UNIVERSIDADE ABERTA

Modelação Linear e Extensões:

Aplicação da Regressão Logística no Estudo de Câncer da Mama

Fernando Emílio da Cruz Ribeiro Vaz

Mestrado em Estatística, Matemática e Computação

Área de especialização em Estatística Computacional

2020

UNIVERSIDADE ABERTA

Modelação Linear e Extensões:

Aplicação da Regressão Logística no Estudo de Câncer da Mama

Fernando Emílio da Cruz Ribeiro Vaz

Mestrado em Estatística, Matemática e Computação

Área de especialização em Estatística Computacional

Dissertação orientada pela Prof. Doutora Maria do Rosário Ramos

2020

“A modelagem em Ciência ainda permanece, pelo menos em parte, uma arte.”

(MC CULLAGH e NELDER, 1989)

Resumo

Uma pesquisa realizada pela Agência Internacional de Pesquisa em Câncer (IARC) indica que o câncer de mama é um dos três tipos de maior incidência, e é o que mais atinge as mulheres a nível mundial, representando um sério problema de saúde pública com predomínio de diagnóstico em fase avançada da doença em países em desenvolvimento como Cabo Verde. Neste trabalho, foi feito um estudo quantitativo do câncer de mama com o objetivo de elaborar um modelo que possa auxiliar no diagnóstico de câncer de mama. Foi feito um apanhado das metodologias com ênfase nos estudos caso-controlo e nos modelos lineares generalizados (GLM), destacando o modelo de regressão logística e apresentando procedimentos computacionais, no âmbito de pesquisas relacionadas ao câncer de mama. O modelo de regressão logística é amplamente utilizado em estudos na área médica dada a sua flexibilidade e tipo de resultados, que podem ser interpretados de forma bastante prática, uma vez que, possibilita identificar fatores de risco para um desfecho da variável de interesse, desde que o modelo esteja bem ajustado. Para ilustrar o uso da regressão logística foi utilizado um banco de dados real com 116 pacientes do sexo feminino no âmbito de um estudo sobre o câncer de mama usando diversos biomarcadores antropométricos, obtidos por meio de análise ao sangue. Com isso, recorreu-se a técnicas de seleção de variáveis implementadas no R-Studio e no SPSS, seleção stepwise, para construir modelos que permitiram selecionar as variáveis que influenciam o sucesso no diagnóstico de câncer de mama. Para avaliar o desempenho do modelo construído recorreu-se à Curva ROC analisando a respetiva área. Este trabalho permitiu identificar que a Idade, o índice de massa corporal, o nível de glicose no sangue e a resistina são fatores de risco na previsão do câncer de mama. Os resultados indicaram que a presença do câncer de mama nas mulheres pode ser predito com uma chance de acerto de 77,6% e uma sensibilidade e especificidade de 75 e 79,7%, respetivamente.

Palavras chave: Câncer de mama, modelos lineares generalizados, regressão logística, biomarcadores, previsão.

Abstract

Research carried out by the International Agency for Research on Cancer (IARC) indicates that breast cancer is one of the three types with the highest incidence, and is the one that most affects women worldwide, representing a serious public health problem with a prevalence of diagnosed at an advanced stage of disease in developing countries like Cabo Verde. In this work, was performed a quantitative study of breast cancer in order to elaborate a model that may help in the diagnosis of breast cancer. An overview of the methodologies was made with an emphasis on case-control studies and generalized linear models (GLM), highlighting logistic regression presenting computational procedures, in the scope of research related to breast cancer. Logistic regression is growingly linked to the medical field, as it aims to predict risk factors for an outcome of the variable of interest, as long as the model is well adjusted. To illustrate the use of logistic regression, a real database with 116 female patients was used as part of a study on breast cancer using several anthropometric biomarkers, obtained through blood analysis. Stepwise technique for variable selection were implemented in R-Studio and SPSS, in order to build significant prediction models. To assess the performance of the built model, the ROC Curve was used to analyze the respective area. This work allows us to find that Age, BMI, blood Glucose level and Resistin are risk factors in predicting breast cancer. It was concluded that the presence of breast cancer in women can be predicted with a chance of success of 77.6% and a sensitivity and specificity of 75 and 79.7%, respectively.

Key words: Breast cancer, generalized linear models, logistic regression, biomarkers, prediction.

Agradecimentos

Gostaria de agradecer à minha família, sempre muito presente, me apoiando e dando força.

Agradeço a professora Doutora Maria do Rosário Ramos, minha orientadora, pela disponibilidade em o ser, e pelas sugestões e críticas que foram importantes para a concretização desta dissertação.

A todos que de alguma forma contribuíram para a realização deste trabalho, muito obrigado!

Índice

Lista de Gráficos

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas

ACLCC	Associação Cabo-verdiana da Luta Contra o Cancro
ADA	American Diabetes Association
AIC	Critério de informação de Akaike
AUC	Area under Curve
BIC	Critério de Informação de Bayes
EMV	Estimativa de Máxima Verossimilhança
gl	Graus de Liberdade
HOMA	Homeostatic Model Assesment
IC	Intervalo de confiança
K-NN	K-Nearest Neighbour
ML	Machine Learning

MLG	Modelos Lineares Generalizados
OMS	Organização Mundial da Saúde
OR	Odds Ratio
RF	Random Forest
RL	Regressão Logística
ROC	Receiver Operating Characteristc
SVM	Support Vector Machine
VIF	Variance Inflation Factor

CAPÍTULO 1

1. Introdução

O câncer é a principal causa de morte em todo o mundo, respondendo por 13% de todas as mortes. Para as mulheres, o câncer de mama é uma das principais causas de morte nos países desenvolvidos e em desenvolvimento. Em 2012, o número de casos de câncer de mama em todo o mundo foi estimado em 14,1 milhões de novos casos e 8,2 milhões de mortes. Estima-se que a incidência de câncer de mama tenha aumentado 20% desde 2008 e a mortalidade 14% (Boughorbel, Al-Ali, & Elk, 2016). De acordo com os dados da Organização Mundial Saude (OMS, 2019) para 2018, estima-se que 627.000 mulheres morreram de câncer de mama, o que representa, aproximadamente 15% de todas as mortes por câncer entre as mulheres. Estima-se ainda que, em 2020, o número de novos casos anuais de câncer será de 15 milhões, sendo que 60% desses ocorrerão em países em desenvolvimento.

Em Cabo Verde, por exemplo, a Associação Cabo-verdiana da Luta Contra o Cancro – ACLCC estima que o câncer da mama é o segundo com maior taxa de incidência nas mulheres com idades compreendidas entre os 40 e 50 anos, constituindo este intervalo um fator de risco. De acordo com Pinheiro et al. (2013), estudos demonstram que existem diversos fatores de risco relacionados a esta neoplasia, entre os quais: idade, duração da atividade ovariana, hereditariedade, hábitos de vida (tipo de alimentação, consumo de bebida alcoólica e de tabaco), medicamentos (anticoncepcionais, repositores hormonais), localização geográfica, entre outros.

Segundo o American Cancer Society (ACS) um fator de risco é algo que aumenta as chances de contrair uma doença, como o câncer. Mas ter um fator de risco, ou mesmo muitos, não significa que você terá a doença. Embora você não possa alterar alguns fatores de risco para câncer de mama - histórico familiar e envelhecimento, por exemplo, existem alguns fatores de risco que você pode controlar, por exemplo, os relacionados ao estilo de vida.

O câncer de mama, possivelmente, é a neoplasia mais temida pelas mulheres, uma vez que a sua ocorrência causa grande impacto psicológico, funcional e social, atuando negativamente nas questões relacionadas à autoimagem e à perceção da sexualidade (Pinheiro, et al., 2013).

Neste trabalho é apresentado um estudo caso-controlo que tenta identificar fatores de risco para o câncer de mama, comparando indivíduos em que está presente (casos) ou ausente (controlos), de forma retrospetiva na tentativa de encontrar possível associação. A razão de chances (odds ratio) é usada como medida estatística de associação. Neste contexto, as técnicas computacionais auxiliam no diagnóstico e prognóstico do câncer de mama, permitindo interpretações ou conclusões que não poderiam ser feitas usando procedimentos estatísticos padrão. Dada a natureza dos dados disponíveis, tal é conseguido recorrendo aos **Modelos Lineares Generalizados** – GLM – Nelder & Wedderburn (1972).

Devido ao grande número de modelos que englobam e à facilidade de análise associada ao rápido desenvolvimento computacional que se tem verificado nas últimas décadas, os MLG têm vindo a desempenhar um papel cada vez mais importante na análise estatística, apesar das limitações ainda impostas, nomeadamente por manterem a estrutura de linearidade, pelo facto das distribuições se restringirem à família exponencial e por exigirem a independência das respostas. Há já atualmente, na literatura, muitos desenvolvimentos da teoria da modelação estatística onde estes pressupostos são relaxados, mas, o não acompanhamento dos modelos propostos com software adequado à sua fácil implementação, faz com que se anteveja ainda, por algum tempo, um domínio dos MLG em aplicações de natureza prática.

1.1. Motivação e Objetivos

O objetivo deste estudo foi utilizar a Modelação Estatística por via de um Modelo Linear Generalizado no problema de pesquisa de biomarcadores para previsão do diagnóstico de câncer de mama, com base em dados antropométricos e parâmetros que podem ser coletados em análises de sangue de rotina. Aprofundar o estudo do Modelo Linear Generalizado e aplicar a uma base de dados da área médica de pacientes com câncer de mama e identificar variáveis importantes para a variável desfecho.

1.2. Estrutura da dissertação

O presente capítulo, que constitui a Introdução, enquadra os aspetos gerais do trabalho, a importância do tema de aplicação, os modelos estatísticos abordados, bem como a restante estrutura da dissertação;

Capítulo 2: Apresenta uma síntese da pesquisa sobre estudos anteriores.

Capítulo 3: Segue a revisão abordando os Modelos Lineares Generalizados, onde serão discutidos os métodos de estimação dos parâmetros nessa classe, propriedades assintóticas dos estimadores de máxima verossimilhança, teste de hipóteses, critérios para seleção de modelos, alguns parâmetros que permitem verificar a qualidade do ajuste e a analise dos resíduos. Ainda no capítulo são abordados, resumidamente, os modelos de Poisson e Binomial Negativa

O **Capítulo 4** aborda o Modelo Linear Generalizado de uma forma mais específica, descrevendo o modelo de regressão Logística.

No **Capítulo 5** é apresentado uma análise descritiva dos dados utilizando os softwares estatísticos SPSS, Excel e R, assim como todo o tratamento realizado na base de dados para a sua utilização nos modelos e, além disso, são apresentados os resultados obtidos do modelo usado.

Por fim, no **Capítulo 6** são descritas as análises e considerações finais do trabalho**.** Em Anexo são apresentados alguns resultados numéricos dos modelos e análises realizadas, como informação complementar, em detalhe.

CAPÍTULO 2

2. Revisão da literatura

Nesta secção apresentam-se alguns trabalhos realizados no âmbito do estudo do câncer de mama. Os estudos têm mostrado que as ocorrências do câncer de mama estão associadas a diversos fatores, dos quais podemos citar os fatores individuais, fatores ambientais e comportamentais, fatores da história reprodutiva e hormonal individual e fatores genéticos e hereditários. Relativamente a fatores individuais parece ser mais usual as variáveis como a faixa etária. A nível ambiental e comportamental parecem ser comuns as variáveis como o sobrepeso, o sedentarismo, o consumo de bebidas alcoólicas ou a exposição frequente radiações ionizantes (Raio-X) (American Cancer Society, 2019). Relativamente a fatores da história reprodutiva e hormonal podem-se destacar o uso de contracetivos hormonais (estrogênio-progesterona), o facto de não ter filhos e ter feito reposição hormonal pós-menopausa, principalmente por mais de cinco anos. E, em relação a fatores genéticos e hereditários destacam-se os casos de câncer de mama na família, principalmente antes dos 50 anos, a história familiar de câncer de ovário ou a alteração genética, especialmente nos genes BRCA1 e BRCA2 (Coelho, et al., 2018).

2.1. Câncer de mama

Segundo a OMS, o câncer, também denominado de neoplasia (do grego “new growth”) ou tumor maligno, é caracterizado por um crescimento anormal e incontrolável de células. A gênese do câncer é um processo complexo que envolve alterações genéticas, eventos epigenéticos em oncogenes, genes supressores de tumor e genes anti-metástases.

O processo de carcinogénese apresenta algumas etapas essenciais para que haja o surgimento do tumor e a sua proliferação. São elas: iniciação: fase em que os genes sofrem ação de fatores cancerígenos; promoção: fase em que os agentes oncopromotores atuam na célula já alterada; e progressão: caracterizada pela multiplicação descontrolada e irreversível da célula (INCA, 2011).

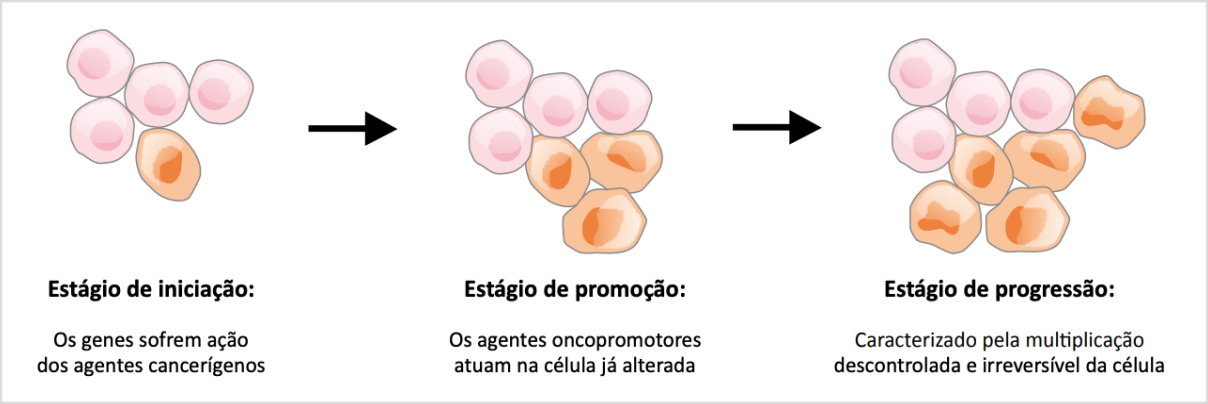

Figura 1. Passo a passo do processo de carcinogénese.
Fonte: (INCA, 2011)

Todos os anos, um milhão de mulheres são diagnosticadas com câncer de mama, de acordo com o relatório da organização mundial de saúde metade delas acabam por morrer, porque geralmente é tarde quando os médicos detetam o câncer. O câncer de mama pode ser categorizado em: câncer de mama maligno e benigno.

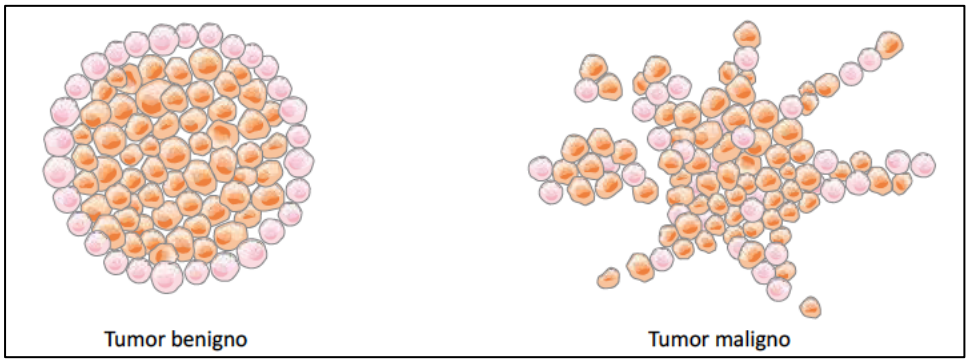

Figura 2. Diferenças entre tipos de tumores.
Fonte: (INCA, 2011)

O câncer de mama se origina a partir de células epiteliais das unidades terminais ductolobulares. Eventos em série irão determinar o surgimento da neoplasia mamária, sendo que as alterações no tecido normal da mama resultam em uma hiperplasia que evolui para um carcinoma in situ (proliferação de células epiteliais malignas), limitado aos ductos e lóbulos mamários, podendo evoluir com 30% de chance, para um carcinoma invasivo da mama (Previato, 2013).

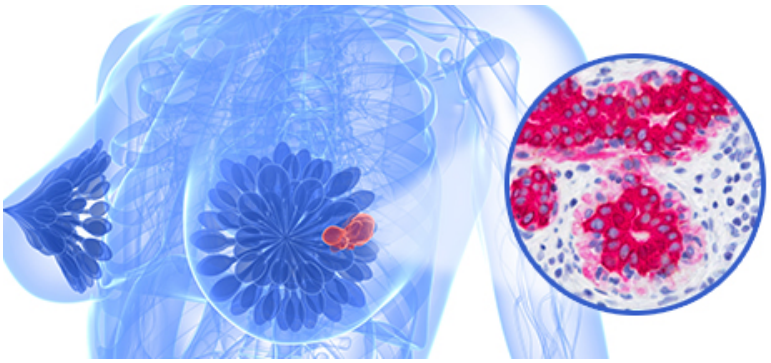

Figura 3. Ilustração pictórica do câncer de mama.

Fonte: (Portal do Médico, 2016)

Todas as neoplasias de mama têm origem genética. Acredita-se que 90-95% delas possam ser esporádicas (não-familiares) e decorrer de mutações somáticas que se verificam no decorrer da vida, 5-10%, hereditárias (familiares) devido à herança de uma mutação germinativa ao nascimento, que confere as mulheres suscetibilidade ao câncer de mama. De uma forma geral, esta neoplasia pode resultar da interação de fatores genéticos com estilo de vida, fatores reprodutivos ou hormonais e meio ambiente (Previato, 2013).

2.2. Estudos anteriores

Os trabalhos que se seguem são de pesquisas realizados no campo biomédico relacionados à classificação do câncer de mama usando algoritmos de classificação.

Anene (2019) propôs um modelo de ML para a classificação do câncer de mama. Para isso, propôs a regressão logística afim de comparar o seu desempenho com outros modelos ML existentes, nomeadamente Support Vector Machine (SVM), Naïve Bayes (NB) e Multilayer Perceptron (MLP), assim, concluiu que a LR, dada sua simplicidade e baixa complexidade é um bom modelo para prever o câncer de mama comparativamente com os outros ML.

Patricio et al. (2018) propuseram modelos preditivos para deteção de câncer de mama com base em nove biomarcadores, construídos com três algoritmos de classificação: regressão logística, SVM e RF. Obtiveram evidências promissoras de que os modelos que combinam idade, IMC e parâmetros metabólicos podem ser uma ferramenta poderosa e eficaz para a predição de câncer de mama.

