(y_i, \hat{\pi}_j)^2 \tag{2.96}$$

A soma dos quadrados dos resíduos da regressão linear, no modelo logístico é denominada deviance (D) e o resíduo de D é definido da seguinte forma:

$$d(y_{j,\hat{\pi}_j}) = \pm \left[2\left(y_j ln\left(\frac{y_j}{\hat{\pi}_j}\right) + (1-y_j) ln\left(\frac{1-y_j}{1-\hat{\pi}_j}\right)\right)\right]^{1/2} \tag{2.97}$$

A estatística de teste é dado por:

$$D = \sum_{j}^{n} d(y_j, \hat{\pi}_j)^2 \tag{2.98}$$

Curva ROC

A análise ROC (Receiver Operating Characteristic) teve origem na teoria de decisão estatística e foi desenvolvida entre 1950 e 1960 para avaliar a deteção de sinais em radar e na psicologia sensorial. A potencial utilidade da análise ROC em avaliar diagnósticos médicos, foi desde então utilizada por vários autores e, subsequentemente, foi aplicada com sucesso a uma grande variedade de testes de diagnóstico e em particular no diagnóstico de imagem médica. Consequentemente, as várias aplicações da análise ROC à medicina estimularam o desenvolvimento de metodologias de análise estatística dos dados ROC [2].

Quando a variável resposta é binária é necessário determinar uma regra de predição, dado que a probabilidade estimada $\hat{\pi}$ está compreendida entre 0 e 1. Poderá ser intuitivo supor que se $\hat{\pi}_i$ apresentar um valor de próximo de 1 deveremos considerar que $\hat{Y}_i = 1$ e se $\hat{\pi}_i$ corresponder um valor pequeno ou próximo de zero, deveremos considerar que $\hat{Y}_i = 0$.

Para determinar o ponto que para os valores acima dele o indivíduo é classificado como evento ($\hat{Y}_i = 1$).e valores abaixo dele o indivíduo é classificado como não evento $\hat{Y}_i = 0$. Esse ponto é denominado como ponto de corte.

Uma forma bastante utilizada para determinar o ponto de corte é através da Curva ROC.

Uma curva ROC é a representação gráfica dos pares sensibilidade ou FVP (ordenadas) e especificidade ou FFP (abcissas), resultantes da variação do valor de corte ao longo de um eixo de decisão, x, a representação gráfica assim resultante é designada por curva ROC no plano unitário. Com efeito, uma curva ROC é uma descrição empírica da capacidade do sistema de diagnóstico poder discriminar entre dois estados num universo, onde cada ponto da curva representa um compromisso diferente entre a FVP e a FFP que pode ser adquirido pela adoção de um diferente valor de corte de anormalidade ou nível crítico de confiança no processo de decisão [2].

A escolha do ponto de corte deve basear-se numa combinação ótima de sensibilidade e de especificidade, pois parte-se do pressuposto que classificar o indivíduo como evento dado que ele é não evento (falso positivo) e classificar o indivíduo como não evento dado que ele é evento (falso negativo) acarreta prejuízos para um investigador. Pela análise da curva ROC, escolhe-se o ponto de corte referente a combinação da sensibilidade e 1-especificidade que mais se aproxima do canto superior esquerdo do gráfico [28].

Foram criados parâmetros numéricos cuja denominação é a seguinte: capacidade de precisão, sensibilidade, especificidade, verdadeiro preditivo positivo e verdadeiro preditivo negativo.

Capítulo 3

Análise Exploratoria dos Dados

Se os resultados obtidos com base em dados reais são importantes, não menos importante é a informação das características dessas bases de dados, pois a validade dos resultados está dependente das condições em que os mesmos são obtidos. A informação contida neste capítulo é relevante não só para validar os resultados, mas também para se perceber alguns procedimentos nas metodologias desenvolvidas.

3.1 Caracterização do Estudo

Os dados reais, disponibilizados para os trabalhos desenvolvidos nesta dissertação, dizem respeito aos registos dos pacientes internados com COVID-19, do Hospital Universitário Agostinho Neto (HUAN). O estudo foi realizado com a autorização do Conselho de Administração do HUAN.

Para este estudo foram selecionados os dados referentes ao período de maio de 2020 a novembro de 2021. Os dados estavam agrupados pelas ocorrências de cada mês, alocados em uma folha de Excel, que continha 527 registros de internamentos dispostos em 9 colunas.

Algumas dessas colunas apresentavam informações repetidas, bem como falta de informações relevantes o que poderia comprometer um pouco os resultados do estudo. Portanto, o primeiro passo foi uma reestruturação da base de dados para o processo de seleção das variáveis.

3.2 Manipulação dos Dados

A manipulação de dados é essencial na ciência de dados, consiste em detectar e corrigir ou remover valores incondizentes com uma base de dados que está sendo tratada.

Ao analisar os dados disponibilizados, encontramos alguns valores das variaveis em falta. Foram aplicadas técnicas de validação de dados e identificação de outliers.

A identificação e limpeza são complementares, então a extração dos dados é tratada diretamente na identificação de alguma informação. Após isso, ficamos com uma amostra

de 502 internados.

3.3 Variáveis Analisadas

Foi feita uma conversão de algumas das variáveis, que estavam na forma texto e deveriam ser representadas em variáveis categóricas. Neste sentido, foram agrupadas e analisadas as seguintes variáveis:

- Gênero: foram agrupadas em 2 categorias {Feminino e Masculino};
- Faixa etária: Foram agrupadas em 5 categorias $\{[0-20], [21-41], [42-62], [63-83], [84-101]\}$;
- Dias: calculada a partir da data de internamento e data de destino;
- Óbitos: codificado com variável binárias, com 2 categorias {0 - Altas e 1 - Obitos};
- Doenças associadas: foram todas criadas como variáveis binárias com 2 categorias {0 - não e 1 - Sim}. No total de 34 doenças associadas;
- Oxigênio: foram agrupadas em 2 categorias {Sim e Não};
- Ventilação: foram agrupadas em 2 categorias {Sim e Não}.

No primeiro estudo, a variável dependente escolhida foi número de dias de internamento, que representava tal como o nome indica o número de dias de internamento. Como variáveis independentes foram escolhidas as outras variáveis como (género, faixa etária, óbitos, doenças associada, oxigênio, ventilação) No segundo estudo a variável dependente escolhida foi o óbito, que representa a associação da ocorrência de morte. Como variáveis independentes foram escolhidas as outras variáveis como (gênero, faixa etária, doenças associadas, oxigênio, ventilação).

3.4 Análise Exploratória

Antes de iniciar a definição dos modelos é necessário proceder à caracterização dos dados.

Sendo a amostra de 502 internamento, verifica-se que, o sexo masculino é o mais representativo (57,8%), sendo a maioria com idade compreendida entre 42 a 62 anos.