Sultana e Jilani (2018) propuseram o método de regressão logística e os multi-classificadores para predizer o câncer de mama. Concluíram que, em termos de precisão, a regressão logística produz a máxima precisão e melhor modelo para prever o câncer de mama.

Boughorbel et al. (2016) compararam a performance de oito modelos para a predição de câncer de mama, baseado na área abaixo da curva ROC para diversos períodos de diagnostico de câncer de mama, e concluíram que, através do teste t pareado, o algoritmo Random Forest (RF) apresenta um maior poder de discriminação. Entretanto, do conjunto de variáveis utilizados, os autores afirmam que a variável linfonodo é preditor mais importante para o diagnóstico de câncer de mama independentemente do período do diagnóstico.

Chhatwal et al. (2009) criaram dois modelos de estimativa de risco de câncer de mama com base em alguns descritores para auxiliar os radiologistas no diagnóstico do câncer de mama usando regressão logística para a tomada de decisões para a detecção precoce do câncer de mama, com o objetivo de discriminar entre doenças mamárias benignas e malignas e identificar os aspectos mais importantes associados ao câncer de mama. Os resultados mostraram que a combinação de um modelo de regressão logística e avaliação de radiologistas tem um desempenho melhor em discriminar entre lesões benignas e malignas.

Abdolmaleki et al. (2004) utilizaram o método discriminante logístico para diferenciar câncer de mama maligno e benigno em um grupo de pacientes com lesões mamárias comprovadas com base em parâmetros ultrassônicos. Neste sentido, utilizaram uma base de dados que inclui imagens ultrassonográficas de 273 pacientes, compostas por 14 variáveis quantitativas. Os resultados obtidos mostraram que o método discriminante logístico foi capaz de classificar corretamente 67 dos 72 casos apresentados na amostra de validação. Os resultados indicam uma precisão diagnóstica notável de 93%.

No artigo de Zhou et al. (2004) “Cancer classification and prediction using logistic regression with Bayesian gene selection”, foi proposto uma abordagem bayesiana para seleção e classificação de genes para os dados de câncer de mama usando o modelo de regressão logística. Os resultados experimentais obtidos mostram que efetivamente o método pode encontrar genes com alta precisão e sensibilidade.

CAPÍTULO 3

3. Modelos Lineares Generalizados

3.1. Conceitos de base

Os Modelos Lineares Generalizados surgiram com o objetivo de solucionar problemas em que o objeto de estudo não é uma variável quantitativa ou não se adapta às exigências do modelo de regressão linear clássico. Fazendo uso de distribuições da família exponencial, que partilham certas características, são uma extensão dos modelos de regressão linear. Englobam, para a variável resposta (**componente aleatória**), as distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens. Pressupõe um conjunto de variáveis independentes ou explanatórias que descrevem a estrutura linear do modelo (**componente sistemática**) e uma **função de ligação** (f) entre a média da variável de resposta (μ) e a estrutura linear (η). Assim, tem-se que

$$f(\mu) = \eta \qquad (3.1)$$

No modelo clássico de regressão linear, tem-se, recorde-se

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \varepsilon \qquad (3.2)$$

sendo Y o vetor, de dimensões $n \times 1$, da variável resposta, $\boldsymbol{X\beta}$, o componente sistemático, X a matriz do modelo, de dimensão $n \times p$, $\beta = (\beta_1, ..., \beta_p)^T$, o vetor de parâmetros desconhecidos e $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, um vetor de erros aleatórios com distribuição que se supõe $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $i = 1, \dots, n$. Assim, tem-se a distribuição normal $\mathrm{N}(\mu, \sigma^2\mathrm{I})$ de Y que define o componente aleatório e o vetor de médias μ da distribuição normal é igual ao preditor linear que representa o componente sistemático. Essa é a forma mais simples de ligação entre esses dois componentes,

sendo denominada de função de ligação identidade (Cordeiro & Demétrio, Modelos Lineares Generalizados e Extensões, 2013).

Durante o planeamento do experimento é estabelecido o componente sistemático, fundamental para a obtenção de conclusões mais fiáveis, o que resulta em modelos de regressão, modelos de análise de variância e de análise de covariância. Definidas as medidas a serem realizadas, que podem ser contínuas ou discretas, o componente aleatório é especificado, o que exige o ajuste de diferentes distribuições. De acordo com McCullagh e Nelder (1989), o MLG é uma extensão do modelo linear clássico. Schmidt (2003) afirma que os MLG representam a união de modelos lineares e não-lineares com uma distribuição da família exponencial.

Em suma, os MLG unificam, tanto do ponto de vista teórico, como conceptual, a teoria da modelação estatística até então desenvolvida. São, pois, casos particulares dos modelos lineares generalizados os seguintes modelos (Turkman, 2000): modelo de regressão linear clássico; modelos de análise de variância e covariância; modelo de regressão logística; modelo de regressão de Poisson; modelos log-lineares para tabelas de contingência multidimensionais; modelo probit para estudos de proporções, etc.

Retomando a abordagem de McCullagh e Nelder (1989), um MLG assenta sobre três componentes:

1) **Componente aleatória**

Dado o vetor de covariáveis x_i, as variáveis y_i são (condicionalmente) independentes com distribuição pertencente à família exponencial, com $(y_i|\boldsymbol{x_i}) = \mu_i$,para $i = 1, \ldots, N$.

2) **Componente estrutural ou sistemática**

Consiste numa combinação linear de variáveis preditoras, tendo p variáveis preditoras e N observações, tal que:

$$\eta = \boldsymbol{X\beta} = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + \beta_3 x_3(i) + \cdots + \beta_p x_p(i), \text{ para } i = 1, \ldots, N. \qquad (3.3)$$

3) Função de Ligação

A função de ligação entre a variável resposta e as variáveis preditoras é uma função monótona e diferenciável que relaciona a média da resposta ao preditor linear, estabelecendo também uma relação entre a componente aleatória e a componente sistemática do modelo. Assim

$$\eta_i = g(\mu_i). \qquad (3.4)$$

onde *g(.)* é chamada de função de ligação. As componentes aleatória e sistemática encontram-se assim "ligadas" por *g(.)*.

Algumas literaturas afirmam que a utilização da função de ligação canônica, ou seja, utilizando a Função Identidade como ligação, implica algumas propriedades interessantes para os estimadores, porém não quer dizer que deva ser utilizada sempre. Essa escolha é conveniente porque, além de simplificar a obtenção das estimativas dos parâmetros do modelo, também facilita o cálculo do intervalo de confiança para a média da variável resposta. Contudo, a conveniência não implica necessariamente em qualidade de ajuste do modelo.

3.2. Família Exponencial

A família exponencial é uma importante classe de distribuições que partilham, entre si, certas caraterísticas. Assim, por exemplo, pertencem a esta família as distribuições mais utilizadas e comuns, entre as quais se destacam a normal, a binomial, a binomial negativa, gama, Poisson, normal inversa, multinomial, beta, logarítmica, entre outras.

O conceito de família exponencial foi introduzido na Estatística por Fisher[1].

[1] **Fisher** foi um estatístico britânico, biólogo evolucionista e geneticista.

Figura 4. Sir Ronald Aylmer Fisher (17/02/1890 - 29/07/1962).
Fonte: http://www-history.mcs.st-andrews.ac.uk/PictDisplay/Fisher.html

Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial se a sua função densidade de probabilidade puder escrita na forma

$$f\left(y\middle|\theta,\phi\right)=exp\left\{\frac{y\theta-b\left(\theta\right)}{a\left(\phi\right)}+c\left(y,\phi\right)\right\} \qquad (3.5)$$

onde θ e ϕ são parâmetros escalares, a(.), b(.) e c(.,.) são funções reais conhecidas e y os dados observados (variável de interesse). A introdução de ∅ na função densidade de probabilidade permite contemplar algumas distribuições biparamétricas pertencentes à família exponencial. Dentre as distribuições uni-paramétricas pertencentes à família de distribuição podemos destacar a distribuição binomial (com n conhecido) e a distribuição de Poisson. Já entre as distribuições bi-paramétricas podemos destacar a distribuição normal, a gama, a binomial negativa e a normal inversa.

De acordo com Maxwell (1992), simplesmente calculando as derivadas de primeira e segunda ordem da função $b(\theta)$, pode-se obter a média e a variância da variável y, como demonstrado na equação abaixo.

$$E(y) = \mu = \frac{d^2 b(\theta)}{d\theta}, \text{ e} \tag{3.6}$$

$$var(y) = \frac{d^2 b(\theta)}{d^2 \theta} \tag{3.7}$$

Como se pode observar, a variância de *y* é, na verdade, um produto que depende de sua média e de *a(.)*. A parcela correspondente à segunda derivada de *b(.)* é conhecida como função da variância, *V(μ)*. Por outro lado, podemos facilmente demonstrar que o valor médio e a variância da distribuição da variável aleatória acima apresentada são dados por *b'(.)* e *a(.)b(.)*.

Suponhamos que *y*, uma variável aleatória resposta, segue uma distribuição Normal com valor médio μ e variância σ^2, sua função densidade de probabilidade é dada por

$$\begin{aligned} f\left(y\middle|\mu,\sigma^2\right) &= \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu}{2}\right) - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln\left(2\pi\sigma^2\right)\right)\right\} \end{aligned} \tag{3.8}$$

A função resultante é do tipo da função densidade de probabilidade com $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2}$, $a(\phi) = \sigma^2$ e $c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln\left(2\pi\sigma^2\right)\right)$. Neste caso, pode-se notar que $E(Y) = b'(\theta) = \mu$ e $var = (Y) = a(\phi)\, b'' = \sigma^2$.

Na Tabela 1 são apresentadas algumas distribuições da família exponencial que a variável resposta pode seguir e o tipo de dados no qual a distribuição é aplicada.

Tabela 1. Distribuições e Tipo de Dados.

Distribuição	Tipos de Dados
Poisson	Contagens
Binomial Negativa	Contagens
Normal	Contínuos
Gama	Contínuos Positivos
Normal Inversa	Contínuos Positivos

Adaptado de: Paula (2010).

Existem inúmeras funções que podem ser consideradas como função de ligação, e a decisão de qual função utilizar provém essencialmente do tipo de resposta que se pretende. A Tabela 2 apresenta uma relação entre as distribuições mais comuns, da componente aleatória, com a função de ligação que frequentemente é utilizada.

Tabela 2. Funções de ligação e distribuições mais comuns nos MLG.

Distribuição	Função de Ligação (designação)	Função de Ligação	Função de Variância
Normal	Identidade	μ	1
Poisson	Logit	$log\left(\frac{\mu}{1-\mu}\right)$	$\mu\frac{1-\mu}{n}$
Binomial	Logarítmica	$log(\mu)$	μ
Gamma	Recíproca	$\frac{1}{\mu}$	μ^2
Gaussiana inversa	Quadrática inversa	$\frac{1}{\mu^2}$	μ^3

Adaptado de: Pousinho (2013)

Os Modelos Lineares Generalizados apresentam uma grande flexibilidade na sua aplicação a diferentes estudos estatísticos, sendo a sua única limitação a necessidade de se garantir que as

variáveis explicativas entrem no modelo através de uma combinação linear, com a distribuição da variável dependente pertencente à família exponencial.

3.3. Estimação de parâmetros dos MLG

Para ajustar um modelo de regressão devemos estimar os parâmetros do modelo. Conforme Hauck Jr e Donner (1977), o método mais usual em estatística para a estimação de parâmetros é o método da máxima verossimilhança. Em geral, nos modelos lineares generalizados os resultados obtidos não são lineares e, assim, deve-se obter uma solução numérica utilizando o método de Newton-Rapson ou o escore de Fisher apresentado por Nelder e Wedderburn (1972).

3.3.1. Método da máxima verossimilhança

Pela teoria de estimação de máxima verosimilhança, sabe-se que os estimadores de máxima verosimilhança maximizam a função log-verosimilhança, pelo que retirar as variáveis resultam geralmente num valor pequeno para a log-verosimilhança, a semelhança do que acontece com o R^2 no modelo de regressão clássica.

A estimação da Máxima Verossimilhança, dada por (Turkman e Silva, 2000; Hosmer & Lemeshow, 2000), em função de β é,

$$L(\beta) = \prod_{i=1}^{n} f\left((y_i)|\theta_i, \phi\right) \qquad (3.9)$$

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

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

Recordando que θ é escrito em função de μ, podemos escrever o logaritmo da função de verossimilhança da seguinte forma:

$$ln\left(L\left(\beta\right)\right)=l\left(\beta\right) \tag{3.10}$$

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

$$=\sum_{i=1}^{n}l_i\left(\beta\right)$$

Sendo que l_i é a contribuição de cada observação y_i para verossimilhança.

Desta forma os estimadores de máxima verossimilhança para β podem ser obtidos mediante a resolução do sistema de equações seguinte,

$$\frac{\partial l_i\left(\beta\right)}{\partial\beta}=0=\sum_{i=1}^{n}\frac{\partial l_i\left(\beta\right)}{\partial\beta_j},\quad j=1,2,...,p \tag{3.11}$$

De acordo com Turkman e Silva (2000), para obter estas equações escrevemos:

$$\frac{\partial l_i\left(\beta\right)}{\partial\beta_j}=\frac{\partial l_i\left(\theta_i\right)}{\partial\theta_i}\frac{\partial\theta_i\left(\mu_i\right)}{\partial\mu_i}\frac{\partial\mu_i\left(\eta_i\right)}{\partial\eta_i}\frac{\partial\eta_i\left(\beta\right)}{\partial\beta_j}$$

sendo

$$\frac{\partial l_i\left(\theta_i\right)}{\partial\theta_i}=\frac{\omega_i\left(y_i-b^{'}\left(\theta_i\right)\right)}{\phi}=\frac{\omega_i\left(y_i-\mu_i\right)}{\phi},$$

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

Assim

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\omega_i(y_i - \mu_i)}{\phi}\frac{\phi}{\omega_i var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} z_{ij} \qquad (3.12)$$

e as equações de verossimilhança para β são

$$\sum_{i=1}^{n}\frac{(y_i - \mu_i) z_{ij}}{var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1,\ldots p \qquad (3.13)$$

A função score, definida abaixo como

$$S(\theta) = \frac{\partial l(\theta;\phi,Y)}{\partial \theta} \qquad (3.14)$$

é o vetor p-dimensional

$$S(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{s_i}(\boldsymbol{\beta}) \qquad (3.15)$$

onde $s_i(\beta)$ é o vetor de componentes $\dfrac{\partial l_i(\beta)}{\partial \beta_j}$ obtidos em (3.12).

O elemento genérico de ordem *j* da função score é apresentada na equação (3.14).

A matriz de covariância da função *score*, $\boldsymbol{I}(\boldsymbol{\beta}) = E\left[\frac{\partial S(\beta)}{\partial \beta}\right]$ é conhecida por *matriz de informação de Fisher*. Para obter a matriz de informação de Fisher temos de considerar o valor esperado das segundas derivadas de $l_i(\beta)$.

Tem-se, para famílias regulares, que

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

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

$$-\sum_{i=1}^{n} E\left(\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n}\frac{z_{ij}z_{ik}}{var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \qquad (3.16)$$

Como referido, as equações (3.13), não têm, em geral, solução analítica e, portanto, a sua resolução implica o recurso a métodos iterativos.

Método iterativo

O método de Newton-Raphson é usado para obter as estimativas de máxima verossimilhança nos modelos lineares generalizados. Aplicando o método, a (p+1)-ésima aproximação da estimativa de β é dada por

$$\hat{\beta}^{(p+1)} = \hat{\beta}^{(p)} - \left[\boldsymbol{H}\left(\widehat{\boldsymbol{\beta}}^{(\boldsymbol{p})}\right)\right]^{-1}\boldsymbol{S}\left(\widehat{\boldsymbol{\beta}}^{(\boldsymbol{p})}\right) \qquad (3.17)$$

onde $H\left(\hat{\beta}^{(p)}\right)$ é a matriz Hessiana, quadrada e de ordem (k+1), calculada em $\beta = \hat{\beta}^{(p)}$, com os seguintes elementos:

$h_{jt} = \frac{\partial^2 \ell(\beta)}{\partial \beta_j \partial \beta_t}$, $\quad j, t = 0,1,\ldots, k,$

e $S\left(\hat{\beta}^{(p)}\right)$ um vetor coluna de ordem (k+1), vetor dos scores, com elementos $S_j = \frac{\partial \ell(\beta)}{\partial \beta_j}$ calculado em $\beta = \hat{\beta}^{(p)}$.

De forma alternativa, usa-se o método de Scores de Fisher que envolve a substituição da matriz das segundas derivadas pela matriz dos respectivos valores esperados:

$E\left[\frac{\partial^2 \ell(\beta)}{\partial \beta_j \partial \beta_t}\right]$, $\quad j, t = 0,1,\ldots, k,$

Aplicando o método de Score de Fisher, a equação (3.17) é substituída por,

$$\hat{\beta}^{(p+1)} = \hat{\beta}^{(p)} + \left[\boldsymbol{\mathcal{I}}\left(\widehat{\boldsymbol{\beta}}^{(\boldsymbol{p})}\right)\right]^{-1}\boldsymbol{S}\left(\widehat{\boldsymbol{\beta}}^{(\boldsymbol{p})}\right) \qquad (3.18)$$

onde $\mathcal{I}\left(\hat{\beta}^{(p)}\right)$ é a matriz de informação de Fisher calculada a partir de $\hat{\beta}^{(p)}$.

A equação (3.18) por ser reescrita como:

$$\left[\mathcal{I}\left(\hat{\beta}^{(p)}\right)\right]\hat{\beta}^{(p+1)} = \left[\boldsymbol{\mathcal{I}}\left(\widehat{\boldsymbol{\beta}}^{(\boldsymbol{p})}\right)\right]\hat{\beta}^{(p)} + \boldsymbol{S}\left(\widehat{\boldsymbol{\beta}}^{(\boldsymbol{p})}\right) \qquad (3.19)$$

3.4. Seleção de modelos

A seleção de modelos é uma parte importante da modelação estatística que tem como objetivo encontrar o melhor modelo, ou melhores modelos, que descreva bem os dados observados para

explicar a variabilidade da variável resposta. Na análise e seleção de modelos, o principal objetivo é encontrar um modelo que forneça maior quantidade de informação em relação à variável resposta com o menor número e covariáveis. Conforme Turkman e Silva (2000), na prática, encontrar esse modelo não é fácil, o processo está intimamente relacionado ao problema fundamental da estatística que, segundo Fisher, é "o que se deve fazer com os dados?", pois é preciso encontrar um equilíbrio entre um bom ajuste e obter um modelo menos complexo. Os métodos de seleção foram desenvolvidos com o intuito de identificar um número pequeno de modelos suficientemente bons de acordo com determinados critérios.

De acordo com Oliveira (2010), no processo de seleção, é frequentemente utilizado dois modelos de referência:

- **Modelo completo ou saturado** – este modelo contém n parâmetros linearmente independentes, cuja a matriz do modelo é uma matriz identidade $n \times n$. Este modelo atribui toda a variação dos dados à componente sistemática. O modelo saturado serve de referência para medir a discrepância de um modelo intermédio com $k + 1$ parâmetros.
- **Modelo Nulo** – este modelo é o mais simples por ter apenas um único parâmetro. Neste modelo assume-se que todas as variáveis têm o mesmo valor médio, neste caso, atribui toda a variação dos dados à componente aleatória. A matriz do modelo corresponde a um vetor coluna unitário.

Existem duas questões relacionadas com a seleção do modelo, o critério e o algoritmo de seleção. Na secção seguinte descrevem-se o método stepwise e os critérios e procedimentos comumente usados na seleção do modelo.

3.4.1. Método stepwise

O método *stepwise* é um dos métodos mais aplicados em regressão logística. O método baseia-se na seleção automática das variáveis importantes para o modelo. No conjunto de variáveis independentes podem haver variáveis que pouco influenciam o conjunto de variáveis dependentes (saída). O método *stepwise* é usado para selecionar quais variáveis mais

influenciam o conjunto de saída podendo, assim, diminuir o número de variáveis a compor a equação de regressão.

São três as formas de se realizar uma regressão *stepwise*: (1) forward - quando a equação começa vazia e cada preditor entra, um por um, na equação; (2) backward - quando todos os preditores são incluídos de uma só vez na equação, e depois são retirados, um a um, até que se identifiquem os melhores preditores; (3) blockwise ou setwise – assemelha-se à regressão stepwise forward, mas, ao invés dos preditores serem incluídos individualmente, eles entram na equação em blocos (Abbad & Torres, 2001).

A Figura 3 apresenta um fluxograma que descreve o método *Stepwise.*

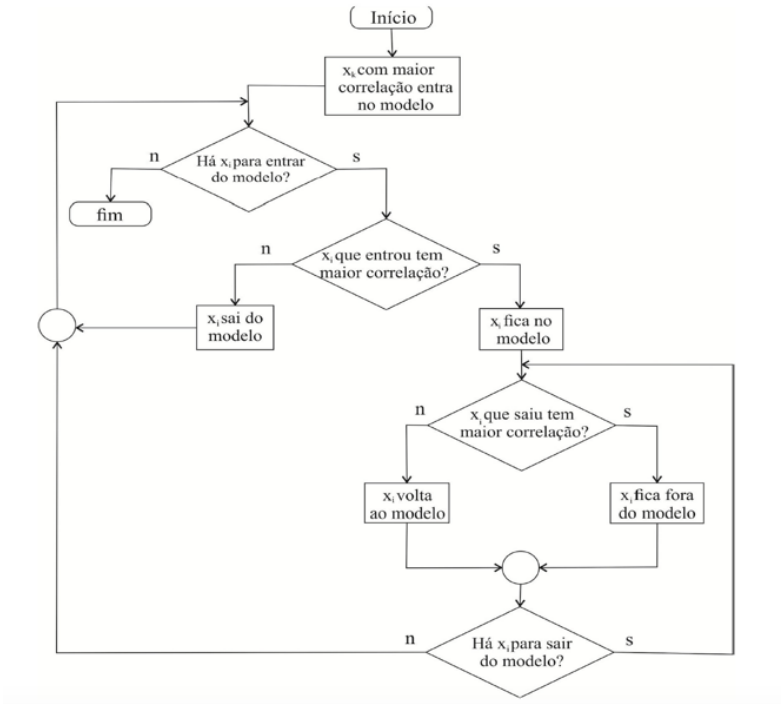

Figura 5. Fluxograma do método stepwise.
Fonte: (Alves, Lotufo, & Lopes, 2013)

Este método utiliza o Critério de Informação de Akaike (AIC - Akaike Information Criterion), na combinação das variáveis dos diversos modelos simulados para selecionar o modelo mais ajustado.

3.4.2. Critérios para a seleção do modelo

Escolher o melhor modelo é controverso, mas um bom modelo deve conseguir equilibrar a qualidade do ajuste e a complexidade, sendo esta, em geral, medida pelo número de parâmetros presentes no modelo; quanto mais parâmetros, mais complexo o modelo, sendo pois mais difícil interpretar o modelo. A seleção do "melhor" modelo torna-se então necessário.

3.4.2.1. Critério de Informação de Akaike

Akaike (1974), propôs utilizar a informação de Kullback-Leibler para a seleção de modelos. Ele estabeleceu uma relação entre a máxima verossimilhança e a informação de Kullback-Leibler desenvolvendo então um critério para estimar a informação de Kullback-Leibler, o posteriormente chamado, Critério de Informação de Akaike (AIC).

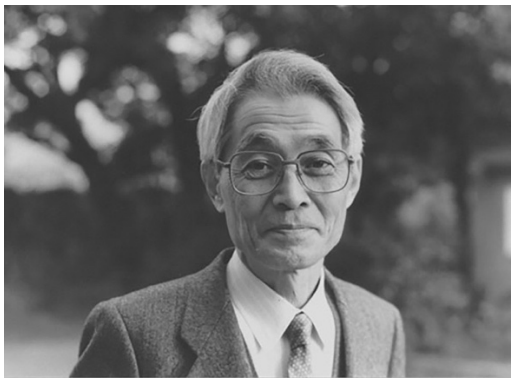

Figura 6. Hirotugu Akaike (Novembro 5, 1927 - Agosto 4, 2009).
Fonte: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1460

O Critério de Informação de Akaike realiza um processo de minimização que não envolve testes estatísticos e pode ser expresso em função do desvio do modelo, e é baseado na função de verossimilhança. Fundamenta-se no conceito de entropia, oferecendo uma medida relativa das informações perdidas, quando um determinado modelo é usado para descrever a realidade.

Com isso, Akaike (1974) definiu seu critério de informação como

$AIC = -2$ (Função suporte maximizada) + 2 (número de parâmetros),

$$AIC = -2\ln(L) + 2K \qquad (3.20)$$

onde *K* é o número de parâmetros no modelo estatístico, e *L* é o valor maximizado da função de verossimilhança para o modelo estimado.

O AIC não é uma prova sobre o modelo, no sentido de testar hipóteses, mas uma ferramenta para a seleção de modelos; não é um teste de hipóteses, não há significância e nem valor-p. Dado um conjunto de dados e vários modelos concorrentes, pode-se classificá-los de acordo com o seu AIC, quanto menor o valor de AIC, possivelmente melhor será o modelo. Não se deve atribuir um valor cima do qual um determinado modelo é "rejeitado".

Por si só, o valor do AIC para um determinado conjunto de dados não tem qualquer significado. O AIC torna-se útil quando são comparados diversos modelos.

3.4.2.2. Critério de informação bayesiano

O Critério de informação Bayesiano (BIC), também chamado de Critério de Schwarz, foi proposto por Schwarz (1978), e é um critério de avaliação de modelos definido em termos da probabilidade a posteriori.

Schwarz (1978) definiu o Critério de Informação de Bayes (BIC) como

$$BIC = -2\ln(L) + 2K.\ln(n) \qquad (3.21)$$

onde *n* é o número de observações.

Assim, os melhores modelos serão aqueles que apresentarem valores baixos destes critérios.

3.5. Testes de Significância do modelo

Após a estimação dos coeficientes do modelo, avalia-se a qualidade da estimação, ou seja, testa-se se as variáveis explicativas pertencentes ao modelo são significativas para explicar o comportamento da variável resposta.

De acordo com Paula (2013), os testes de hipóteses para os modelos lineares generalizados baseiam-se em três estatísticas: teste de razão de verossimilhança, teste de Wald e o teste de Escore.

No livro de Buse[2] (1982) podemos encontrar uma forma bastante didática para a interpretação geométrica dos testes da razão de verossimilhanças, escore e Wald. A estatística de razão de verossimilhança é o critério mais poderoso. A estatística de Wald, por exemplo é, geralmente, mais adequada em hipóteses referentes a um único coeficiente β_j.

3.5.1. Teste de Razão de Verossimilhanças

Para realizar o teste da razão da verossimilhança, compara-se os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável em questão, baseado no logaritmo da verossimilhança.

Segundo Aldrich (1997), em 1912, Fisher empregou a função de verossimilhança $f(\theta|x)$ com a ideia de que o(s) valor(es) de θ que maximizam a probabilidade dos dados observados (x) seria bom estimador de θ.

Pela teoria de estimação de máxima verosimilhança, sabe-se que os estimadores de máxima verosimilhança maximizam a função log-verosimilhança. O teste de razão de verosimilhança avalia se o valor de log-verosimilhança é suficientemente grande para concluir que as variáveis retiradas são importantes para o modelo.

[2] A. Buse é professor do Departamento de Economia da Universidade de Alberta, Edmonton, Canada.

Para um melhor entendimento, pensa-se em um valor observado da variável resposta também como sendo um valor predito resultante de um modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quanto observações.

A comparação dos valores observados com os valores preditos usando a função de verossimilhança baseia-se na seguinte expressão:

$$D = -2ln\left[\frac{Verossimilhança\ do\ Modelo\ Ajustado}{Verossimilhança\ do\ Modelo\ Saturado}\right]$$

Para assegurar a significância de uma variável independente, compara-se o valor de *D* com e sem a variável na equação. A mudança em *D* devido a inclusão da variável no modelo é obtida conforme a expressão seguinte:

$G = D$ (verosimilhança sem as m variáveis) $- D$ (verosimilhança com as m variáveis) $\langle = \rangle$

$$G = -2ln\left(L_s\right) + 2ln\left(L_c\right) \qquad (3.22)$$

em que L_S é a verossimilhança do modelo sem a covariável e L_c é a verossimilhança do modelo com a covariável.

O objetivo é testar a seguinte

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Sob a hipótese nula, a estatística *G* tem uma distribuição assintótica de um chi-quadrado com *m* grau de liberdade.

3.5.2. Teste de Wald

Para a realização do teste de Wald compara-se a estimativa de máxima verossimilhança do parâmetro $\hat{\beta}_1$ e a estimativa de seu erro padrão. Este teste é utilizado para avaliar se um

parâmetro é estatisticamente significativo e tem a distribuição normal. Assim, as hipóteses para o teste são as seguintes:

$$H_0: \hat{\beta}_j = \hat{\beta}_j^*$$

$$H_1: \hat{\beta}_j \neq \hat{\beta}_j^* \quad (j = 2, ..., k)$$

A estatística de teste é dada pela expressão

$$W = \frac{\hat{\beta}}{\sqrt{var\left(\hat{\beta}\right)}} \sim N(0,1) \qquad (3.23)$$

onde $\sqrt{var(\hat{\beta})}$ é o desvio padrão estimado do estimador $\hat{\beta}_j$. A estatística de Wald apresenta uma distribuição qui-quadrado com número de graus de liberdade igual ao número de restrições.

De acordo com Hauck e Donner (1977) a estatística de Wald se comporta de maneira estranha, em determinadas situações; frequentemente não rejeitando a hipótese nula quando o coeficiente é significativo. Com isso, é recomendado a utilização do teste da razão de verossimilhança para testar se realmente o coeficiente não é significativo quando o teste de Wald não rejeita a hipótese nula.

3.5.3. Teste Escore de Rao

Obtido a partir da função escore, o teste tem sido muito utilizado nas aplicações em Bioestatística. O teste de escore, conhecido como teste de multiplicadores de Lagrange, estima o modelo com restrições e avalia o declive da função log verosimilhança na restrição.

A estatística de teste para o teste de Escore é dado pela seguinte expressão

$$TS = \frac{\sum_{i=1}^{n} x_i \left(y_i - \bar{y} \right)}{\sqrt{\left(\bar{y}\left(1-\bar{y}\right) \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right)}} \qquad (3.24)$$

em que $\bar{y} = \hat{\pi}$ (proporção de sucessos na amostra).

No Teste Score também o objetivo é testar

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

A estatística score é útil em situações em que já se calculou um estimador restrito para β. Tem a vantagem em relação à estatística de razão de verossimilhanças de não requerer o cálculo do estimador não restrito. Além disso, tal como a estatística de Wald pode ser utilizada em modelos com parâmetro de sobredispersão, já que para o seu cálculo só há necessidade de conhecer os momentos de 1 ª e 2ª ordem.

3.6. Diagnóstico do Ajuste

A análise de diagnóstico é uma etapa fundamental no ajuste de modelos de regressão. Quando um modelo é ajustado a um conjunto de dados tem que se ter em atenção a análise de medidas das diferenças entre os seus valores observados da variável resposta, *y*, e os resíduos. Vale destacar que os resíduos têm papel fundamental na verificação do ajuste de um modelo.

Segundo McCullagh & Nelder (1989), o ajuste de um modelo a um conjunto de dados observados $\boldsymbol{y}$ pode ser encarado como uma maneira de se substituir $\boldsymbol{y}$ por um conjunto de valores estimados $\widehat{\boldsymbol{\mu}}$ para um modelo com um número de parâmetros relativamente pequeno. Logicamente os $\hat{\mu}^{'s}$ não serão exatamente iguais aos *y's*, e a questão, então, que aparece é em quanto eles diferem.

Para analisar a adequação do modelo são utilizados os resíduos de Pearson e os *Deviance* residuals. MacCullagh e Nelder (1989) definem vários tipos de resíduos, entretanto, para a família exponencial, o resíduo *deviance* é o que mais se aproxima da distribuição normal.

3.6.1. Resíduo de Pearson

Segundo Schmidt (2003), o resíduo de Pearson é capaz de comparar a distribuição observada com a determinada pelo modelo, sendo definido pela seguinte expressão

$$r\left(y_i,\hat{\pi}_i\right)=r_j=\frac{y_i-\hat{\pi}_i}{\sqrt{\hat{\pi}_j\left(1-\hat{\pi}_j\right)}},\ j=1,2,...,n \qquad (3.25)$$

Este resíduo tem a desvantagem da sua distribuição ser bastante assimétrica para modelos não-normais.

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

$$\chi^2=\sum_{i=1}^{n}\frac{\left(y_i-\hat{\pi}_i\right)^2}{V\left(\hat{\pi}_i\right)}, \qquad (3.26)$$

com distribuição Qui-Quadrado para *n-(p+1)* g.l., em que *p* é o número de covariáveis do modelo, em que $V\left(\hat{\pi}_i\right)$ é a função de variância estimada sob o modelo que está sendo ajustado. O teste é usado para testar a hipótese H_0: "O modelo ajustado é adequado".

3.6.2. *Deviance*

A **função desvio**, uma medida da distância entre o modelo saturado e o modelo corrente, avalia a qualidade de um MLG e, segundo Paula (2013) é expressa por

$$D^*\left(y;\hat{\mu}\right)=\frac{D\left(y,\hat{\pi}\right)}{a\left(\phi\right)}=2\left\{l\left(y,y\right)-l\left(\hat{\pi};y\right)\right\}, \tag{3.27}$$

em que $l\left(y,y\right)$ é o logaritmo da função de verossimilhança do modelo saturado, e $l\left(\hat{\pi};y\right)$ é o estimador de máxima verossimilhança (E.M.V) do modelo corrente. Na Tabela 3 apresentam-se as funções desvios das principais distribuições.

Tabela 3. Funções Desvios para algumas distribuições da família exponencial.

Distribuição	**Desvio**
Poisson	$D_p=2\sum_{i=1}^{n}\left[y_i ln\left(\frac{y_i}{\hat{\pi}_i}\right)+\left(\hat{\pi}_i-y_i\right)\right]$
Binomial	$D_p=2\sum_{i=1}^{n}\left[y_i ln\left(\frac{y_i}{\hat{\pi}_i}\right)+\left(m_i-y_i\right)ln\left(\frac{m_i-y_i}{m_i-\hat{\pi}_i}\right)\right]$
Binomial Negativo	$D_p=2\sum_{i=1}^{n}\left[y_i ln\left(\frac{y_i}{\hat{\pi}_i}\right)+\left(y_i+k\right)ln\left(\frac{\hat{\pi}_i+k}{y_i+k}\right)\right]$
Normal	$D_p=\sum_{i=1}^{n}\left(y_i-\hat{\pi}_i\right)^2$
Normal Inversa	$D_p=\sum_{i=1}^{n}\frac{\left(y_i-\hat{\pi}_i\right)^2}{y_i\hat{\pi}_i^2}$
Gama	$D_p=2\sum_{i=1}^{n}\left[ln\left(\frac{\hat{\pi}_i}{y_i}\right)+\frac{y_i+\hat{\pi}_i}{\hat{\pi}_i}\right]$

Adaptado de: Cordeiro e Demétrio (2013).

A *deviance* (*D*) para a regressão logística é dada pela seguinte expressão

$$d = \left(y_j, \hat{\pi}_j \right) = \pm \left[2 \left[y_j ln \left(\frac{y_j}{\hat{\pi}_j} \right) + \left(1 - y_j \right) ln \left(\frac{1 - y_j}{1 - \hat{\pi}_j} \right) \right] \right]^{1/2} \quad (3.28)$$

A estatística de teste baseada na *deviance* é

$$D = \sum_{j=1}^{n} d\left(y_j, \hat{\pi}_j \right)^2 \quad (3.29)$$

A estatística *D* é usada para testar a hipótese de que o modelo ajustado é correto. A estatística de teste tem uma distribuição assintótica χ^2 com *n-(p+1)* g.l., sendo que *p* é o número de covariáveis do modelo.

3.6.3. Teste de Hosmer-Lemeshow

Hosmer e Lemeshow (2000) propuseram um teste para verificar a qualidade de ajuste do modelo de regressão logística, que consiste no cálculo da estatística χ^2 de Pearson. O teste de Hosmer-Lemeshow testa se o modelo obtido explica de forma adequada os dados observados. Estes dados são separados em *g* grupos de acordo com as probabilidades previstas. Hosmer e Lemeshow (1980) propõe que seja utilizado *g* = 10. Os grupos são criados de maneira que o primeiro tenha probabilidade predita entre 0,0 e 0,1, e o segundo, entre 0,1 e 0,2 e assim por diante até que o décimo grupo tenha valores de probabilidade predita entre 0,9 e 1,0.

O teste avalia o modelo ajustado através das distâncias entre as probabilidades ajustadas e as probabilidades observadas.