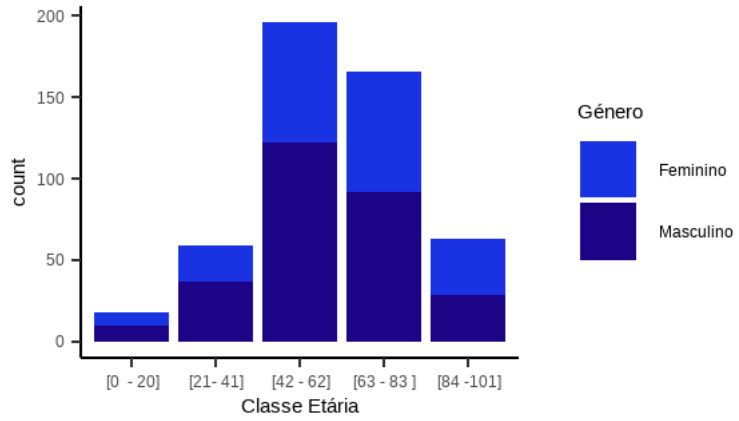

Fig. 3.1: Classe Etária por Género

Avaliando o número de dias por classe etária e sexo, verifica-se que, as classes apresentam níveis médios diferentes, nota-se uma certa variabilidade. Há 4 indivíduos discrepantes.

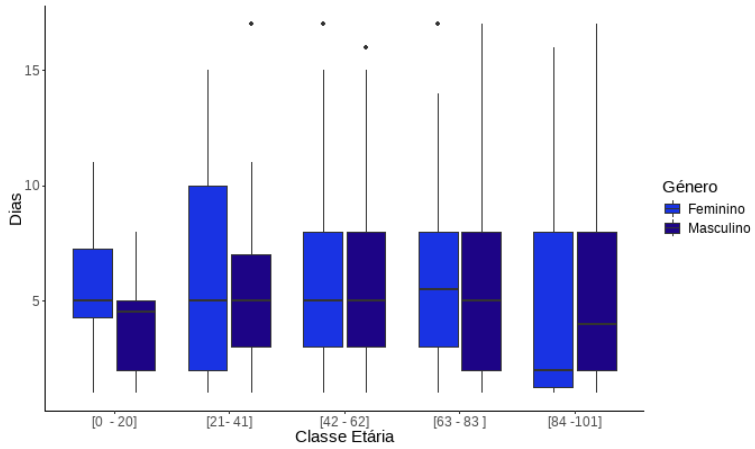

Fig. 3.2: Dias por Classe Etária e Género

Analisando o óbito, predomina o sexo masculino com 56,7% e a faixa etaria [63 - 83] com 42,3%.

Tab. 3.1: Óbitos por Género Classe e Etária

Óbitos	**Gênero**		**Classe Etária**					**Total**
	F	**M**	**[0 -20]**	**[21 - 41]**	**[42 - 62]**	**[63 - 83]**	**[84 - 101]**	
Não	170	235	16	53	174	125	37	405
Sim	42	55	2	6	22	41	26	97

Relativamente às doenças associadas, a maioria com pneumonia, hipertensão arterial (HTA), diabetes mellitus (DM), síndrome e sepsia de aflição respiratória aguda (ARDS).

Tab. 3.2: Doenças Associadas por Genero Classe e Etária

Doenças	**Género**		**Classe Etária**					**Total**
	F	**M**	**[0 -20]**	**[21 - 41]**	**[42 - 62]**	**[63 - 83]**	**[84 - 101]**	
Obesidade	17	19		1	17	13	5	36
Pneumonia	134	190	5	29	117	129	44	324
Bronquite	1	1		2				2
EAP	2	3		2	1	2		5
TEFBilateral		1			1			1
Hematoma		1				1		1
ARDS	36	53		7	40	31	11	89
Abdómen	1			1				1
Abscesso		1				1		1
Alcoolismo	2	9	1	1	6	2	1	11
Anemia	1	2		2	1			3
Asma	2	3			2	3		5
Ataxia		1		1				1
AVC	14	24		2	10	20	6	38
Câncer	1	3			1	2	1	4
Cardiopatia		5		2		1	2	5
Celulite		1		1				1
Diabete	2	2			1	3		4
Cirrose	1	1			2			2
Colites	1	1			2			2
Demência		1				1		1
DM	64	79		6	48	71	18	143
IRC	17	21		1	16	15	6	38
HTA	82	112		6	65	89	34	194
IUT	2				1		1	2
DPOC	1	3			2		2	4
Etilismo		8		3	5			8
Epilepsia	1			1				1
Esquizofrenia		2			1	1		2
FACRVR	1						1	1
Gravidez	6			5	1			6
IRA	12	13		3	7	9	6	25
HIV	3	5		3	5			8
ICC	14	13		1	7	9	10	27

No total dos 10 internados que foram submetidos a oxigênio, por sexo, 60% é masculino e 40% feminino e por classe etária, 50% com idade entre 42 a 62 anos.

Foram 45 internados submetidos a ventilação, sendo 5,8% do sexo masculino e 3,4%

com idade entre 63 a 83 anos.

Tab. 3.3: Oxigênio e Ventilação por Genero Classe e Etária

Variáveis	**Género**		**Classe Etária**					**Total**
	F	**M**	**[0 -20]**	**[21 - 41]**	**[42 - 62]**	**[63 - 83]**	**[84 - 101]**	
Oxigenio	6	4	0	1	5	4	0	10
Ventilação	16	29	2	4	15	17	7	45

Foi calculado o Coeficiente de Correlação Linear de Pearson, para as variáveis dias e doenças associadas, e para óbitos e doenças associadas. Verificou-se que, as variáveis evidenciam na sua totalidade uma relação muito fraca. Entretanto, para as variáveis com coeficiente fraco, realizou-se um teste de hipóteses, para avaliar se esses resultados são significativos.

Entre dias e doenças associadas, como os valores dos p-value são pequenos, conclui-se que os valores dos Coeficiente de Correlação Linear tem alguma significância estatística.

Tab. 3.4: Correlação entre as variáveis Dias e Doenças Associadas

Doenças	**Correlação**	**Valor p-value**
ARDS	0.2192923	6.992e-07
ITU	0.1040047	0.01977
IRA	0.1270751	0.00435

Entre óbitos e doenças associadas, os valores dos p-value também são pequenos, conclui-se que os valores dos Coeficiente de Correlação Linear tem alguma significância estatística.

Tab. 3.5: Correlação entre as variáveis Óbitos e Doenças Associadas

Doenças	**Correlação**	**Valor p-value**
EAP	0.1541428	0.0005288
ARDS	0.2483948	1.699e-08
Diabetes	0.1263816	0.004569
IRA	0.2358594	8.944e-08
ICC	0.1964256	9.281e-06

Capítulo 4

Modelos de Regressão: Aplicação e Interpretação dos Resultados

Para a obtenção dos resultados práticos, será utilizado o Software R, que é uma linguagem de programação multi-paradigma orientada a objetos, programação funcional, dinâmica, fracamente tipada, voltada à manipulação, análise e visualização de dados. Foi criado originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da Universidade de Auckland, Nova Zelândia.

Serão apresentados dois estudos. No primeiro estudo foi utilizado o Modelo de Regressão de Poisson, enquanto que no segundo considerámos o Modelo de Regressão Logística.

4.1 Análise de Dias de Internamento

Nesta seção será utilizada a Regressão de Poisson para estudar os dias de internamento.

Com vista a perceber se a classe etária e género influencia o número de dias de internamento, foi realizada uma análise de simples, que representa a relação entre a variável dependente (Dias), e as variáveis independentes.

A classe etária $[63 - 83]$ influencia significativamente os dias de internamento, com um valor de significância de 0,049 $(< 0,05)$.

Tab. 4.1: Modelo de Regressão de Poisson - Dias de Internamento por Classe Etária

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	1.5285	0.1098	13.925	<2e-16 ***
Classe etária[21- 41]	0.1931	0.1228	1.572	0.1158
Classe etária[42 - 62	0.2100	0.1138	1.846	0.0649 .
Classe etária[63 - 83	0.2254	0.1144	1.970	0.0488 *
Classe etária[84 - 101]	0.1091	0.1230	0.887	0.3750

É importante avaliar a sua adequação aos dados através da análise de resíduos. Se

os resíduos forem aleatórios e tiverem uma distribuição normal então é válido considerar adequado o modelo obtido.

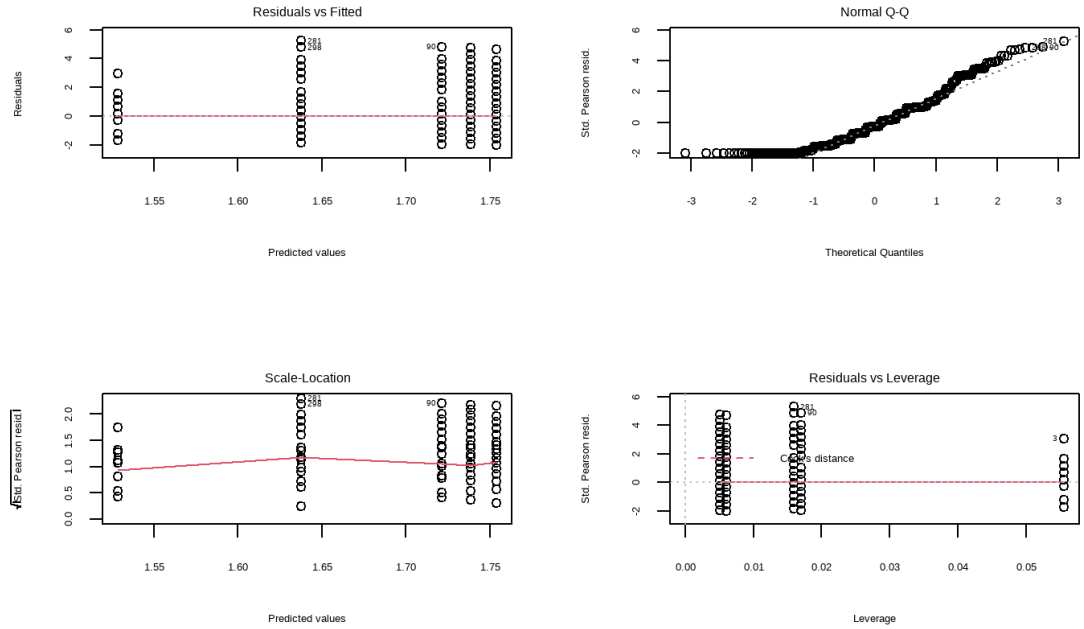

Fig. 4.1: Análise de Resíduos - Dias de Internamento por Classe Etária

De acordo com a estatística de teste de Durbin-Watson obtida (DW = 1.8717) não rejeitamos a independência dos resíduos para um nível de significância de 0.05.

Para o teste de Breush-Pagan o p-value obtido (0,2977) indica-nos que não devemos rejeitar a hipótese nula do teste de hipóteses. Assim, conclui-se que os resíduos são independentes.

O género não influencia significativamente os dias de internamento, porque o valor de significância de 0,183 é maior que 0,05.

Tab. 4.2: Modelo de Regressão de Poisson - Dias de Internamento por Género

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	1.75166	0.02861	61.233	<2e-16 ***
Género Masculino	-0.05068	0.03805	-1.332	0.183

Avaliando a adequação dos dados através dos gráficos da análise de resíduos.

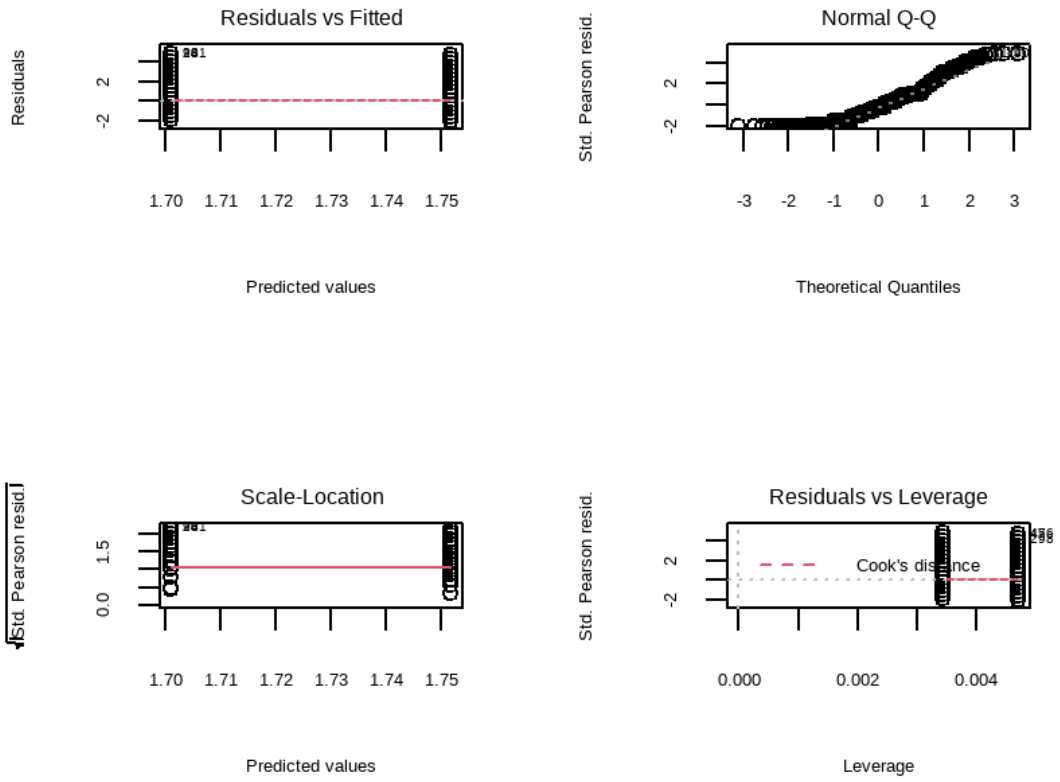

Fig. 4.2: Análise de Resíduos - Dias de Internamento por Género

De acordo com a estatística de teste de Durbin-Watson obtida (DW = 1.8734) e do teste de Breush-Pagan o p-value obtido (0,6765) não rejeitamos a hipótese nula, os resíduos são independentes.