Neste caso, a hipótese a testar é

$$\begin{cases} H_0: \text{ O modelo ajusta-se bem aos dados} \\ \text{vs} \\ H_1: \text{ O modelo não se ajusta bem aos dados} \end{cases}$$

A estatística de teste é dada pela seguinte expressão

$$C = \sum_{k=1}^{g} \frac{\left(o_k - e_k\right)^2}{e_k\left(1 - \frac{e_k}{n_k}\right)} \qquad (3.30)$$

onde

- $o_k = \sum_{j=1}^{n_k} y_{kj}$ é o número de casos registados no *k*-ésimo decil;
- $e_k = \sum_{j=1}^{n_k} \hat{\pi}_{kj}$ é o número esperado de casos no *k*-ésimo decil;
- y_{kj} e $\hat{\pi}_{kj}$ correspondem aos valores previstos e às probabilidades previstas para a observação *j* no grupo *k* de decil de risco.

O teste de Hosmer e Lemeshow segue, aproximadamente, uma distribuição qui-quadrado com *g-2* g.l., quando o modelo está especificado corretamente. Hosmer e Lemeshow (2000) verificaram, através de simulações, que a distribuição nula assintótica de C pode ser bem aproximada por uma distribuição qui-quadrado com $(g - 2)$ graus de liberdade.

3.6.4. Predição – Curva ROC

Quando a variável resposta é binária é necessário escolher uma regra de predição, visto que, a probabilidade estimada $\hat{\pi}$ está compreendida entre 0 e 1. Intuitivamente, pode-se pensar que se $\hat{\pi}_i$ for grande, $\hat{Y}_i = 1$, caso contrário $\hat{Y}_i = 0$. Mas como determinar o ponto que para os valores acima dele o indivíduo seja classificado como "evento" $\left(\hat{Y}_i = 1\right)$, e valores abaixo dele o indivíduo é classificado como "não evento" $\left(\hat{Y}_i = 0\right)$? Esse ponto é conhecido como ponto de corte. O valor do ponto de corte é escolhido arbitrariamente pelo pesquisador entre os valores possíveis para a variável de decisão, acima da qual o paciente é classificado positivo e abaixo

do qual é classificado como negativo. A curva ROC (Receiver Operating Characteristic) é bastante utilizada para determinar esse ponto.

A curva ROC é muito utilizada na medicina para avaliar o desempenho de diagnósticos médicos, em que se procura identificar a presença ou ausência de certa doença, associando a respetiva probabilidade de erro. Em termos estatísticos, esta é uma ferramenta que permite avaliar a capacidade preditiva de um modelo de regressão. A análise do desempenho do modelo pode ser realizada por meio de um gráfico simples e robusto, que fornece a variação de dois indicadores importantes do desempenho para diferentes para diferentes valores de corte, nomeadamente a Sensibilidade e a Especificidade do modelo (ou do teste diagnóstico), conceitos que serão apresentados a seguir.

No Figura 7 é mostrado um exemplo da Curva ROC.

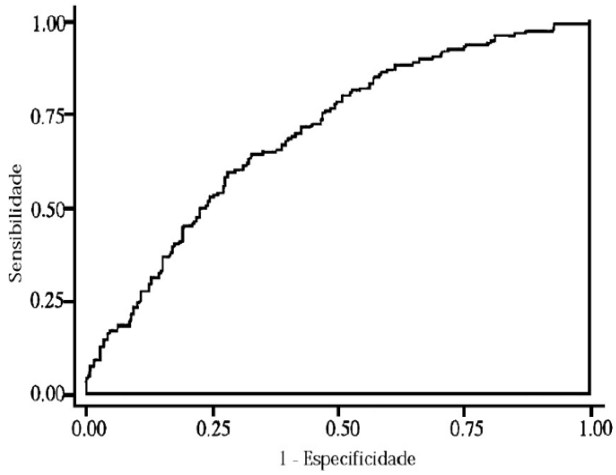

Figura 7. Exemplo Curva ROC.

Fonte: *(Nakamura, 2013)*

3.6.4.1. Área abaixo da Curva ROC

A área abaixo da curva ROC (*auc*) é uma forma bastante generalizada de aferir uma medida de discriminação entre os indivíduos que apresentam a característica de interesse versus aqueles

que não apresentam, e o seu valor varia entre 0 e 1. Algumas literaturas indicam que *auc* nunca deve ser inferior a 0,5.

Seja *auc*, o valor que corresponde à área abaixo da curva de ROC, como regra geral temos as seguintes linhas de orientação:

Se *auc* = 0, 5 não há discriminação;

Se 0,7≤ *auc* <0,8 a discriminação é aceitável;

Se 0,8≤ *auc* <0,9 Discriminação excelente;

Se *auc* ≥ 0,9 Discriminação excecional.

Após o ajuste de um modelo e a determinação do ponto de corte, é importante avaliar o poder de discriminação do modelo, isto é, discriminar os eventos dos não eventos. Para tal foram criados parâmetros numéricos cuja denominação é a seguinte: Capacidade preditiva Sensibilidade, Especificidade, Verdadeiro Preditivo Positivo e Verdadeiro Preditivo Negativo

Tabela 4. Representação geral de um teste diagnóstico/Matriz de confusão.

	Positivo (+)	**Negativos (-)**
Positivo (+)	VP Verdadeiros positivos	FP Falsos positivos
Negativos (-)	FN Falsos negativos	VN Verdadeiros negativos
Total	VP+FN	FP+VN

Adaptado de: (Silva, 2011).

Acurácia: indica uma performance geral do modelo. Dentre todas as classificações, a proporção de predições corretas realizadas pelo modelo;

$$AC = \frac{VP + VN}{VP + VN + FP + FN} \qquad (3.31)$$

Precisão: indica dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas;

$$Pr = \frac{VP}{VP + FP}$$

Sensibilidade: probabilidade de o teste gerar um resultado positivo, dado que o indivíduo é realmente portador da "patologia", ou seja, a proporção de verdadeiros positivos. Isto é, a avaliação da capacidade do modelo em classificar um indivíduo como evento $\hat{Y} = 1$ dado que realmente ele é evento (Y=1):

$$Sens. = \frac{VP}{VP + FN} \qquad (3.32)$$

Especificidade: probabilidade de o teste gerar um resultado negativo quando o indivíduo não é portador da "patologia", ou seja, proporção de verdadeiros negativos. Isto é, a avaliação da capacidade de o modelo predizer um indivíduo como não evento $\hat{Y} = 0$ dado que ele realmente é não evento (Y=0).

$$ESPEC = \frac{VN}{VN + FP} \qquad (3.33)$$

A Sensibilidade e a Especificidade não são calculadas usando os mesmos indivíduos, ou seja, enquanto Sensibilidade usa apenas os "doentes", a Especificidade utiliza os "não doentes", assim, Sensibilidade e Especificidade são medidas independentes entre si.

Verdadeiro preditivo positivo: É a proporção de verdadeiros positivos em relação a todas as predições positivas, isto é, o indivíduo ser evento (Y=1) dado que o modelo o classificou como evento $\hat{Y} = 1$.

$$VPP = \frac{VP}{VP + FP} \tag{3.34}$$

Verdadeiro preditivo Negativo: É a proporção de verdadeiros negativos em relação a todas predições negativas, ou seja, o indivíduo ser não evento (Y=0) dado que o modelo o classificou como não evento $\hat{Y} = 0$.

$$VPN = \frac{VN}{VN + FN} \tag{3.35}$$

3.6.4.2. Comparação de Modelos

Os gráficos que representam duas ou mais curvas ROC associadas a diferentes testes de diagnósticos permitem uma comparação de desempenhos entre eles.

A razão crítica é um método utilizado para verificar se as diferenças entre duas áreas abaixo das curvas ROC provenientes de amostras independentes são significativas.

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2}} \sim N(0,1) \tag{3.36}$$

onde A_1 e A_2 correspondem as áreas e SE_1 e SE_2 correspondem aos erros estimados para a curva ROC, respetivamente para os testes diagnósticos 1 e 2. A aproximação à estatística de *Wilcoxon-Mann-Whitney* é usada para determinar as áreas e os respetivos erros padrão.

Os erros padrão associados às áreas, podem ser obtidos através da seguinte expressão:

$$SE(A) = \sqrt{\frac{A(1-A) + (n_A - 1)(Q_1 - A^2) + (n_N - 1)(Q_2 - A^2)}{n_A n_N}} \tag{3.37}$$

onde Q_1 é a probabilidade de duas observações anormais, aleatoriamente escolhida serem classificadas com maior desconfiança do que uma observação normal aleatoriamente escolhida, e Q_2 corresponde à probabilidade de uma observação anormal, aleatoriamente escolhida ser classificada com maior desconfiança do que duas observações normais aleatoriamente escolhidas e n_A e n_N correspondem, respetivamente à dimensão dos pacientes anormais e normais.

3.6.5. Distância de Cook

A distância de Cook[3] (1977), originalmente desenvolvida para modelos normais lineares, foi rapidamente assimilada e estendida para diversas classes de modelos.

Figura 8. Dennis Cook.
Fonte: http://users.stat.umn.edu/~rdcook/

Em 1986, Cook apresentou uma proposta inovadora na área de diagnóstico em modelos de regressão, que propõe avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo ou nos dados, ao invés da avaliação pela remoção individual ou conjunta de pontos. Essas propostas seriam inicialmente derivadas para o modelo normal linear clássico sendo estendidas em seguida para os MLGs.

[3] **Cook** é professor no School of Statistics, University of Minnesota - Twin Cities.

Alguns autores afirmam que a distância de Cook é uma métrica da distância entre a estimativa β pelo método dos mínimos quadrados baseada nas *n* observações, e a estimativa obtida quando o i-ésimo ponto for removido.

A distância de Cook é dada por

$$D_{(i)} = \frac{\boldsymbol{h_i}}{(1-\boldsymbol{h_i})^2}\frac{1}{p}rsi_i^2 \qquad (3.38)$$

onde h_i representa o elemento da diagonal da matriz de projeção *H*, dada por $H = X(X^T X)^{-1}X^T$, *p* o numero de parâmetros e rsi_i^2 o resíduo para a i-ésima observação.

3.6.6. Técnicas Gráficas

De acordo com Turkman e Silva (2000), o uso de representações gráficas é uma ferramenta informal, porém bastante utilizada e útil, na análise de resíduos. Através delas podemos encontrar desvios tanto no componente aleatório como no componente sistemático. Os gráficos variam conforme os desvios que se pretende encontrar. Os mais recomendados são os seguintes:

a. Gráfico do desvio residual versus a ordem das observações, ou versus os valores ajustados. Este gráfico permite identificar observações consideradas *outliers*, observações que estão fora do limite considerado para a distribuição dos resíduos;
b. Gráfico normal de probabilidades para resíduos com envelope permite avaliar o pressuposto da normalidade dos resíduos e da escolha da distribuição para a variável resposta. No caso da regressão logística é mais útil para avaliar se o modelo em análise é ou não adequado. Se o modelo ajustado é o correto, existe grande probabilidade de que todos os pontos estejam dentro do envelope.
c. Gráfico de h_i e $D_{(i)}$ contra a ordem das observações para identificar as observações influentes.

3.7. Testes de Autocorrelação

3.7.1. Análise de Colinearidade e Multicolinearidade

A escolha de um determinado método múltiplo é determinada segundo os objetivos da investigação a ser realizada. Quando trabalhamos com mais de uma variável regressora, é muito importante verificar se essas variáveis explicativas são correlacionadas. Desta forma, se não existir nenhum tipo de correlação entre elas, dizemos que são ortogonais.

Na prática, é muito difícil que as variáveis independentes não tenham nenhuma correlação, no entanto, a correlação não é impeditiva da aplicação da regressão, desde que as variáveis não sejam muito correlacionadas, caso contrário, as inferências baseadas no modelo de regressão podem ser errôneas ou pouco confiáveis.

Por isso, é necessário verificar se as variáveis são altamente correlacionadas. Algumas literaturas afirmam que os termos Colinearidade (Multicolinearidade) são utilizados para indicar a existência forte de correlação entre duas (ou mais) variáveis independentes. A Figura 9 mostra a visualização gráfica da presença ou não de multicolinearidade em um conjunto de dados.

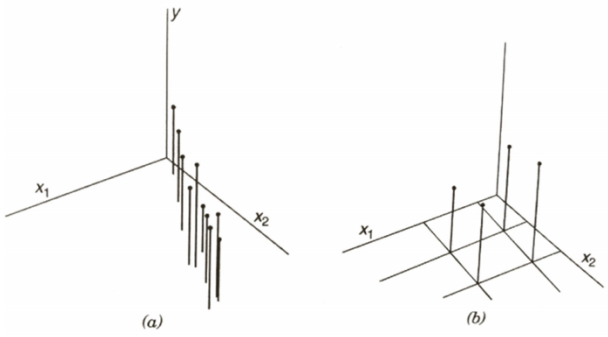

Figura 9. (a) Presença de multicolinearidade; (b) Ausência de multicolinearidade.
Fonte: (Montgomery, Peck, & Vining, 2001)

Conforme Montgomery et al. (2001), existem quatro fontes primárias de multicolinearidade:

a. Método de coleta de dados utilizado: Quando o pesquisador coleta amostras somente no subespaço da região dos regressores definidos.

b. Restrições no modelo ou na população: Pode haver alguma razão para que haja a restrição do modelo. Por exemplo, um modelo para avaliar o consumo de energia elétrica em função da renda e do tamanho da casa, há uma restrição física, pois famílias com maiores casas têm maior renda.
c. Escolha do modelo: as variáveis escolhidas para compor o modelo podem ser linearmente dependentes, causando assim multicolinearidade.
d. Modelo com excesso de termos: O modelo possui mais variáveis regressoras do que observações. Estes modelos são comumente encontrados em pesquisas médicas. Neste caso é comum eliminar algumas variáveis regressoras para lidar com a multicolinearidade.

Quando detetada a multicolinearidade, algumas literaturas sugerem a **Remoção das Variáveis**, porém, esta ação não ajuda a avaliar os efeitos da variável independente, pois nenhuma informação é obtida sobre a variável removida, e também porque o valor do coeficiente de regressão para a variável independente remanescente no modelo é afetado pelas variáveis independentes correlacionadas não incluídas no modelo, outros sugerem ainda a **Ampliação da Amostra** ou a **Ridge Regression** que consiste no uso de estimadores tendenciosos para os coeficientes ou ainda os **Componentes Principais** que permite que todas as variáveis independentes participam de certa forma do modelo, assim é possível reduzir um grande número de variáveis independentes em um número razoavelmente pequeno de novas variáveis independentes, que são chamadas de componentes e são determinadas pela combinação linear das variáveis originais. Estas novas variáveis (ou componentes principais) são não correlacionadas e são usadas para integrar o modelo de regressão.

3.7.2. Fator de Inflação da Variância (VIF)

Supondo que as variáveis estão centradas e padronizadas, tem-se que $R = (\boldsymbol{X^T X})^{-1}$ em que os elementos da diagonal dessa matriz são chamados de fatores de inflação de variância (VIF) e representam o incremento da variância devido à presença de multicolinearidade (Montgomery, Peck, & Vining, 2001).

O VIF pode ser calculado pela seguinte expressão:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad j = 1, 2, \dots, p \tag{3.39}$$

Em que p é o número das variáveis preditoras, R_j^2 é o coeficiente de correlação múltipla, resultante da regressão de X_j nos outros *p-1* regressores. De acordo com Montgomery et al. (2001), valores de VIF maiores do que 5 indicam multicolinearidade moderada e valores de VIF maiores do que 10 implicam em multicolinearidade severa.

3.7.3. Diagnóstico de Homocedasticidade

O termo homocedasticidade é utilizado para designar variância constante dos erros experimentais para observações distintas.

3.7.3.1. Teste de Breusch-Pagan

O teste de Breusch-Pagan é baseado no multiplicador de Lagrange. O teste é utilizado para testar a hipótese nula de que as variâncias dos erros são iguais (homocedasticidade) versus a hipótese alternativa de que pelo menos uma variância é diferente. A estatística de teste de BP é obtida da seguinte forma:

Inicialmente, ajustamos o modelo de regressão linear (simples ou múltiplo) e encontramos os resíduos $e_i = \hat{e}_i = y_i - \hat{y}_i$ e os valores ajustados $\hat{y}_i = (\hat{y}_1, \dots, \hat{y}_n)$. Em seguida, consideramos os resíduos ao quadrado e padronizamos de modo que a média do vetor de resíduos padronizados, que denotaremos por u, seja 1. Esta padronização é feita dividindo cada resíduo ao quadrado pela SQE/n em que SQE é a Soma de Quadrados dos Resíduos do modelo ajustado e n é o número de observações. Desta forma, temos que cada resíduo padronizado é dado por

$$u_i = \frac{e_i^2}{SQE/n}, \qquad i = 1, \dots, n \tag{3.40}$$

Por fim, realiza-se regressão entre $u = (u_1, \dots, u_n)$ (variável resposta) e o vetor $\hat{y}$ (variável explicativa) e obtemos a estatística do teste χ^2 calculando a Soma de Quadrados da Regressão de u sobre $\hat{y}$ e dividindo o valor encontrado por 2. Sob a hipótese nula, esta estatística tem distribuição Qui-quadrado com 1 grau de liberdade.

3.7.4. Teste de Durbin-Watson

A estatística Durbin-Watson é usada para testar a presença de autocorrelação nos erros de um modelo de regressão. A autocorrelação significa que os erros de observações adjacentes são correlacionados. As hipóteses do teste são as seguintes:

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$$

A estatística de teste de Durbin Watson (D_W) envolve o cálculo de um teste estatístico baseado nos resíduos do método de regressão de mínimos quadrados, e é dada pela expressão

$$D_W = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (3.41)$$

em que $0 \leq D_W \leq 4$, onde $e_i = y_i - \hat{y}_i$ e y_i e $\hat{y}_i$ são, respetivamente, os valores observados e preditos para a variável resposta.

3.8. Modelo de Regressão de Poisson

O modelo de Poisson se torna muito importante quando desejamos analisar dados em forma de contagens, pelas características que possui. O modelo é um tipo específico dos Modelos Lineares Generalizados e não lineares, com origem por volta de 1970, quando Wedderburn (1974) desenvolveu a teoria da quase verossimilhança. A teoria foi analisada posteriormente com mais detalhes por McCullagh (1983).