Então, os modelos obtidos podem ser considerados estatisticamente válidos para a realização de inferência estatística.

Passamos agora a considerar o modelo de Possion estimado através de todas as variáveis. O que pretendemos averiguar é de que forma as variáveis independentes influenciam a variável dependente (dias de internamento).

Para chegar ao modelo adicionou-se todas as variáveis de interesse da base de dados, e foram retiradas as que não apresentaram significância pela análise do valor-p.

É possível perceber que, ITU e oxigênio são os coeficientes com maior valor, mantendo tudo constante, o aumento de 1 paciente com ITU, aumenta o número de dias de internamento. As variáveis explicativas influenciam significativamente os dias de internamento ($p<0,05$).

Tab. 4.3: Modelo de Regressão de Poisson - Dias de Internamento

	Estimate	**Std. Error**	**z value**	**p -value**
(Intercept)	1.60329	0.02321	69.064	<2e-16***
Obesidade	-0.22889	0.07803	-2,933	0.003355**
Bronquite	-1.68023	0.60888	-2,760	0.005788**
ARDS	0.30302	0.04756	6.372	1.87e-10***
Asma	0.48241	0.19400	2.487	0.012896*
Cirose	-1.33487	0.42071	-3.173	0.001509**
ITU	0.88162	0.20544	4.291	1.78e-05***
IRA	0.28928	0.07634	3.789	0.000151***
HIV	0.38530	0.12827	3.004	0.002665**
Oxigénio Sim	0.68515	0.10306	6.648	2.97e-11***
Ventilação Sim	0.25347	0.06010	4.217	2.47e-05***

Em seguida procede-se à obtenção e análise de valores associados aos parâmetros do modelo, através do Intervalo de Confiança (IC).

Os valores referentes à variância dos parâmetros, podem ser determinados através da matriz de covariância dos mesmos parâmetros. Assim, a variância estimada para os parâmetros é, aproximadamente: 0.001, 0.006, 0.371, 0.002, 0.038, 0.177, 0.042, 0.006, 0.016, 0.011 e 0.004.

Tab. 4.4: IC(95%) - Modelo de Regressão de Poisson

	2.5%	**97.5%**
(Intercept)	1.55747231	1.64847512
Obesidade	-0.38513246	-0.07906332
Bronquite	-3.11277284	-0.63921442
ARDS	0.20912291	0.39557539
Asma	0.07723913	0.84072346
Cirrose	-2.27623141	-0.59806841
ITU	0.45071202	1.25961125
IRA	0.13658670	0.43599194
HIV	0.12366283	0.62730770
Oxigénio Sim	0.47695121	0.88134416
Ventilação Sim	0.13423691	0.36989563

Tab. 4.5: ANOVA - Modelo de Regressão de Poisson

	Deviance Resid.	**Df**	**Resid. Dev**
NULL		501	1341.0
Obesidade	2.340	500	1338.7
Bronquite	8.651	499	1330.0
ARDS	63.795	498	1266.2
Asma	3.747	497	1262.5
Cirrose	4.595	496	1257.9
ITU	12.673	495	1245.2
IRA	17.122	494	1228.1
HIV	8.163	493	1219.9
Oxigénio Sim	36.635	492	1183.3
Ventilação Sim	16.890	491	1166.4

É importante avaliar a adequação dos dados através da análise de resíduos. Podemos afirmar que os resíduos e valores ajustados apresentam uma distribuição algo padronizada em torno da reta horizontal que representa o erro zero. Os resíduos seguem uma distribuição normal devido ao seu alinhamento com a reta representada. Esta reta está relacionada com a frequência acumulada dos valores de uma distribuição normal para as abcissas.

O gráfico Scale - Location apresenta os valores ajustados contra a raiz dos valores absolutos dos resíduos. Os pontos não apresentam nenhum padrão na sua distribuição.

O gráfico Residuals vs Leverange, apresenta a distância de Cook, em que distâncias curtas sugerem que a remoção da observação correspondente terá pouco efeito no modelo linear e, distâncias superiores a um podem sugerir a presença de um outlier.

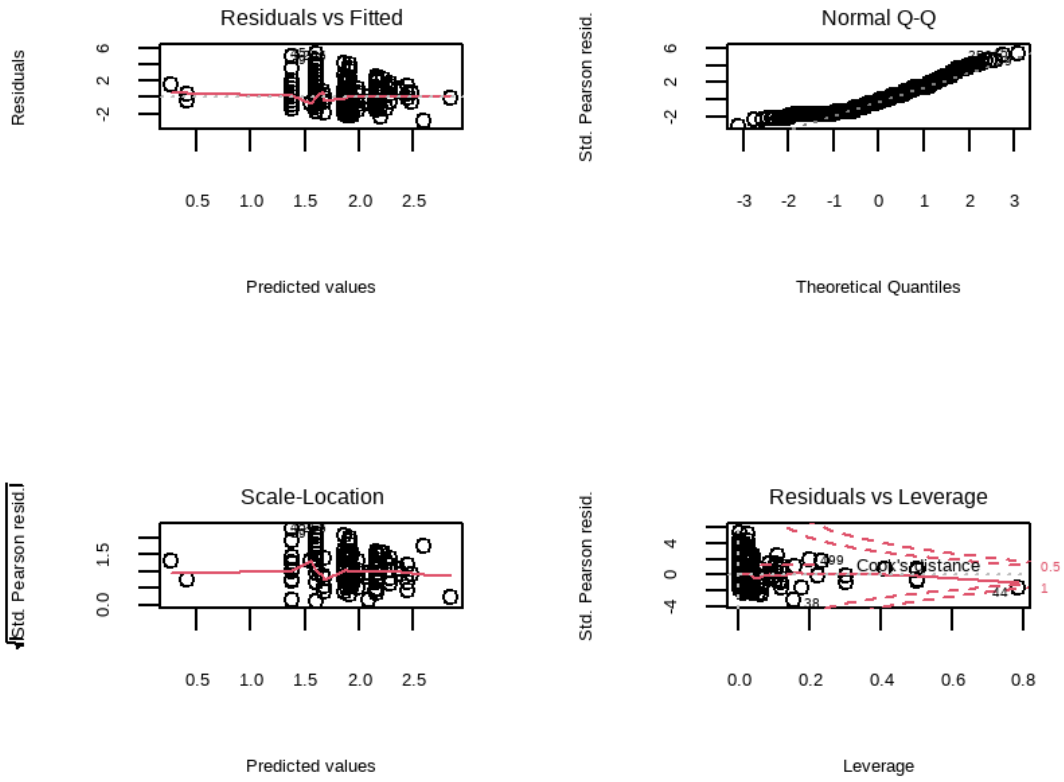

Fig. 4.3: Análise de Resíduos - Moelodo de Regressão de Poisson

Para testar a homocedasticidade dos resíduos foi utilizado o teste de Breusch-Pagan. O p-value obtido (0,6608) para um nível de confiança de 0,05 indica-nos que não devemos rejeitar a hipótese nula do teste de hipóteses. Assim, conclui-se que os resíduos são independentes.