Conforme Cordeiro (1986), as principais características do modelo de regressão de Poisson são:

- Proporciona, em geral, uma descrição satisfatória de dados experimentais cuja variância é proporcional à média;
- Pode ser deduzido teoricamente de princípios elementares com um mínimo de restrição;
- Se eventos ocorrem independente e aleatoriamente no tempo, com taxa média de ocorrência constante, o modelo determina o número de eventos, num intervalo de tempo especificado. O modelo de Poisson desempenha na análise de dados categorizados, o mesmo papel.

O modelo de Poisson desempenha na análise de dados categorizados, o mesmo papel que o modelo normal ocupa na análise de dados contínuos. A diferença primordial é que a estrutura multiplicativa para as médias do modelo de Poisson é mais apropriada do que a estrutura aditiva das médias do modelo normal (Cordeiro G. , 1986).

A Regressão de Poisson foi introduzida por Siméon Denis Poisson em 1837, conhecida ainda como Modelo Log-Linear de Poisson, faz parte da família de Modelos Lineares Generalizados.

Figura 10. Siméon-Denis Poisson (21 de Junho 1781 – 25 de Abril de 1840)
Fonte: https://en.wikiquote.org/wiki/Sim%C3%A9on_Denis_Poisson

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. A sua função de probabilidades é expressa como

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \text{ para } y = 0,1,2,... \qquad (3.42)$$

onde y é o valor observado da variável resposta Y, com $\mu > 0$, sendo o parâmetro desconhecido, valor médio da variável Y. Uma das suas propriedades é que seu parâmetro corresponda ao seu valor esperado e sua variância $E(Y) = Var(Y) = \mu$.

A distribuição de Poisson pertence à família exponencial com os seguinte elementos $\phi = 1$, $\theta = ln(\mu)$, $b(\theta) = e^{\theta}$, $c(y, \phi) = -ln(y!)$, $\mu(\beta) = e^{\theta}$ e $V(\mu) = \mu$. Dos elementos citados, considera-se que o parâmetro canônico se relaciona com a média mediante a transformação logarítmica. Com isso,

$$ln(\mu_i) = \mathrm{x}_i'\beta,$$

em que

$$E(Y_i|X_i) = e^{\boldsymbol{x}_i'\boldsymbol{\beta}},$$

sendo $x_i = (1, x_{i,1}, ..., x_{i,k})$' o vetor de covariáveis e β o vetor $(k+1)$ de parâmetros de regressão. Assim a log-verossimilhança é expressa como

$$l(\beta) = \sum_{1}^{n}\left[y_i\boldsymbol{x}_i'\boldsymbol{\beta} - exp(\boldsymbol{x}_i'\boldsymbol{\beta}) - \ln(y_i!)\right] \qquad (3.43)$$

Em casos práticos, é comum encontrar dados de contagens que apresentem variância menor / maior que a média, o que pode limitar a aplicação do modelo Poisson. Quando o modelo é utilizado para dados não equidispersos resulta em erros padrões não confiáveis o que acarreta em inferências incorretas. Por este motivo, outros modelos devem ser considerados.

3.8.1. Estimação dos Coeficientes do Modelo

Para estimar os parâmetros, utiliza-se o método de estimação de máxima verossimilhança, assim, o logaritmo da verosimilhança para o modelo de Poisson é dado por

$$l(\beta)=\sum_{i=1}^{n}\left[\left(y_i ln\left(\mu(x_i)\right)\right)-\mu(x_i)-ln(y_i!)\right], \tag{3.44}$$

substituindo $ln(\mu(x_i))$ e $\mu(x_i)$ por $(\mathrm{x}'\beta)$ e $\left(e^{\mathrm{x}'\beta}\right)$ respetivamente, obtém-se

$$l(\beta)=\sum_{i=1}^{n}\left[y_i\beta_1+y_i\beta_2 x_{i2}+\ldots+y_i\beta_p x_{ip}-e^{\beta_1+\beta_2 x_{i2}+\beta_p x_{ip}}-ln(y_i!)\right]$$

Para encontrar a estimativa dos parâmetros é necessário a utilização de métodos iterativos.

3.8.2. Qualidade do Ajuste

Como forma de avaliar a qualidade de ajuste do modelo de Poisson com p parâmetros independentes aos dados y_1, …, y_n, utiliza-se a medida *AIC*, a razão de verossimilhanças e o Qui-Quadrado de Pearson.

A *deviance* para o modelo de regressão de Poisson é dada pela seguinte expressão

$$D=2\sum_{j=1}^{n}\left(y_j ln\left(\frac{y_j}{\hat{\pi}_j}\right)-\left(y_j-\hat{\pi}_j\right)\right) \tag{3.45}$$

O Qui-Quadrado de Pearson para a regressão de Poisson é dada por

$$\chi^2=\sum_{j=1}^{n}\frac{\left(y_j-\hat{\pi}_j\right)^2}{\hat{\pi}_j}. \tag{3.46}$$

Com isso, ainda podemos calcular a percentagem da *deviance*, ou seja, a variabilidade nos dados explicados pelo modelo de Poisson, através da expressão seguinte

$$\left(1-\frac{D\left(\text{Modelo ajustado}\right)}{D\left(\text{Modelo nulo}\right)}\right)\times 100. \qquad (3.47)$$

3.9. Modelo Binomial Negativo

A distribuição binomial negativa foi desenvolvida por Bernoulli em seu tratado Ars Conjectand, publicado em 1713.

O modelo binomial negativo tem diferentes parametrizações, sendo uma generalização da distribuição de Poisson. Algumas representam o número total de tentativas até obter *r* sucessos em uma sequência de ensaios binomiais, outras como o número total de falhas até obter *r* sucessos em uma sequência de ensaios binomiais. Utilizaremos a distribuição binomial negativa como uma distribuição discreta de propósito geral que pode ser usado para modelar valores inteiros.

A sua função densidade de probabilidade é dada por

$$f\left(y;\mu,k\right)=\frac{\Gamma\left(k+y\right)}{\Gamma\left(k\right)y!}\frac{\mu^{y}k^{k}}{\left(\mu+k\right)^{k+y}}, \qquad (3.48)$$

em que $k > 0$, $\mu > 0$ e $y = 0,1, \ldots$ Assim a função pode ser reescrita como

$$\begin{aligned} f\left(y;\mu,k\right) &= exp\left[ln\left(\frac{\Gamma\left(k+y\right)}{\Gamma\left(k\right)y!}\right)+y\ ln\left(\mu\right)+k\ ln\left(k\right)-\left(k+y\right)ln\left(\mu+k\right)\right] \\ &= exp\left[y\left(ln\left(\mu\right)-ln\left(\mu+k\right)\right)+k\left(ln\left(k\right)-ln\left(\mu+k\right)\right)+ln\left(\frac{\Gamma\left(k+y\right)}{\Gamma\left(k\right)y!}\right)\right]. \end{aligned}$$

Utilizando a notação da família exponencial:

$$\begin{cases} \phi = 1 \\ \theta = ln\left(\dfrac{\mu}{\mu + k}\right) \\ b(\theta) = -k\ ln\left(1 - e^{\theta}\right) \\ c(y, \phi) = ln\left(\dfrac{\Gamma(k + y)}{\Gamma(k)\, y!}\right) \end{cases}$$

A esperança e a variância são dadas por

$$E(Y) = \frac{k\ e^{\theta}}{1 - e^{\theta}}, \tag{3.49}$$

$$Var(Y) = \frac{k\ e^{\theta}}{\left(1 - e^{\theta}\right)^2}. \tag{3.50}$$

O modelo de regressão com resposta binomial negativa pode ser especificado da seguinte forma:

$$E(Y_i | \boldsymbol{x_i}) = exp(\boldsymbol{x_i'\beta}) \tag{3.51}$$

em que $x_i = (1, x_{i,1}, ..., x_{i,k})$ é o vetor de covariáveis e β o vetor $(k+1)$ de parâmetros de regressão.

Aplicando a transformação logarítmica, o modelo de regressão Binomial Negativa é expresso como

$$ln(\mu_i) = \mathrm{x}_i'\beta. \tag{3.52}$$

3.9.1. Estimação dos Coeficientes do Modelo

Para estimar os coeficientes de regressão é utilizado o método da máxima verossimilhança. O logaritmo do método para o modelo de regressão Binomial Negativa é expresso por

$$l(\beta)=\sum_{i=1}^{n}\left(y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)-\left(\frac{1}{\alpha}\right)ln\left(1+\alpha\mu_i\right)+ln\left(\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma\left(y_i+1\right)\Gamma\left(\frac{1}{\alpha}\right)}\right)\right). \tag{3.53}$$

Para encontrar a estimativa dos parâmetros (α e β) é necessário a utilização de métodos iterativos.

3.9.2. Qualidade de Ajuste

Como forma de avaliar a qualidade de ajuste do modelo de regressão Binomial Negativa com p parâmetros independentes aos dados y_1, …, y_n, utilizam-se as mesmas estatísticas utilizadas para o modelo de Poisson.

A *deviance* para o modelo de regressão Binomial Negativa é expressa por

$$D=2\sum_{j=1}^{n}\left[y_j ln\left(\frac{y_j}{\hat{\pi}_j}\right)-\left(\frac{1}{\alpha}+y_j\right)ln\left(\frac{1+\alpha y_j}{1+\alpha\hat{\pi}_j}\right)\right]. \tag{3.54}$$

O Qui-Quadrado de Pearson para o modelo de regressão Binomial Negativa é dado pela expressão

$$\chi^2=\sum_{j=1}^{n}\frac{\left(y_j-\hat{\pi}_j\right)^2}{\hat{\pi}_j+\alpha\hat{\pi}_j^2}. \tag{3.55}$$

CAPÍTULO 4

4. Modelo de Regressão Logística

4.1. Introdução

A regressão logística é um modelo linear generalizado, muito utilizada para estudar variáveis *dummys*, que são aquelas que são compostas apenas por duas opções de eventos, como "sim" ou "não", e tem-se constituído num dos principais métodos de modelação estatística de dados.

Conforme Fávero et al. (2009), a regressão logística é uma técnica estatística utilizada para descrever o comportamento entre uma variável dependente binária e variáveis independentes métrica e não métricas. 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. De acordo com Belfiore (2015), a variável dependente segue uma distribuição de Bernoulli. Fávero et al. (2009) afirma que as variáveis independentes, tanto podem ser categóricas ou não. As variáveis categóricas são aquelas variáveis que podem ser mensurados usando um número limitado de valores ou categorias. A regressão logística permite-nos estimar a probabilidade de um determinado evento ocorrer em face de um conjunto de variáveis explanatórias.

Hosmer e Lameshow (2000) apontam pelo menos duas razões para a utilização do modelo logístico na análise de variáveis-resposta dicotómicas:

- De um ponto de vista matemático, é bastante flexível e fácil de ser utilizado;
- Permite uma interpretação de resultados bastante rica e direta.

4.2. Função Logit

De acordo com Cramer (2003), os primeiros trabalhos publicados sobre logit foram feitos no final das décadas de 1950 e 1960 em estatística e epidemiologia; na estatística havia uma

vantagem analítica na transformação do logit em lidar com saídas binárias, uma vez, que todos os cálculos eram realizados a mão. Na epidemiologia o estudo do logit se deu ainda mais cedo (1950), uma vez que estava diretamente ligada à razão de chances de probabilidades. Corrar et al. (2009) acrescentam que essa técnica foi desenvolvida para tentar realizar predições ou tentar explicar a ocorrência de determinados fenômenos quando a variável dependente é de natureza binária.

Para Corrar et al. (2009) um dos motivos que as funções de ligação vêm sendo largamente utilizadas, para realizar predições quando a variável dependente é dicotômica, é devido ao pequeno número de restrições que são elas: incluir todas as variáveis para que se obtenha maior estabilidade; valor esperado do erro deve ser zero; inexistência de autocorrelação entre os erros; inexistência de correlação entre os erros e as variáveis independentes e; ausência de multicolinearidade perfeita entre as variáveis independentes.

Cramer (2003) e Corrar et al. (2009) acrescentam que existe um problema quando não se tem variáveis independentes normais no caso linear, mas como a variável dependente é do tipo dicotômica (com distribuição de Bernoulli) e no caso das funções de ligação logit e probit não há essa restrição. Os autores ainda informam que quando se trabalha com o logit, devem-se obter amostras maiores que no caso linear, mas essas funções de ligação possuem a vantagem de acolher mais facilmente variáveis dependentes binárias.

O logit, também chamado de logito, equivale ao logaritmo natural da chance, dada pela seguinte expressão:

$$logit(p) = log\left(\frac{p}{1-p}\right) \qquad (4.1)$$

Pelo gráfico seguinte é possível observar que a função tende para zero, mas não toca o eixo de y, e o mesmo ocorre quando tende para 1, ou seja, a função está no intervalo de 0 a 1.

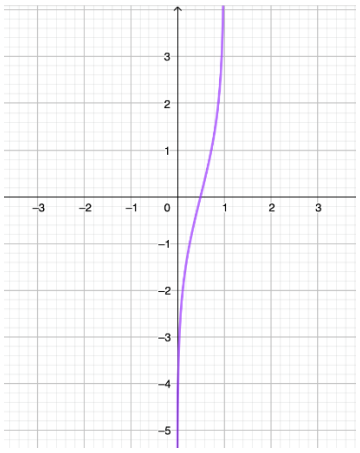

Figura 11. Gráfico da função logit (p).
Fonte: Elaborado pelo autor.

A função logística é dada pelo logito-inverso (anti-logit) que nos permite transformar o logito em probabilidade:

$$p = \frac{exp(x)}{1 + exp(x)} \qquad (4.2)$$

4.3. Odds Ratio

O odds ratio é dado pelo quociente entre a odds do acontecimento de interesse ocorrer (Y = 1) nos indivíduos com $x = 1$ e a odds desse acontecimento ocorrer nos indivíduos com $x = 0$. O odds do acontecimento de interesse ocorrer nos indivíduos com $x = 1$ é definida por $\frac{\pi(1)}{1-\pi(1)}$. Analogamente, a odds do acontecimento de interesse ocorrer nos indivíduos com $x = 0$ é definida por $\frac{\pi(0)}{1-\pi(0)}$. Assim, o odds ratio é uma forma de comparar se a probabilidade do acontecimento de interesse ocorrer é a mesma para os indivíduos com $x = 1$ ou $x = 0$.

$$\text{Odds} = \frac{\text{Probabilidade de um evento ocorrer}}{\text{Probabilidade de um evento não ocorrer}}$$

A medida de associação Odds Ratio (OR), denominada razão de chances, é utilizada usualmente na regressão logística univariável para complementar o teste à significância da covariável (x).

A codificação de *x* permite a interpretação trivial dos parâmetros. Na prática é usado a tabela de contingência para o cálculo de odds ratio. As probabilidades do evento de interesse ocorrer para as duas categorias de *x*, são dadas pelas expressões

$$\pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \qquad \pi(0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \tag{4.3}$$

Consequentemente, o valor do odds ratio é dado pela expressão

$$OR = \frac{\pi(1)[1 - \pi(0)]}{\pi(0)[1 - \pi(1)]} \tag{4.4}$$

Pontos importantes sobre Odds Ratio:

- OR > 1 indica que a probabilidade de o evento ocorrer é maior do que a probabilidade do evento não ocorrer.
- OR < 1 indica uma diminuição na ocorrência do evento.
- OR = 1 indica que a probabilidade de o evento ocorrer não afeta a probabilidade do evento não ocorrer.
- O intervalo de confiança (IC) e o *p-value* indica a significância do valor.

Contudo, em situações do quotidiano é um pouco diferente, por exemplo, um OR maior que 1 não significa um resultado estatisticamente significante. Para confirmar a significância do resultado é necessário considerar o intervalo de confiança e o *p-value*. No entanto, este resultado pode não aplicar a toda a população, uma vez que existem vários fatores a serem considerados.

4.4. Regressão logística simples

A regressão logística simples é usada para o caso de regressão com uma variável explicativa.

Sendo x_i uma variável explicativa e y_i o número de ocorrências de um certo evento, onde $i=1,2,...,n$ seja o número de observações, e assumindo que a variável resposta tenha uma distribuição binomial $\left(Y_i \sim B(\pi_i)\right)$ com $\pi_i = E(Y_i)$, então

$$P\left[Y_i = y_i\right] = \pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i} \tag{4.5}$$

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

$$\pi\left(x_i\right) = \frac{e^{\beta_0+\beta_i x_i}}{1+e^{\beta_0+\beta_i x_i}},\ i = 1,...,n \tag{4.6}$$

Para facilitar a obtenção dos parâmetros (β_0, β_i) 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* de probabilidade, que é dada como

$$g\left(x\right) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_i x_i \tag{4.7}$$

4.4.1. Parâmetros do Modelo Simples

De acordo com Hair et al. (2006), a regressão linear utiliza dos métodos dos mínimos quadrados ordinários para realizar a estimação de seus coeficientes, esse método consiste em minimizar a soma de quadrados das diferenças entre os valores observados e os previstos. Na regressão não linear o método da máxima verossimilhança é utilizado de forma iterativa para que sejam encontradas as estimativas mais prováveis dos parâmetros. Ao invés de minimizar os desvios quadrados, a regressão não linear maximiza a probabilidade de que um evento ocorra.

Para Casella e Berger (2010) quando se usa regressão linear, a técnica de mínimos quadrados é uma opção para o cálculo dos estimadores; nos modelos não lineares não há uma conexão direta entre a variável dependente (Y_i) e o componente sistemático $(\beta_0+\beta_1 x_i)$, assim o método dos mínimos quadrados não é mais uma opção, sendo a estimação realizada por meio do 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 pode ser escrita da seguinte forma

$$P\left[Y_i = y_1,\ldots,y_n \middle| \beta_0,\beta_1\right] = \quad (4.8)$$
$$= \prod_{i=1}^{n} \pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i}$$
$$= \prod_{i=1}^{n}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}\left(1-\pi_i\right)^{1-y_i}$$

Aplicando logaritmo dos dois lados, a expressão fica

$$\ln L(\beta_0,\beta_1|(x_i,y_i)) = \sum_{i=1}^{n} y_i\,(\beta_0+\beta_1 x_i) - \sum_{i=1}^{n}\ln\left(1+e^{\beta_0+\beta_1 x_i}\right) \quad (4.9)$$

Os estimadores de máxima verossimilhança para os parâmetros β_0,β_1 são os valores $\hat{\beta}_0,\hat{\beta}_1$ 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

$$\frac{\partial}{\partial\beta_0}\ln L\left((\beta_0,\beta_1)\middle|(x_i,y_i)\right) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \quad (4.10)$$

$$\frac{\partial}{\partial \beta_1} \ln L\big((\beta_0, \beta_1)\big|(x_i, y_i)\big) = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \qquad (4.11)$$

Segundo Casella e Berger (2010), o método da máxima verossimilhança é definido como sendo os valores dos parâmetros que geram, com maior frequência, a amostra observada. Para a realização do procedimento, deve-se maximizar a função de verossimilhança com relação à $\hat{\beta}$ assim iguala-se a zero as derivadas parciais da função de verossimilhança e determinar $\hat{\beta}$ que solucione o conjunto de equações.

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}} = 0$$
$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}} = 0$$

Por serem não-lineares, estas equações são resolvidas recorrendo métodos iterativos, como por exemplo Newton-Raphson, e o resultado 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

$$I\left(\hat{\beta}\right) = \begin{bmatrix} \sum_{i=1}^{n} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{\left(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}\right)^2} & \sum_{i=1}^{n} x_i \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{\left(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}\right)^2} \\ \sum_{i=1}^{n} x_i \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{\left(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}\right)^2} & \sum_{i=1}^{n} x_i^2 \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{\left(1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}\right)^2} \end{bmatrix} \qquad (4.12)$$

Após obter as estimativas dos parâmetros do modelo podemos calcular as probabilidades estimadas

$$\pi_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}} \qquad (4.13)$$

A interpretação dos parâmetros da regressão logística é semelhante aos da regressão linear, usando a função odds ratio – OR (Razão de chances).

Seja a função de ligação

$$g(x) = \frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x}$$

Ao tomarmos dois valores distintos da variável explicativa x_j e x_{j+1}, obtemos

$$OR = \frac{g(x_{j+1})}{g(x_j)} = \frac{e^{\beta_0+\beta_1 x_{j+1}}}{e^{\beta_0+\beta_1 x_j}} \qquad (4.14)$$

Assim, temos que

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

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

$$ln(OR) = ln(e^{\beta_1}) = \beta_1 \qquad (4.15)$$

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

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

4.4.2. 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 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$, ou seja, 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 $\hat{\sigma}^2(\hat{\beta}_j)$. Os demais elementos da matriz I^{-1} são as covariâncias entre $(\hat{\beta}_j\ \hat{\beta}_u)$ com $j \neq u$.

Com isso, o desvio padrão é definido como

$$D\hat{P}\left(\hat{\beta}_j\right) = \sqrt{\hat{\sigma}^2\left(\hat{\beta}_j\right)} \qquad (4.16)$$

4.4.3. Intervalos de Confiança

A formulação das estimativas do intervalo de confiança para os parâmetros baseia-se na mesma teoria estatística utilizada para a elaboração de testes de significância do modelo. Os intervalos de confiança para a inclinação e intercepto baseia-se nos testes de Wald ao nível de significância $100(1-\alpha)\%$ para o parâmetro β_1, assim

$$IC_{\beta_1,1-\alpha} = \left[\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} DP\left(\hat{\beta}_1\right)\right] \qquad (4.17)$$

e para o intercepto β_0, fica

$$IC_{\beta_0,1-\alpha} = \left[\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} DP\left(\hat{\beta}_0\right) \right] \tag{4.18}$$

em que $z_{1-\frac{\alpha}{2}}$ é o valor critico da distribuição normal padrão correspondente a $100\left(1-\frac{\alpha}{2}\right)\%$.

4.4.3.1. Intervalo de Confiança para o Logit

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

$$IC_{\hat{g}(x),1-\alpha} = \left[\hat{g}\left(x\right) \pm z_{1-\frac{\alpha}{2}} DP\left(\hat{g}\left(x\right)\right) \right] \tag{4.19}$$

onde $\hat{g}\left(x\right) = \hat{\beta}_0 + \hat{\beta}_1 x$ é o estimador para *logit* e $DP\left(\hat{g}\left(x\right)\right)$ é a raiz quadrada de $\hat{Var}\left[\hat{g}\left(x\right)\right] = \hat{Var}\left(\hat{\beta}_0\right) + x^2 \hat{Var}\left(\hat{\beta}_1\right) + 2x\hat{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)$.

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

$$IC_{\pi,1-\alpha} = \left[\frac{e^{\hat{g}(x) \pm z_{\left(1-\frac{\alpha}{2}\right)} DP\left(\hat{g}(x)\right)}}{1+e^{\hat{g}(x) \pm z_{\left(1-\frac{\alpha}{2}\right)} DP\left(\hat{g}(x)\right)}} \right] \tag{4.20}$$

Desta forma pode-se também calcular o intervalo de confiança para Odds Ratio, dado os limites inferiores $\left(\beta_I\right)$ e superiores $\left(\beta_S\right)$, pela seguinte expressão

$$IC_{Odds\ Ratio,\ 1-\alpha} = \left[e^{\beta_I}, e^{\beta_S} \right] \qquad (4.21)$$

4.5. Regressão Logística Múltipla

Assim como no modelo de regressão linear, no modelo de Regressão Logística Múltipla pode-se ajustar um modelo para a variável resposta levando em conta mais de uma variável explicativa (covariável).

Consideremos um conjunto de p covariáveis, x_1,..., x_p, onde a probabilidade do acontecimento de interesse ocorrer é representada pela definição $P\left(Y = \frac{1}{X} \right) = \pi\left(X \right)$ e a função de ligação ou logit é dado, respetivamente, por

$$E\left[Y \right] = \pi\left(X \right) = \frac{e^{g(X)}}{1 + e^{g(X)}} \qquad (4.22)$$

e

$$g\left(X \right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \qquad (4.23)$$

Dependendo das áreas de aplicação da regressão logística, existem várias possibilidades de escolha para as covariáveis, nomeadamente, podem ser sexo, cor dos olhos, grupo de tratamento etc. Assim, surge a necessidade de atribuir valores numéricos, meramente identificativos, a cada categoria da variável, contudo, é necessário utilizar as variáveis *dummy*. Estas variáveis são definidas da seguinte forma

$$D = \begin{cases} 0, \textit{característica Ausente} \\ 1, \textit{característica Presente} \end{cases}$$

4.5.1. Parâmetros do Modelo Múltipla

Para obter as estimativas dos componentes do vetor, será utilizado o método da máxima verossimilhança, assim

$$L\left(\beta_0, \beta_1, ..., \beta_p \middle| \left(x_i, m_i, y_i\right)\right) = \sum_{i=1}^{n}\left[y_i g\left(X\right) - ln\left(1 + e^{g(X)}\right)\right] \qquad (4.24)$$

onde $g\left(X\right)$ é a função de ligação e $\left(x_i, y_i\right)$ são os dados observados.

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

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

2) $\sum_{i=1}^{n} y_i x_i \left(1 + e^{g(X)}\right) - \sum_{i=1}^{n} x_i e^{g(X)} = 0$

Aplicando o método iterativo dos mínimos quadrados ponderados é possível encontrar as raízes destas 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.

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \qquad (4.25)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i) \qquad (4.26)$$

onde *j, l = 0,1,...,p* em que $\pi_i = \pi(x_i)$.

Consideremos uma matriz *(p+1)* x *(p+1)* que contém os termos negativos das derivadas parciais das equações (4.25) e (4.26) denotada por *I(β)*, denominada de Matriz de Informação de Fisher. As variâncias e covariâncias dos coeficientes estimados são obtidos a partir da inversa da matriz *I(β)*.

A matriz de informação de Fisher estimada pode ser obtida por

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

Seja $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$ um vetor de parâmetros de dimensão *k*. O estimador de máxima verossimilhança de $(\beta_1, \beta_2, \ldots, \beta_k)$ são obtidos como soluções das equações:

$$\frac{\partial L(\boldsymbol{\beta}; x)}{\partial \beta_i} \qquad (4.27)$$

4.5.2. 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.

A estatística mais utilizada para verificar a discrepância entre a probabilidade de sucesso observada, π_i, e as probabilidade ajustadas, $\hat{\pi}_i$, considerando a função de verossimilhança, é a *deviance*, cuja a distribuição é assintoticamente χ^2 com (n + p) graus de liberdade, onde n representa o número de observações e p o número de parâmetros do modelo corrente.

Já 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. O teste de Wald averigua se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente.

A existência de uma relação significativa entre a variável dependente e as variáveis independentes ou explicativas pode ser avaliada pelo seguinte teste de hipóteses:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: $\exists \beta_i \neq 0$

A estatística de teste é dada pela expressão (3.22).

4.5.3. Qualidade de Ajuste

A regressão logística binária e multinomial é comumente usada, e os pesquisadores há muito procuram uma medida interpretável da força de um modelo logístico específico.

4.5.3.1. Pseudo R² de McFadden

Os modelos de regressão logística são ajustados usando o método da máxima verossimilhança, ou seja, as estimativas de parâmetros são aqueles valores que maximizam a probabilidade dos dados observados. O pseudo R^2 de McFadden é definida como,

$$R^2_{McFadden} = 1 - \frac{log(L_c)}{log(L_{null})} \quad (4.28)$$

onde L_c representa o valor da máxima verossimilhança do modelo ajustado e L_{null} o valor para o modelo nulo.

O pseudo R² de McFadden é um dos mais utilizados na literatura. McFadden sugere que valores de R^2 entre 0,2 e 0,4 devem ser considerados para representar um bom ajuste do modelo (Louviere, Hensher, & Swait, 2000).

4.5.3.2. Pseudo R² de Cox & Snell e de Nagelkerke

O pseudo R^2 de Cox & Snell e o pseudo R^2 de Nagelkerke descrevem a proporção da variável dependente que é explicada pela variação das variáveis preditoras.

O pseudo R^2 de Cox & Snell baseia-se no logaritmo da verossimilhança para o modelo ajustado em comparação com o logaritmo da verossimilhança para o modelo nulo. No entanto, com resultados categóricos, ele tem um valor máximo teórico inferior a 1.

O pseudo R^2 de Cox & Snell é definida como,

$$R^2_{\text{Cox \& Snell}} = 1 - \left(\frac{L_{null}}{L_c}\right)^{2/n} \quad (4.29)$$

onde *n* representa o número de observações, L_c o valor da máxima verossimilhança do modelo ajustado e L_{null} o valor para o modelo nulo.

O pseudo R^2 de Nagelkerke é uma versão ajustada do pseudo R^2 de Cox & Snell de modo que o intervalo de valores possíveis se situe no intervalo de 0 a 1.

O pseudo R^2 de Nagelkerke é dado por,

$$R^2_{\text{Nagelkerke}} = \frac{1 - \left(\frac{L_{null}}{L_c}\right)^{2/n}}{1 - (L_{null})^{2/n}} \qquad (4.30)$$

onde L_c reo valor da máxima verossimilhança do modelo ajustado e L_{null} o valor para o modelo nulo.

CAPÍTULO 5

5. Análise de Dados de Câncer de Mama

Neste capítulo é apresentado o caso de estudo que foi tratado com o modelo aprofundado, ou seja, o Modelo de Regressão Logística. Na análise dos dados recorreu-se a métodos descritivos e inferenciais conforme apresentado nos capítulos anteriores, bem como a softwares estatísticos para apoio à implementação dos métodos. Nas análises univariada e multivariada dos dados foram utilizados o Microsoft Excel para Mac, o RStudio versão 1.2.5033 para Mac e diversos packages deste ambiente obtidos de https://cran.r-project.org/, e ainda o software estatístico IBM SPSS versão 25 para Mac. O nível de significância adotado nos testes de significância foi $\alpha = 0{,}05$, por ser um valor comum e que permite a comparação com os resultados de um estudo sobre os mesmos dados.

5.1. Descrição da Base de Dados

A base de dados utilizada na análise foi escolhida por ter características de interesse para este tema, nomeadamente: por ser uma pesquisa relacionada com o câncer, incluindo a fisiopatologia, a prevenção, o tratamento e o diagnóstico; porque as variáveis utilizadas resultam de exames comuns de sangue; por permitir prever o desfecho de uma variável categórica dicotômica e por fim devido à sua acessibilidade.

Os dados utilizados nesta análise se referem a um estudo publicado por Patrício et al. (2018) na revista BMC Cancer.

A base de dados contém informação sobre a presença/ausência de câncer de mama em 116 pacientes do sexo feminino, dos quais 52 não apresentam a patologia, enquanto que 64 apresentam a patologia. Este dado está registado internamente na variável Class, uma variável categórica dicotômica. Toma o valor 0 caso a paciente não tenha a patologia, ou o valor 1 caso contrário. A base de dados contém ainda 9 biomarcadores, são eles: Idade, Índice de massa

corporal (IMC), Glicose, Insulina, Homa, Leptina, Adiponectina, Resistina e MCP.1, que são utilizados na área médica como relevantes para predizer a variável resposta de interesse, neste caso ter ou não câncer de mama.

Os dados utilizados estão disponíveis no arquivo dataR2.xlxs em "https://archive.ics.uci.edu/ml/machine-learning-databases/00451/".

5.2. Análise Descritiva Univariada

As 116 pacientes foram divididas em dois grupos: Grupo I, referente às pacientes que não apresentam câncer de mama (Ausência / Controle) e Grupo II, referente aos pacientes com câncer de mama (Presença / Paciente).

Tabela 5. Resumo dos grupos.

Grupo I	Grupo II	Total
0 (Controle)	1 (Paciente)	
52	64	116
45%	55%	100%

Desta forma, das 116 pacientes incluídos na análise, 45% não foram identificados com a patologia enquanto que 55% foram identificados com a patologia. Todos os casos foram incluídos na análise, pois não existem informações omissas.

Consideramos todas as 9 covariáveis disponíveis na base de dados para estudar a variável de interesse, ou seja, o diagnóstico do câncer de mama, com base em dados antropométricos e parâmetros coletados em análises de sangue. No anexo 1, encontramos a descrição sucinta das variáveis em estudo assim como a indicação da nomenclatura usada para as definir.

O objetivo da nossa análise é verificar a importância que cada variável explicativa tem para a variável resposta, criando um modelo para o efeito, a partir do modelo da regressão logística.

Para uma análise preliminar, analisou-se o comportamento das variáveis em estudo com recurso a histogramas para as variáveis contínuas, medidas descritivas básicas e tabelas de frequências

para as restantes variáveis. Em seguida, prosseguiu-se com a análise univariada na qual foi aplicado o teste de normalidade de Kolmogorov-Smirnov, uma vez que existem na nossa base de dados 116 pacientes (n > 50) e por fim, um modelo de regressão logística simples como análise exploratória da relação entre a patologia e cada uma das covariáveis.

Idade

A Tabela 6 mostra o sumário da variável idade para os dois grupos de indivíduos.

Tabela 6. Sumário da variável Idade.

	Mínimo	Q1	Mediana	Média	Desvio Padrão	Q3	Máximo
Grupo I (Controle)	24	41,75	65	58,08	18,96	75	89
Grupo II (Paciente)	34	45	53	56,67	13,49	68	86

Dos 116 indivíduos incluídos na análise, constatou-se que a idade mínima é de 24 anos e a máxima é de 89 anos. No grupo dos pacientes contatou-se que 50% dos indivíduos tinham 53 anos ou mais. Analisando a média de idades dos indivíduos do grupo II, observamos um perfil alto para indivíduos com câncer, facto que pode ser explicado pela presença de um alto número de indivíduos com idades acima de 56 anos.

O grupo de controle apresenta uma assimetria negativa, uma vez que mediana é maior do que a média, enquanto que, o grupo dos pacientes apresenta uma assimetria positiva.

Tabela 7. Coeficiente de assimetria para a variável idade.

	Assimetria	Erro padrão	Coeficiente de assimetria
Controle	-0,276	0,330	-0,836
Paciente	0,532	0,299	1,779

A variável idade foi classificada em 3 classes, a primeira classe corresponde aos indivíduos com idade inferior a 40 anos, a segunda classe refere-se aos indivíduos com idade compreendida

entre os 40 aos 69 anos e a terceira classe corresponde aos indivíduos com mais de 70 anos. A tabela seguinte mostra a caracterização da idade para os grupos e a respetiva frequência.

Tabela 8. Tabela de frequências para a idade.

Idade (anos)	Grupo I		Grupo II		Total	
Inferior a 40	13	87%	2	13%	15	100%
40 - 69	22	31%	49	69%	71	100%
Acima de 70	17	57%	13	43%	30	100%
Total	52	45%	64	55%	116	100%

Na amostra global apenas 15 indivíduos pertencem à primeira classe (12,9%), e a classe com maior quantidade de indivíduos é a segunda classe (61,2%). Observou-se que a maior percentagem de indivíduos com câncer de mama se encontra na segunda classe (69%). Considerou-se a primeira classe como sendo a classe de referência, por ser a categoria que teoricamente possui menor risco de ter a patologia segundo o Instituto Brasileiro de Câncer.

Analisando o Boxplot seguinte, verificou-se que não existem outliers no conjunto de dados.

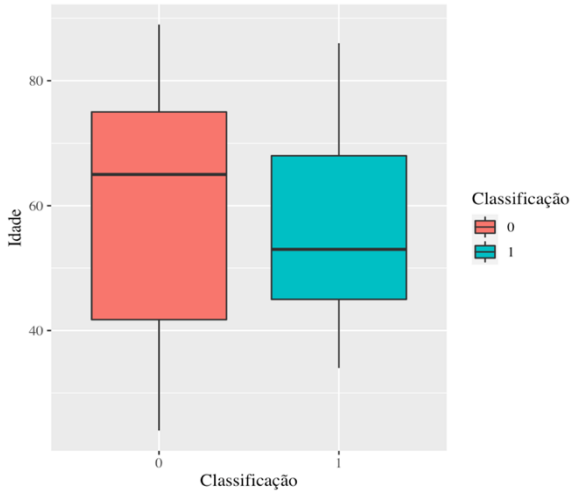

Figura 12. Boxplot para a variável Idade.

Observou-se ainda que a idade dos indivíduos do grupo II tem menor variabilidade que a dos do grupo I, uma vez que os intervalos interquartis, no anexo 3, não sofrem influência de outliers.

Tendo em conta as possíveis análises futuras, realizou-se um teste de normalidade à variável idade no global da amostra e por grupo. Assim, procedeu-se ao teste não paramétrico de Kolmogorov-Smirnov que permite testar a hipótese da normalidade dos dados.

A tabela seguinte apresenta os valores da estatística de teste e da significância do teste realizado no global e por grupos.

Tabela 9. Testes de Normalidade de Kolmogorov-Smirnov para a variável idade.

		Estatística	**df**	**Sig.**
Idade	Controle	0,162	52	0,002
	Paciente	0,153	64	0,001
	Global	0,102	116	0,005

Assim, analisando a significância do teste, podemos afirmar com nível de significância de 5% que a amostra no global não segue uma distribuição normal.

Realizou-se o teste U de Mann-Whitney, para averiguar se as medianas das duas distribuições são iguais.

Tabela 10. Resumo do teste U de Mann-Whitney para a variável idade.