Logo, o modelo apresentado, cumpre os pressupostos necessários para a validação estatística do modelo e para a realização de inferência estatisticamente válida.

4.2 Análise de Óbitos

Nesta secção pretende-se ajustar um modelo de Regressão Logística com a variável resposta dependente binária (óbitos), tratando os grupos de interesse com valores de 0 e 1, selecionando o melhor modelo de ajustamento.

Foi aplicada a Regressão Logística ao modelo completo, no sentido de selecionar as variáveis com nível de significância $< 0{,}05$.

Após isso, foi gerado um novo modelo (Model 1), sendo os valores estimados dos coeficientes presentes, pode-se concluir que todos (com exceção do alcoolismo) são estatisticamente significativos para o modelo, os p-value são inferiores a um nível de significância de 0,05, logo a hipótese nula que está subjacente à criação de um modelo de regressão, que é definida pela falta de preponderância das variáveis explicativas no modelo, em relação a todas as variáveis (completo), não é aceite.

Tab. 4.6: Modelo de Regressão Logística - Model 1

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-2.4022	0.1842	-13.040	<2e-16 ***
ARDS	0.9726	0.3213	3.027	0.002467**
Alcoolismo	1.3830	0.7309	1.892	0.058478.
Diabete	3.3028	1.1810	2.797	0.005164**
IRA	1.8566	0.5023	3.696	0.000219***
ICC	1.5964	0.4895	3.261	0.001109**
Ventilação Sim	3.0961	0.4204	7.365	1.77e-13***

Sendo o alcoolismo não significativo (<0,05) para o modelo, foi necessário proceder à análise de um outro modelo (Model 2), sendo os valores estimados dos coeficientes presentes, são todas estatisticamente significativos. Os resultados são ilustrados na tabela:

Tab. 4.7: Modelo de Regressão Logística - Model 2

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-2.3574	0.1804	-13.068	<2e-16 ***
ARDS	0.9386	0.3185	2.947	0.00321**
Diabete	3.2634	1.1797	2.766	0.00567**
IRA	2.0186	0.4877	4.139	3.49e-05***
ICC	1.5513	0.4889	3.173	0.00151**
Ventilação Sim	3.0932	0.4193	7.377	1.62e-13***

Posteriormente compararam-se os dois modelos com o propósito de escolher o melhor que estimasse a variável resposta, através do Teste da Verosimilhança.

Dos resultados, observa-se que p-value (0,07077) é superior a um nível de significância de 0,05. Deve-se rejeitar a hipótese de que o Model 2 é melhor que o Model 1. Apesar do alcoolismo não ter um efeito significativo na variável dependente, a sua inclusão no modelo contribui para uma redução seignificativa da Deviance.

Tab. 4.8: Teste de Verosimilhança - Model 1 e Model 2

	Resid. DF	**Resid. Dev**	**Df**	**Deviance**	**Pr(>Chi)**
Model 1	495	358.81			
Model 2	496	362.07	-1	-3.265	0.07077

Model 1: Óbitos ARDS + Alcoolismo + Diabete + IRA + ICC + Ventilação
Model 2: Óbitos ARDS + Diabete + IRA + ICC + Ventilação

Uma outra forma de avaliar os dois modelos é observar os valores do Critério de Informação de Akaike (AIC) para cada um dos modelos. Como para o Model 1 se observou um AIC de 372,81, que é inferior ao do Model 2 (374,07), conclui-se que se rejeita a hipótese, que o Model 2 é significativamente melhor que o Model 1.

Tab. 4.9: AIC - Model 1 e Model 2

	df	**AIC**
Model 1	7	372.8061
Model 2	6	374.0712

Concluímos que, o Model 1 é o melhor modelo ajustado, neste sentido estudou-se esse modelo.

Procede-se à obtenção e análise de valores associados aos parâmetros do modelo, através do Intervalo de Confiança (IC) , representado na tabela.

Tab. 4.10: IC (95%) - Model 1

	2.5%	**97.5%**
(Intercept)	-2.7812518	-2.057115
ARDS	0.3312753	1.595720
Alcolismo	-0.1240286	2.786042
Diabete	1.1855763	6.336616
IRA	0.8600970	2.849054
ICC	0.6081523	2.545068
Ventilação [Sim]	2.3074412	3.970920

Utilizando o teste de análise de variância ANOVA pode-se observar a variância com o modelo nulo (501) e à medida que as variáveis explicativas foram sendo incluídas, estas reduziram a variância do modelo para 495, contribuindo para o modelo.

Tab. 4.11: ANOVA - Model 1

	Df	**Deviance Resid.**	**Df**	**Resid. Dev**
NULL	501	492.83		
ARDS	1	26.874	500	465.96
Alcolismo	1	5.109	499	460.85
Diabete	1	6.010	498	454.84
IRA	1	16.614	497	438.22
ICC	1	13.122	496	425.10
Ventilação	1	66.296	495	358.81

Para obter os Odds Ratio (OR) vamos exponenciar os valores obtidos. Na tabela seguinte podem observar-se os valores exponenciados das estimativas para os parâmetros e os respectivos I.C´s.

Tab. 4.12: OR - Model 1

	OR	**2.5%**	**97.5%**
(Intercept)	0.090515	0.061961	0.1278
ARDS	2.644929	1.392743	4.9319
Alcolismo	3.986763	0.883355	16.2167
Diabete	27.187725	3.272572	564.8816
IRA	6.401768	2.363390	17.2714
ICC	4.935389	1.837034	12.7441
Ventilação [Sim]	22.112608	10.048679	53.0333

Para testar a adequabilidade do modelo, foram realizadas análise e testes sobre resíduos de Pearson, Deviance e outros.

A estatística de teste para a qualidade de ajuste a partir dos resíduos de Pearson, o valor do p-value determinado foi aproximadamente 0,55. Assim, não se rejeita a hipótese nula de adequação do modelo.

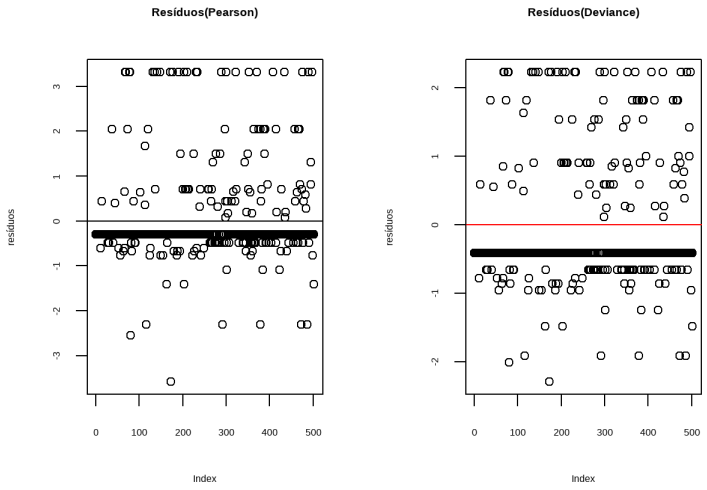

Fig. 4.4: Resíduos de Pearson e Deviance - Model 1

Podemos também identificar os pontos influentes, quando temos valores em que os resíduos ultrapassam 2 ou -2 possivelmente são valores influentes.

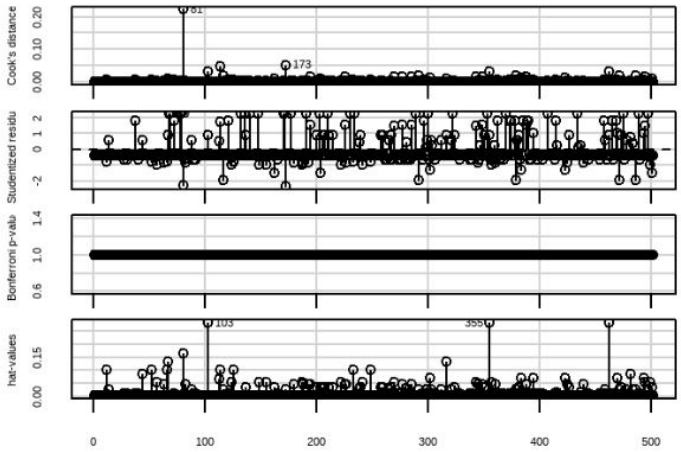

Fig. 4.5: Pontos Influentes - Model 1

Nesse gráfico, mostra as mesmas informações anteriores, mas podemos ver aqui quais são os valores que estão se distanciando mais dos limites.

Avaliando cada um dos pontos que ultrapassam os limites 2 e -2, verificamos que os pontos não são influentes, ou seja, ao retirar os pontos não muda muito o formato do modelo.

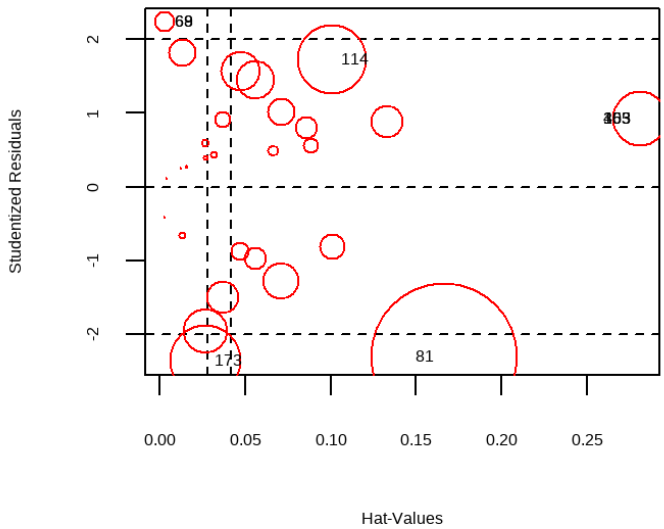

Fig. 4.6: Pontos Influentes - Model 1

Verificando a presença da multicolinearidade, observa-se que os valores são abaixo de

5.

Tab. 4.13: Multicolinearidade - Model 1

ARDS	Alcoolismo	Diabete	IRA	ICC	Ventilação
1.015471	1.030761	1.006518	1.026316	1.013548	1.017642

Os gráficos de cada variável, ilustram os valores de probabilidade de acordo com cada uma das variáveis.