N Total	116
U de Mann- Whitney	1536
Wilcoxon W	3616
Estatística de teste	1536
Erro padrão	180,058
Estatística de teste padronizado	-0,711
Sig. assintótico (teste de 2 lados)	0,477

Apesar das diferenças que se constata entre os valores da variável idade nos dois grupos, o teste não identifica diferenças significativas na distribuição da idade entre os dois grupos (*sig. = 0,477* > 0,05).

IMC

De acordo com a classificação definida pela Organização Mundial de Saúde (OMS), válida somente para pessoas adultas, o índice de massa corporal pode ter as seguintes classificações:

Tabela 11. Classificações do índice de massa corporal.

IMC	Classificação
Abaixo de 18,5	Abaixo do Peso Normal
18,5 - 25	Peso Normal
25 - 30	Acima do peso
Acima de 30	Obeso

A tabela seguinte apresenta o sumário para a variável IMC para os dois grupos de indivíduos.

Tabela 12. Sumário da IMC.

	Mínimo	Q1	Mediana	Média	Desvio Padrão	Q3	Máximo
Grupo I	18,670	23,096	27,694	28,317	5,427	32,328	38,579
Grupo II	18,370	22,789	27,408	26,985	4,620	30,810	37,109

Analisando o grupo I e o grupo II, observou-se que 50% dos indivíduos estão acima do peso ou obeso, e o grupo I apresenta pelo menos um indivíduo abaixo do peso normal. A média de IMC do grupo I é superior a mediana (assimetria positiva), o que indica maior frequência de indivíduos com valores de IMC maiores que a média, enquanto que no grupo II a média é inferior à mediana (assimetria negativa). O grupo II apresenta maior frequência de indivíduos com IMC menores do que a média.

Tabela 13. Coeficiente de assimetria para a variável IMC.

	Assimetria	Erro padrão	Coeficiente de assimetria
Controle	0,152	0,330	0,461
Paciente	0,056	0,299	0,187

Como mostrado na tabela seguinte, a classe *baixo peso* é a classe de menor frequência (0,9%).

As classes com maior frequência pertencem aos indivíduos com o *peso normal* e aos indivíduos com *obesidade* (33,6%). Quanto ao endpoint, observou-se que a grande maioria dos indivíduos se verificam na classe acima do peso (62%), uma vez que a categoria *baixo peso* possui somente 1 indivíduo.

Tabela 14. Tabela de frequências para o índice de massa corporal.

IMC	Grupo I		Grupo II		Total	
Baixo peso (< 18,5)	0	0%	1	100%	1	100%
Peso normal (18,5 - 24,9)	17	44%	22	56%	39	100%
Acima do peso (25-30)	14	38%	23	62%	37	100%
Obesidade (Acima de 30)	21	54%	18	46%	39	100%
Total	52	45%	64	55%	116	100%

Analisando o diagrama seguinte, pode-se observar que não existem outliers no conjunto de dados. O índice de massa corporal dos indivíduos do grupo II apresenta uma menor variabilidade que a dos do grupo I.

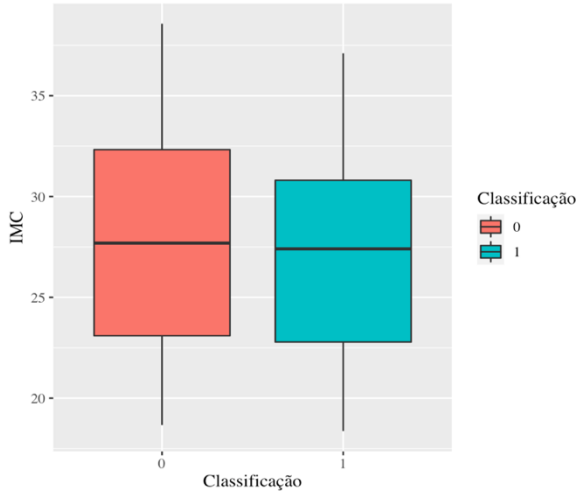

Figura 13. Boxplot para a variável IMC.

Assim, procedeu-se ao teste não paramétrico de Kolmogorov-Smirnov. Na tabela seguinte pode-se encontrar os valores da estatística de teste e da significância do teste realizado no global e por grupos.

Tabela 15. Testes de Normalidade de Kolmogorov-Smirnov para a variável IMC.

		Estatística	df	Sig.
IMC	Controle	0,124	52	0,044
	Paciente	0,094	64	0,20
	Global	0,096	116	0,01

Assim, dado que *p-value* = 0,01 < 0,05, concluiu-se que os dados não seguem uma distribuição normal. Por fim, procedeu-se ao teste U de Man-Whitney e obteve-se os seguintes resultados:

Tabela 16. Resumo do teste U de Mann-Whitney para a variável IMC.

N Total	116
U de Mann- Whitney	1433,5
Wilcoxon W	3513,5
Estatística de teste	1433,5
Erro padrão	180,131
Estatística de teste padronizado	-1,28
Sig. assintótico (teste de 2 lados)	0,201

Observou-se que o teste não identifica diferenças significativas na distribuição da variável IMC entre os dois grupos (*p-valor* = 0,201 > 0,05).

Glicose

De acordo com a ADA, o nível de glicose no sangue pode ser classificado da seguinte forma:

Tabela 17. Classificação da glicose.

Glicose	Normal	Pré-diabetes	Diabetes
Classificação	Inferior a 99	100 - 125	Acima de 125

A tabela seguinte mostra o sumário para a variável Glicose para os dois grupos de indivíduos.

Tabela 18. Sumário para glicose.

	Mínimo	Q1	Mediana	Média	Desvio Padrão	Q3	Máximo
Grupo I	60	82,750	87	88,231	10,192	93,25	118
Grupo II	70	92	98,5	105,563	26,557	109	201

Analisando o grupo II (pacientes), constatou-se que pelo menos 50% dos indivíduos são pré-diabéticos ou diabéticos, enquanto que, no grupo I (controle) constatou-se que não existem indivíduos diabéticos. Observou-se ainda que as distribuições dos dois grupos possuem assimetria positiva A tabela seguinte ilustra o resultado do coeficiente de assimetria para os grupos de indivíduos.

Tabela 19. Coeficiente de assimetria para a variável IMC.

	Assimetria	Erro padrão	Coeficiente de assimetria
Controle	0,302	0,330	0,915
Paciente	2,157	0,299	7,214

Na tabela seguinte encontramos as frequências para as categorias para os dois grupos de indivíduos.

Tabela 20. Tabela de frequências para glicose.

Glicose (mg/dL)	Grupo I		Grupo II		Total	
Normal (inferior a 99)	44	56,4%	34	43,6%	78	100%
Pré-diabetes (100 - 125)	8	29,6%	19	70,4%	27	100%
Diabetes (Superior a 125)	0	0%	11	100%	11	100%
Total	52	45%	64	55%	116	100%

Analisando a classe de diabéticos, observou-se que todos os indivíduos desta classe possuem a neoplasia. Os indivíduos com o nível de glicose entre inferior a 99 mg/dL no sangue é a categoria mais comum com 67,3%, enquanto que a categoria com o nível de glicose acima de 125 mg/dL, a menos comum com 9,5%.

Analisando o diagrama seguinte, observou-se que existem alguns outliers e outliers severos (assinalados por *) no conjunto de dados. Analisando os intervalos interquartis, pode-se verificar que o nível de glicose dos indivíduos do grupo I (controle) tem menor variabilidade que a dos do grupo II (paciente).

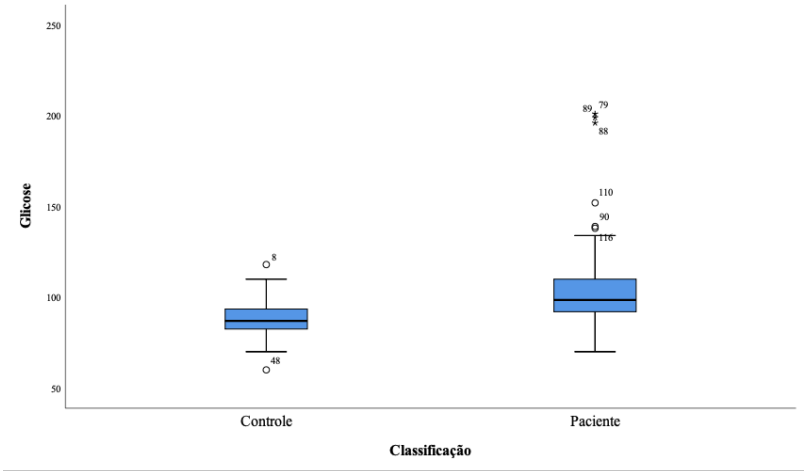

Figura 14. Boxplot para a variável Glicose.

Procedeu-se ao teste não paramétrico de Kolmogorov-Smirnov. A tabela seguinte ilustra os valores obtidos.

Tabela 21. Testes de Normalidade de Kolmogorov-Smirnov para a variável glicose.

		Estatística	df	Sig.
Glicose	Controle	0,087	52	0,200
	Paciente	0,228	64	0,000
	Global	0,202	116	0,000

Assim, concluiu-se que a amostra no global não provém de uma população normal, dado que *p-value* = 0,000 < 0,05.

Seguiu-se com o teste U de Man-Whitney e os seguintes valores foram retornados:

Tabela 22. Resumo do teste U de Mann-Whitney para a variável glicose.

N Total	116
U de Mann- Whitney	2544,5
Wilcoxon W	4624,5
Estatística de teste	2544,5
Erro padrão	180,033
Estatística de teste padronizado	4,891
Sig. assintótico (teste de 2 lados)	0,000

Assim, concluiu-se que existem diferenças significativas na distribuição da variável glicose entre os dois grupos (*sig.* = 0,000 < 0,05).

Insulina

A tabela seguinte mostra o sumário para a variável Insulina para os dois grupos de mulheres.

Tabela 23. Sumário da insulina.

	Mínimo	**Q1**	**Mediana**	**Média**	**Desvio Padrão**	**Q3**	**Máximo**
Grupo I	2,707	4,304	5,484	6,934	4,860	7,001	26,211
Grupo II	2,432	4,406	7,580	12,513	12,318	16,063	58,460

Constatou-se que a média é maior do que a mediana nos dois grupos, portanto a assimetria é positiva em ambos os grupos.

A tabela seguinte ilustra o resultado do coeficiente de assimetria para os grupos de indivíduos.

Tabela 24. Coeficiente de assimetria para a variável insulina.

	Assimetria	**Erro padrão**	**Coeficiente de assimetria**
Controle	2,414	0,330	7,315
Paciente	1,960	0,299	6,556

Analisando o diagrama seguinte, constatou-se que existem outliers severos no conjunto de dados.

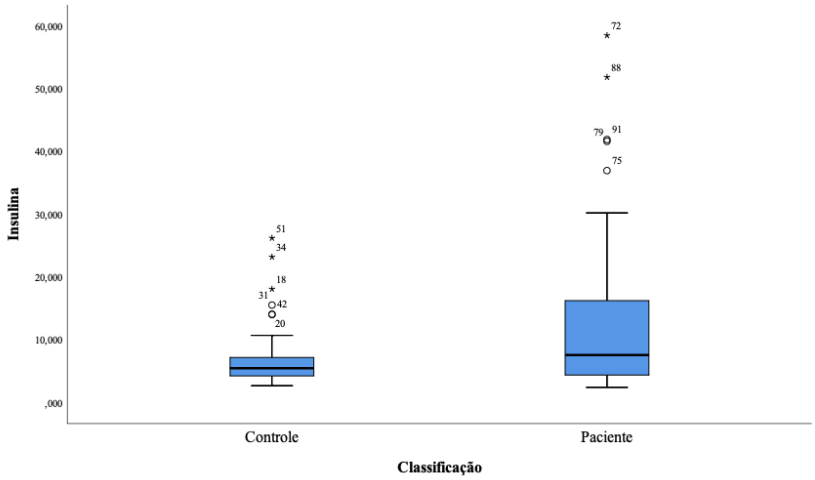

Figura 15. Boxplot para a variável Insulina.

O nível de insulina das mulheres do grupo I apresenta uma menor variabilidade em relação ao grupo II.

A tabela seguinte ilustra os valores do teste não paramétrico de Kolmogorov-Smirnov.

Tabela 25. Testes de Normalidade de Kolmogorov-Smirnov para a variável insulina.

		Estatística	**df**	**Sig.**
Insulina	Controle	0,260	52	0,000
	Paciente	0,207	64	0,000
	Global	0,229	116	0,000

Uma vez que se tem significância do teste *sig.* < 0,05, pode-se afirmar que a amostra no global não provém de uma população normal.

O teste U de Mann–Whitney para a variável insulina retornou os seguintes resultados:

Tabela 26. Resumo do teste U de Mann-Whitney para a variável insulina.

N Total	116
U de Mann- Whitney	2064
Wilcoxon W	4144
Estatística de teste	2064
Erro padrão	180,132
Estatística de teste padronizado	2,221
Sig. assintótico (teste de 2 lados)	0,026

Assim, dado que *sig.*= 0,026 < 0,05, a hipótese de que a distribuição dos dados para a insulina é a mesma entre os dois grupos é rejeitada, ou seja, os grupos não têm distribuições idênticas.

HOMA

O HOMA-IR é um método usado na avaliação da sensibilidade à insulina, relacionando a glicose com a insulina.

Tabela 27. Sumário para HOMA.

	Mínimo	**Q1**	**Mediana**	**Média**	**Desvio Padrão**	**Q3**	**Máximo**
Grupo I	0,467	0,880	1,140	1,552	1,218	1,775	7,112
Grupo II	0,508	1,037	2,052	3,623	4,589	4,461	25,050

Tendo em conta a média e a mediana, observou-se que as distribuições apresentam assimetrias positivas, visto que a média nas duas distribuições é maior do que a mediana. A tabela seguinte mostra os valores dos coeficientes de assimetria.

Tabela 28. Coeficiente de assimetria para a variável HOMA.

	Assimetria	**Erro padrão**	**Coeficiente de assimetria**
Controle	2,693	0,330	8,161
Paciente	2,911	0,299	9,736

Pelo diagrama seguinte, constatou-se que existem alguns outliers severos no conjunto de dados.

Analisando o diagrama, pode-se afirmar que o nível de insulina dos indivíduos do grupo I (controle) tem menor variabilidade que a dos do grupo II (paciente).

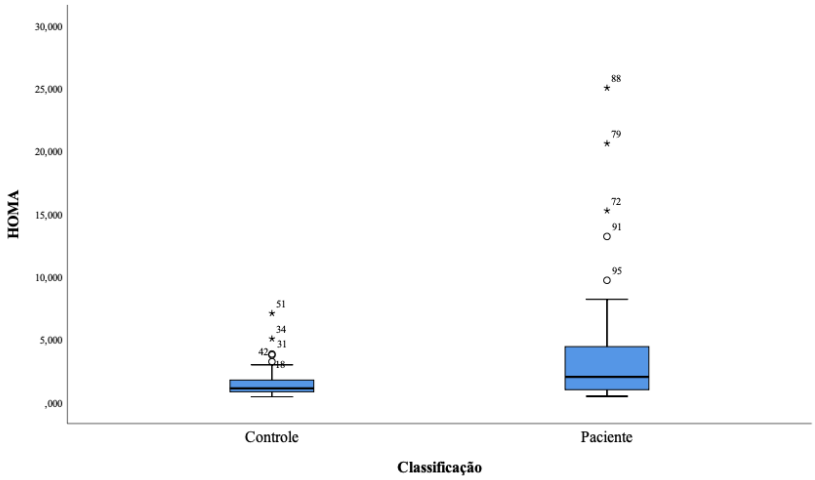

Figura 16. Boxplot para a variável HOMA.

Em seguida, realizou-se um teste de não paramétrico de Kolmogorov-Smirnov. A tabela seguinte ilustra os resultados obtidos.

Tabela 29. Testes de Normalidade de Kolmogorov-Smirnov para a variável HOMA.

		Estatística	df	Sig.
HOMA	Controle	0,240	52	0,000
	Paciente	0,249	64	0,000
	Global	0,270	116	0,000

Assim, concluiu-se que a amostra não provém de uma distribuição normal (sig. = 0,000< 0,05).

O teste U de Mann–Whitney para a variável HOMA retornou os seguintes resultados:

Tabela 30. Resumo do teste U de Mann-Whitney para a variável HOMA.

N Total	116
U de Mann- Whitney	2201
Wilcoxon W	4281
Estatística de teste	2201
Erro padrão	180,133
Estatística de teste padronizado	2,981
Sig. assintótico (teste de 2 lados)	0,003

Dado que *sig.*= 0,003 < 0,05, a hipótese nula é rejeitada, assim há diferenças nas distribuições nos grupos.

Leptina

A tabela seguinte mostra o sumário para a variável leptina para os dois grupos de indivíduos.

Tabela 31. Sumário para leptina.

	Mínimo	**Q1**	**Mediana**	**Média**	**Desvio Padrão**	**Q3**	**Máximo**
Grupo I	4,311	11,846	21,495	26,638	19,335	36,721	83,482
Grupo II	6,334	12,403	18,878	26,597	19,212	37,378	90,28

Analisando a tabela, pode-se observar que a média nas duas distribuições é maior do que a mediana, assim as distribuições apresentam assimetrias positivas. A tabela seguinte mostra os valores dos coeficientes de assimetria.

Tabela 32. Coeficiente de assimetria para a variável leptina.

	Assimetria	**Erro padrão**	**Coeficiente de assimetria**
Controle	1,152	0,330	3,491
Paciente	1,471	0,299	4,920

Analisando o diagrama seguinte, podemos observar que existem outliers no conjunto de dados.

Analisando os intervalos interquartis, podemos afirmar que não há diferença de variabilidade entre os grupos.

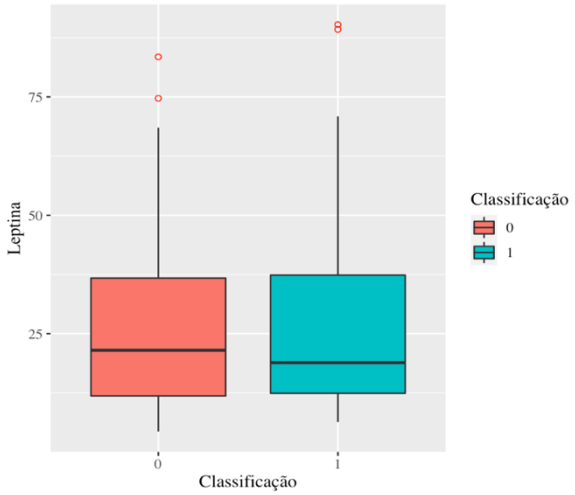

Figura 17. Boxplot para a variável Leptina.