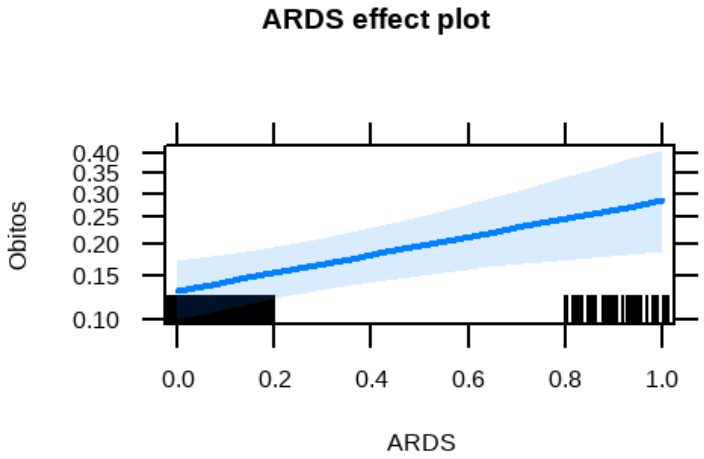

Fig. 4.7: Efeito:ARDS - Model 1

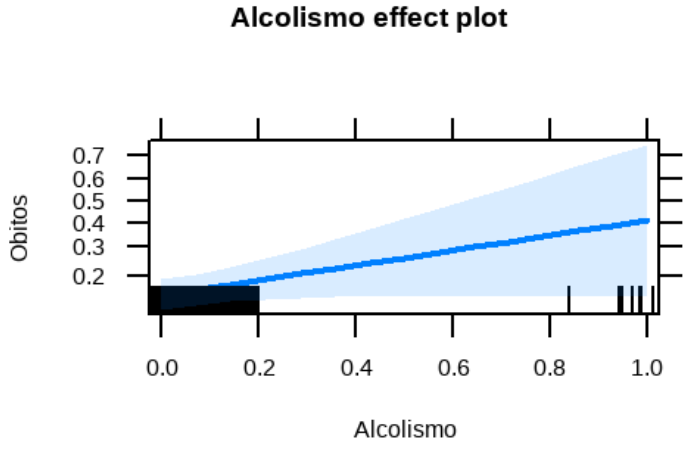

Fig. 4.8: Efeito:Alcoolismo - Model 1

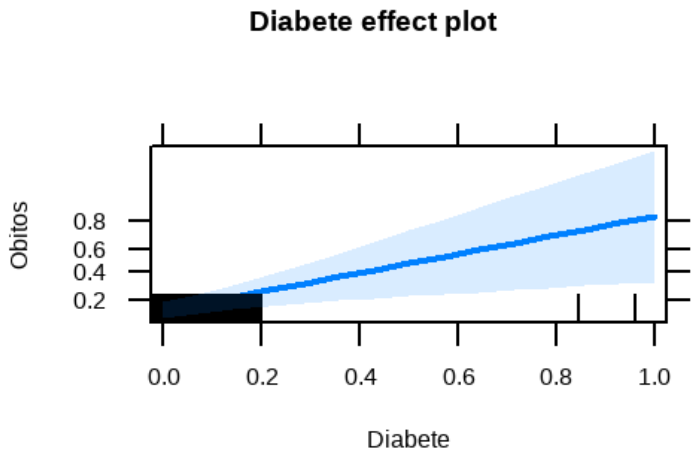

Fig. 4.9: Efeito:Diabete - Model 1

Fig. 4.10: Efeito:Ventilação - Model 1

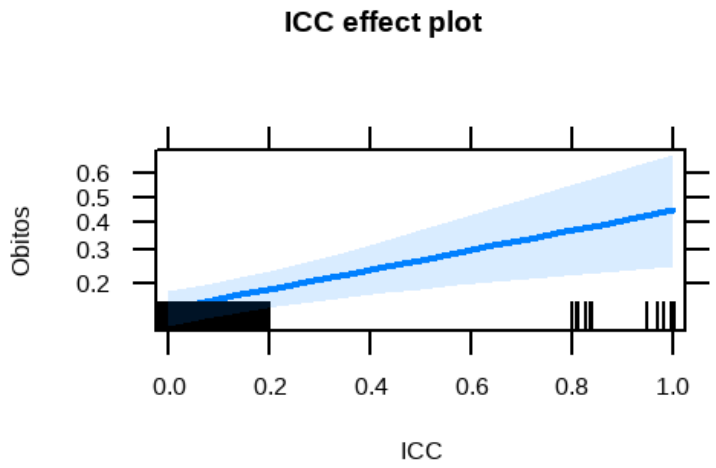

Fig. 4.11: Efeito - Model 1

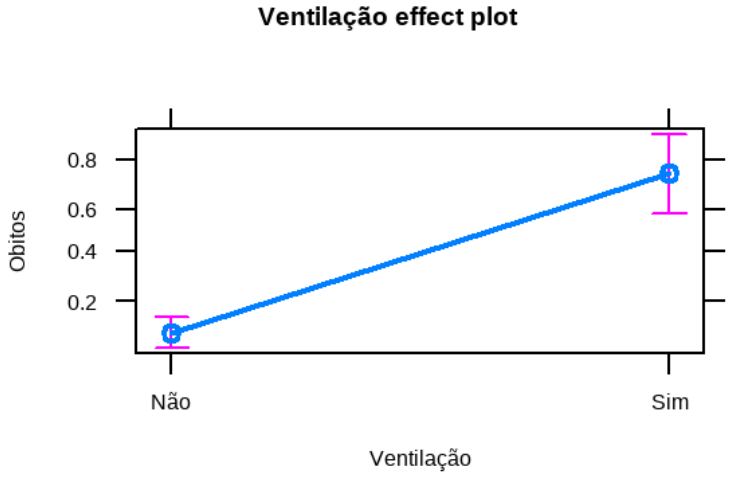

Fig. 4.12: Efeito:IRA - Model 1

O valor obtido para a AUC (Area Under The Curve) é de 0,803. Isto significa que a capacidade de predição do modelo pode ser considerada excelente, pois é um valor acima de 0,70, é fiável quanto à precisão na predictabilidade.

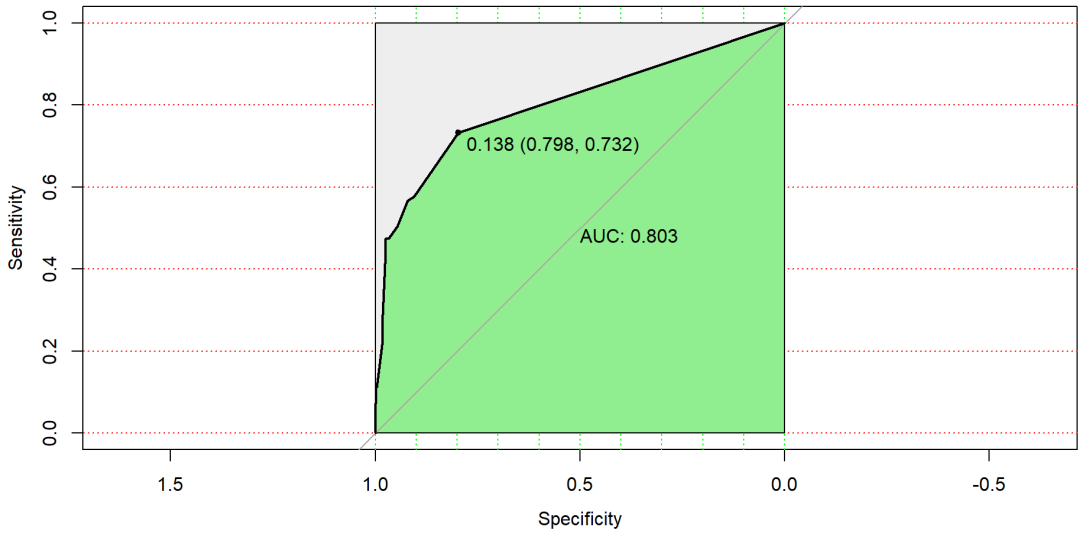

Fig. 4.13: Curva ROC - Model 1

Considerando os resultados obtidos, quer nos testes para a significância dos coeficientes, quer nos testes sobre a qualidade de ajuste do modelo ou ainda a área AUC, deve-se considerar que este modelo é estatisticamente válido para modelar a relação entre as variáveis e também para procedimentos associados à inferência estatística.

Capítulo 5

Conclusão

A análise de dados, tem como finalidade inspecionar, limpar, transformar e modelar os dados objetivando extrair informações úteis, que podem ser utilizados no processo de apoio e tomada de decisão. Devido a esses importantes aspectos, esta área está em franco desenvolvimento mediante o surgimento de novas técnicas e domínios de aplicações.

Em relação a este trabalho os objetivos propostos foram atingidos. O desenvolvimento deste estudo contribui para estudar o número de dias de internamento, em função de alguns dados biométricos e de identificar os fatores associados à morte dos pacientes internados.

Nesta dissertação foram apresentadas e descritas de forma sucinta as principais definições e propriedades de modelos de regressão de Poisson, modelos de regressão linear simples e múltipla e de modelos de regressão logística simples e múltipla, bem como definições de algumas estatísticas de teste associadas aos modelos de regressão. No capítulo da análise dos resultados procedeu-se à aplicação de modelos de regressão Poisson e logística a situações relacionadas com os dias de internamento e número de óbitos respetivamente.

No estudo relacionado com os dias de internamento, numa primeira etapa utilizou-se um modelo de regressão poisson (simples) para relacionar os dias de internamento com a variável classe etária/gênero. Foram criados dois modelos de regressão a partir dos dados. Numa segunda etapa foi abordado o número de dias de internamento, através da modelação de poisson (múltipla) de várias variáveis regressoras. Numa primeira análise verificou-se que apenas as variáveis obesidade, bronquite, ARDS, asma, cirrose, ITU, IRA , HIV, oxigênio e ventilação eram significativas para a explicação dos dias de internamento. A remoção das outras variáveis foi realizada em função da significância estatística dos coeficientes, mas também foi levado em conta o contexto e especificidade do problema. Para além do cálculo dos I.C. para os coeficientes, procedeu-se também à análise de resíduos Breusch-Pagan. Após a obtenção das respectivas estatísticas de teste observou-se que os resíduos são independentes. O modelo obtido foi considerado válido para inferência estatística após a verificação dos principais pressupostos (satisfeitos).

No segundo estudo, foram aplicados modelos de regressão logísticos para modelação do número de óbitos em relação às outras variáveis. Algumas variáveis não apresentavam significância estatística no modelo produzido, pelo que foram removidos, obtendo o Model

1. Sendo o alcoolismo não significativo (<0,05) para o modelo, foi necessário proceder à análise de um outro modelo (Model 2), sendo os valores estimados dos coeficientes presentes, todas estatisticamente significativos. No sentido de escolher o melhor modelo realizou-se a comparação através do teste de verossimilhança e do AIC, os resultados mostraram que o Model 1 é o melhor modelo, sendo a variável retirada (alcoolismo) significativa para o modelo. Neste sentido, procedeu-se ao estudo em torno do Model 1, analisando o IC, a ANOVA, OR, testes sobre resíduos de Pearson, Deviance entre outros. Todos os testes realizados apontaram para a evidência estatística de que o modelo está ajustado corretamente. Finalmente, foi avaliada a capacidade de previsão do modelo através da curva ROC. O valor de corte foi definido em 0,5, isto é, os pacientes cuja probabilidade prevista está acima de 0,5 são definidos como óbitos (Y=1) e os restantes como altas (Y=0). Ao obter o valor de 0,803 para AUC, conclui-se que a capacidade de previsão do modelo é considerada excelente.

Por fim, com base nesse resultado e nas análises técnicas para a validação desse modelo, pode-se afirmar que o ajuste está adequado aos dados. Esses resultados podem contribuir para orientação e planeamento, dos internados de COVID-19, desde o processo de recolha dos dados, tratamento dos pacientes e tambem na gestão hospitalar e de recursos.

5.1 Trabalhos futuros

Os futuros estudos que consideramos importantes são:

- Aplicar análise de regressão linear para dados simbólicos do tipo intervalo;
- Fazer uma análise por região, tentar obter dados mais completos e fidedignos;
- Estudar outras abordagens de classificação como clustering ou SVM.

Bibliografia

[1] Aquino, D. S. Influência do acesso a saneamento básico na incidência e na mortalidade por covid-19: análise de regressão linear múltipla nos estados brasileiros. *Revista Thema 18* (2020), 319–331.

[2] Braga, A. C. *Curva ROC: Aspetos Funcionais e Aplicações*. PhD thesis, 2001.

[3] Casella, George e Berger, R. L. *Inferência estatística*. Cengage Learning, 2010.

[4] De Amo, S. Técnicas de mineração de dados. *Jornada de Atualizaçao em Informatica* (2004), 26.

[5] Fávero, L. P., Belfiore, P., Silva, F. d., and Chan, B. L. Análise de dados: modelagem multivariada para tomada de decisões, 2009.

[6] Franco, M. C. Complicações neuromusculares em indivíduos adultos após infecção por covid-19: Revisão integrativa.

[7] Fundão, N. M. *Modelo de Regressãoão Linear*. PhD thesis, 2018.

[8] Fávero, L. P., and Belfiore, P. *Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata*. Elsevier Brasil, 2017.

[9] Hosmer, David W e Lemeshow, S. e. C. E. Regressão logística aplicada 2ª edição. *Nova York: Jhon Wiley and Sons Inc*.

[10] Hosmer, David W e Lemeshow, S. e. L. S. Regressão logística aplicada. série wiley na seção de textos e referências de probabilidade e estatística.

[11] Hosmer, David W e Lemeshow, S. e. S. R. X. O modelo de regressão logística múltipla. *regressão logística aplicada 1*.

[12] INSP. Conhecimentos, atitudes e práticas sobre à covid-19 na população residente em cabo verde, 2020.

[13] Korkmaz, Muhammet Furkan e Türe, E. e. D. B. A. e. K. Z. B. As características epidemiológicas e clínicas de 81 crianças com covid-19 em um hospital pandêmico na turquia: um estudo de coorte observacional.

[14] Machado, F. N. R. *Big Data O Futuro dos Dados e Aplicações*. Saraiva Educação SA, 2018.

[15] Maroco, J. Análise estatística, com utilização do spss.(2003) edições sílabo. *Lisboa, Portugal*.

[16] McCullagh, Peter e Nelder, J. A. *Modelos lineares generalizados*.

[17] Montgomery, D. C. *Design e análise de experimentos*. John wiley & filhos, 2017.

[18] Paulino, C. D., and da Motta Singer, J. *Análise de dados categorizados*. Editora Blucher, 2006.

[19] Paulo, D. G. *Modelo de egressão de Poisson Generalizado: análise de dados de contagem com sobredispersão e subdispersão*. PhD thesis, 2019.

[20] Penha, R. N. Um estudo sobre regressão logística binária. *Minas Gerais* (2002).

[21] Pereira, H. T. *Estudo da distribuição de Poisson generalizada*. PhD thesis, Universidade de Brasília, 2016.

[22] Pires, A. M., and Branco, J. A. Introduçao aos métodos estatısticos robustos.

[23] Ramalho, J. J. D. S. *Modelos de regressão para dados de contagem*. PhD thesis, Universidade de Evora (Portugal), 1996.

[24] Rodrigues, S. C. A. *Modelo de regressão linear e suas aplicaçõesções*. PhD thesis, Universidade da Beira Interior (Portugal).

[25] Santana, L. Ofício nº 001/2022/covid19/ufam.

[26] Tadano, Y. d. S., Ugaya, C. M. L., and FrAnCo, A. T. Método de regressão de poisson: metodologia para avaliação do impacto da poluição atmosférica na saúde populacional. *Ambiente & Sociedade 12*, 2 (2009), 241–255.

[27] Turkman, M. A. A., and Silva, G. L. Modelos lineares generalizados–da teoria à prática. *Sociedade Portuguesa de Estatística, Lisboa* (2000).

[28] Valle Benavides, A. R. d. Curvas roc (receiver-operating-characteristic) y sus aplicaciones.

[29] Weisberg, S. *Regressão linear aplicada*, vol. 528. John Wiley & Filhos, 2005.