A tabela seguinte ilustra os resultados obtidos pelo teste de Kolmogorov-Smirnov.

Tabela 33. Testes de Normalidade de Kolmogorov-Smirnov para a variável leptina.

		Estatística	df	Sig.
Leptina	Controle	0,154	52	0,003
	Paciente	0,183	64	0,000
	Global	0,149	116	0,000

Ao nível de 5% de significância, pode-se afirmar que a distribuição não provém de uma população com distribuição normal.

O teste U de Mann–Whitney para a variável leptina retornou os seguintes resultados:

Tabela 34. Resumo do teste U de Mann-Whitney para a variável leptina.

N Total	116
U de Mann- Whitney	1676
Wilcoxon W	3756
Estatística de teste	1676
Erro padrão	180,133
Estatística de teste padronizado	0,067
Sig. assintótico (teste de 2 lados)	0,947

Ao nível de significância de 5%, pode-se afirmar que a distribuição de leptina é a mesma entre as categorias de classificação.

Adiponectina

A tabela seguinte mostra o sumário para a variável adiponectina para os dois grupos de indivíduos.

Tabela 35. Sumário para adiponectina.

	Mínimo	Q1	Mediana	Média	Desvio Padrão	Q3	Máximo
Grupo I	2,194	5,454	8,128	10,328	7,631	10,820	38,04
Grupo II	1,656	5,484	8,446	10,061	6,189	12,255	33,75

A distribuição dos dados apresenta uma assimetria positiva para os dois grupos, uma vez que a média nas duas distribuições é maior do que a mediana. A tabela seguinte mostra os valores dos coeficientes de assimetria.

Tabela 36. Coeficiente de assimetria para a variável adiponectina.

	Assimetria	Erro padrão	Coeficiente de assimetria
Controle	1,152	0,330	3,491
Paciente	1,471	0,299	4,920

O diagrama seguinte mostra que existem alguns outliers severos no conjunto de dados.

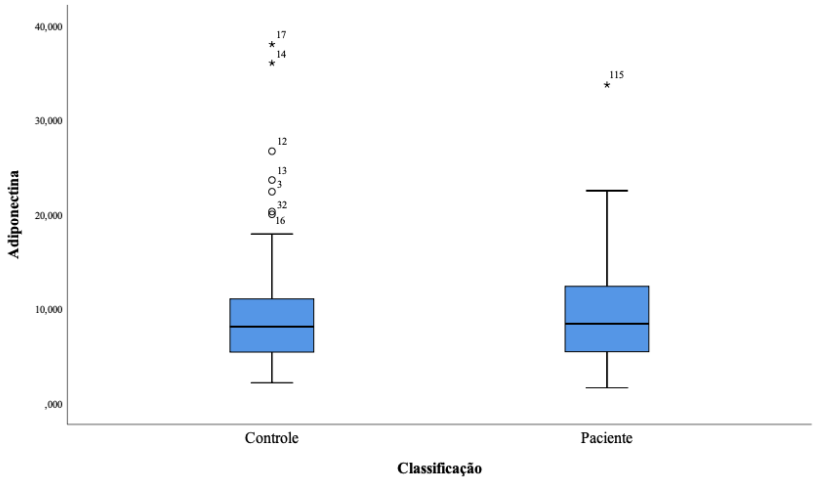

Figura 18. Boxplot para a variável adiponectina.

A tabela seguinte ilustra os resultados obtidos pelo teste de Kolmogorov-Smirnov.

Tabela 37. Testes de Normalidade de Kolmogorov-Smirnov para a variável adiponectina.

		Estatística	df	Sig.
Adiponectina	Controle	0,238	52	0,000
	Paciente	0,143	64	0,002
	Global	0,180	116	0,000

Assim, concluiu-se que a distribuição não provém de uma população normal.

Pelo teste U de Mann–Whitney, pode-se afirmar que os grupos têm distribuições idênticas, como mostrado pela significância do teste (*sig. = 0,764* > 0,05).

Tabela 38. Resumo do teste U de Mann-Whitney para a variável adiponectina.

N Total	116
U de Mann- Whitney	1718
Wilcoxon W	3798
Estatística de teste	1718
Erro padrão	180,133
Estatística de teste padronizado	0,300
Sig. assintótico (teste de 2 lados)	0,764

Resistina

A tabela seguinte mostra o sumário para a variável resistina para os dois grupos de indivíduos.

Tabela 39. Sumário para resistina.

	Mínimo	Q1	Mediana	Média	Desvio Padrão	Q3	Máximo
Grupo I	3,292	6,598	8,929	11,615	11,447	12,809	82,100
Grupo II	3,210	8,114	14,372	17,254	12,637	22,965	55,215

A distribuição dos dados apresenta uma assimetria positiva para os dois grupos. A tabela seguinte mostra os valores dos coeficientes de assimetria.

Tabela 40. Coeficiente de assimetria para a variável resistina.

	Assimetria	Erro padrão	Coeficiente de assimetria
Controle	4,796	0,330	14,533
Paciente	1,526	0,299	5,104

Analisando o diagrama seguinte, verificamos que existe um outliers severo no conjunto de dados pertencente ao grupo de controle. Analisando o diagrama, pode-se afirmar que o grupo I apresenta menor variabilidade em relação ao grupo II.

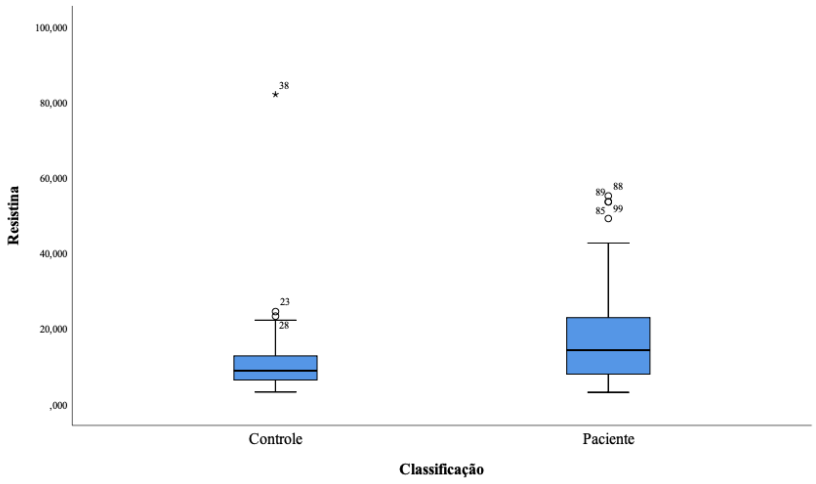

Figura 19. Boxplot para a variável resistina.

A tabela seguinte ilustra os resultados obtidos pelo teste de Kolmogorov-Smirnov.

Tabela 41. Testes de Normalidade de Kolmogorov-Smirnov para a variável resistina.

		Estatística	df	Sig.
Resistina	Controle	0,234	52	0,000
	Paciente	0,149	64	0,001
	Global	0,176	116	0,000

Ao nível de 5% de significância, concluiu-se que a distribuição não provém de uma população normal.

Pelo teste U de Mann–Whitney, pode-se afirmar que a distribuição da resistina não é a mesma entre as categorias de classificação. As diferenças podem ser notadas também no diagrama de caixa.

Tabela 42. Resumo do teste U de Mann-Whitney para a variável resistina.

N Total	116
U de Mann- Whitney	2225
Wilcoxon W	4305
Estatística de teste	2225
Erro padrão	180,133
Estatística de teste padronizado	3,114
Sig. assintótico (teste de 2 lados)	0,002

MCP.1

A tabela seguinte mostra o sumário para a variável MCP.1 para os dois grupos de indivíduos.

Tabela 43. Sumário para MCP.1.

	Mínimo	Q1	Mediana	Média	Desvio Padrão	Q3	Máximo
Grupo I	45,843	260,737	471,323	499,731	292,242	642,934	1256,083
Grupo II	90,090	299,188	465,374	563,017	384,002	737,763	1698,440

A distribuição dos dados apresenta uma assimetria positiva para os dois grupos. A tabela seguinte mostra os valores dos coeficientes de assimetria.

Tabela 44. Coeficiente de assimetria para a variável resistina.

	Assimetria	Erro padrão	Coeficiente de assimetria
Controle	0,742	0,330	2,248
Paciente	1,569	0,299	5,247

Analisando o diagrama observamos outliers no conjunto de dados. Comparativamente com o grupo II, o grupo I apresenta uma variabilidade ligeiramente menor.

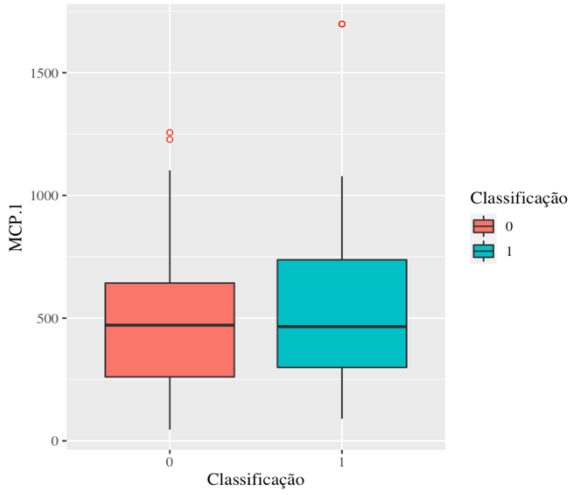

Figura 20. Boxplot para a variável MCP.1.

A tabela seguinte ilustra os resultados obtidos pelo teste de Kolmogorov-Smirnov.

Tabela 45. Testes de Normalidade de Kolmogorov-Smirnov para a variável MCP.1.

		Estatística	df	Sig.
Resistina	Controle	0,103	52	0,200
	Paciente	0,126	64	0,013
	Global	0,103	116	0,004

No global, pode-se afirmar, ao nível de 5% de significância, que a distribuição não provém de uma população normal.

Pela significância do teste U de Mann–Whitney, pode-se afirmar que a distribuição de MCP.1 é a mesma entre as categorias de classificação.

Tabela 46. Resumo do teste U de Mann-Whitney para a variável MCP.1.

N Total	116
U de Mann- Whitney	1785
Wilcoxon W	3865
Estatística de teste	1785
Erro padrão	180,130
Estatística de teste padronizado	0,672
Sig. assintótico (teste de 2 lados)	0,672

Realizou-se ainda uma análise de correlação, ou seja, uma análise bivariada para explorar a correlação entre os potenciais preditores da variável resposta, ter/não ter câncer. A matriz seguinte ilustra a correlação de Pearson entre as variáveis numa escala de −1 a 1.

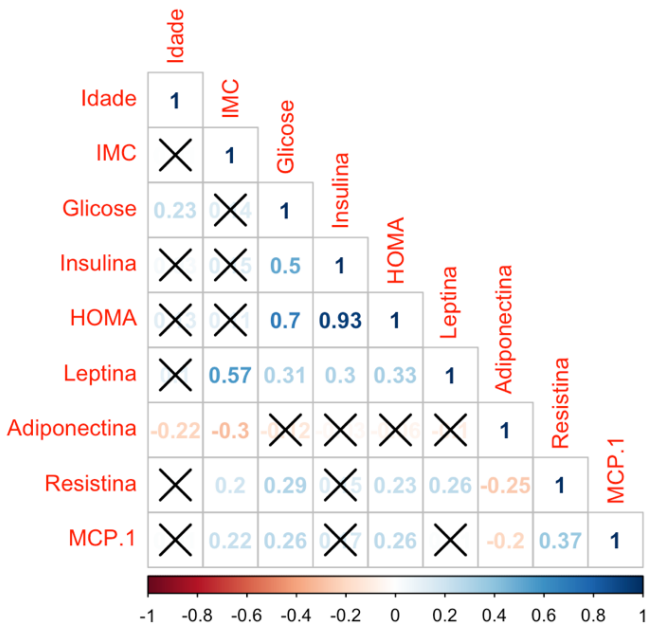

Figura 21. Matriz de correlação de Pearson.

A matriz mostra o valor numérico da correlação entre as variáveis. Os valores assinalados com um X indicam que os valores não são estatisticamente significativos. Assim, podemos observar que as variáveis HOMA e Insulina estão fortemente correlacionados, embora isto não indique causalidade.

Para sustentar os valores anteriores, e atendendo à distribuição dos dados (não normal), realizou-se também uma análise do coeficiente de correlação de Spearman, uma alternativa não paramétrica, e constatou-se mais uma vez que as variáveis HOMA e Insulina estão fortemente correlacionadas.

	Idade	IMC	Glicose	Insulina	HOMA	Leptina	Adiponectina	Resistina	MCP.1
Idade	1								
IMC	0,008	1							
Glicose	**0,184***	0,109	1						
Insulina	0,03	**0,348****	**0,434****	1					
HOMA	0,055	**0,345****	**0,584****	**0,980****	1				
Leptina	0,053	**0,661****	**0,211***	**0,475****	**0,474****	1			
Adiponectina	-0,141	**-0,236***	-0,056	-0,084	-0,085	-0,106	1		
Resistina	-0,111	**0,192***	0,084	0,142	0,134	0,135	**-0,276****	1	
MCP.1	-0,058	**0,271****	-0,039	0,11	0,097	0,084	**-0,227***	**0,488****	1

* A correlação é significativa no nível 0,05 (2 extremidades).

** A correlação é significativa no nível 0,01 (2 extremidades).

Figura 22. Matriz de correlação de Spearman.

5.3. Modelos de Regressão Logística Simples

Nesta secção, realizou-se uma análise exploratória univariada das variáveis Idade, Índice de massa corporal, Glicose, Insulina, Homa, Leptina, Adiponectina, Resistina e MCP.1, utilizando o modelo de regressão logística para uma única variável independente.

Analisou-se inicialmente o modelo de regressão logística contendo apenas o intercepto. A tabela seguinte apresenta a estatística Wald referente à significância do modelo sem considerar as variáveis independentes.

Tabela 47. Coeficiente do modelo nulo.

	B	S.E	Wald (Z)	*Sig.*
Intercepto	0,2076	0,1867	1,112	0,266

A estatística Wald foi de 1,112 com significância de 0,266 (sig. > 0,05), o que significa que o modelo de regressão logística contendo somente o intercepto não contribui para formular previsões sobre o risco de ter câncer de mama.

A tabela de classificação seguinte considera que todos os indivíduos estão dentro da categoria de maior frequência.

Tabela 48. Classificação inicial.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	0	52	0
	Paciente	0	64	100%
Percentagem global				55,2%

Tendo em conta que 64 pacientes têm câncer e 52 não tem câncer, o modelo vai prever que todos os pacientes têm câncer, uma vez que o modelo vai acertar 0% dos que não tem câncer e 100% dos que tem câncer. O modelo tem uma acurácia de 55,2%, classificando de forma correta apenas os indivíduos com câncer. Portanto, faz-se necessário incorporar as variáveis independentes ao modelo, para torna-lo mais assertivo.

Idade

A tabela seguinte apresenta a estatística Wald referente à significância da variável idade incluída no modelo de regressão logística.

Tabela 49. Regressão logística univariada para a variável desfecho com a variável idade.

	B	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**95% C.I. para EXP(B)**	
						Inferior	**Superior**
Inferior a 40	-1,87	0,7596	6,073	0,014	0,154		
40 – 69	2,67	0,8708	11,112	0,001	14,440	3,01	69,68
Acima de 70	1,60	0,8442	3,608	0,058	4,953	0,95	26

Sendo a categoria "Inferior a 40" considerado como categoria de referência, os indivíduos com idades compreendidas entre 40 aos 69 anos apresentam maior chance de ter câncer de mama, 14,44 vezes superior aos indivíduos com idade inferior a 40 anos. Analisando a amplitude do intervalo de confiança, observou-se que a precisão é pequena apesar da classe ser significativa. A classe acima de 70 anos não mostrou significância estatística compatível com o limite estabelecido de 5%. Observou-se que o intervalo de confiança para o odds ratio para a classe ainda inclui o valor 1, o que é indicativo que este previsor não tem contribuição significativa no aumento da chance do desfecho ocorrer. Assim, as classes de idades não são relevantes por si só.

Na tabela seguinte pode-se encontrar os resultados do Teste Omnibus para avaliar a validade do modelo de regressão em realizar previsões sobre risco de o indivíduo ter câncer.

Tabela 50. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável idade.

	Qui-quadrado	df	Sig.
Passo	18,836	2	0,000
Bloco	18,836	2	0,000
Modelo	18,836	2	0,000

O valor do Qui-quadrado foi de 18,836 com significância de 0,000 (*sig.* < 0,05), portanto, rejeita-se a hipótese nula de que todos os coeficientes são nulos. Logo, o modelo de regressão logística contendo a variável idade contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

Procedeu-se ao teste de Hosmer e Lameshow, que avalia a hipótese da inexistência de diferenças entre os resultados previstos e os observados. Assim sendo, a significância deste teste precisa ser superior ao nível de significância adotado no estudo (5%) para que se possa afirmar que os resultados previstos não diferem dos resultados observados.

Tabela 51. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável idade.

Qui-quadrado	df	Sig.
1,1075E-23	8	1

Dado que *sig.* = 1 > 0,05, a hipótese nula não é rejeitada, sendo assim os valores previstos não são significativamente diferentes dos observados, logo o modelo de regressão logística é válido para prever a probabilidade de um indivíduo ter câncer em função da variável independente idade. Na tabela seguinte pode-se observar a classificação final dos casos com o uso do modelo de regressão logística.

Tabela 52. Tabela de classificação do modelo contendo somente a variável idade.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	30	22	57,7%
	Paciente	15	49	76,6%
Percentagem global				68,1%

O modelo tem uma acurácia geral de 68,1%, sendo que classifica corretamente 49 dos 64 casos dos indivíduos que têm câncer (76,6%). Em comparação com o modelo nulo, nota-se que houve uma melhoria na capacidade de acerto das classificações dos indivíduos.

IMC

Para calcular as estimativas dos parâmetros do modelo, e consequentemente, as razões de chances, foram agrupadas as categorias *baixo peso* e *peso normal*, uma vez que a categoria *baixo peso* apresentava um único indivíduo, o que causaria alguns problemas numéricos. A tabela

seguinte apresenta a estatística Wald referente à significância da variável IMC incluída no modelo de regressão logística.

Tabela 53. Regressão logística univariada para a variável desfecho com a variável IMC.

	B	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**95% C.I. para EXP(B)**	
						Inferior	**Superior**
Peso Normal	0,302	0,320	0,893	0,345	1,353		
Acima do peso	0,194	0,466	0,174	0,677	1,214	0,49	3,03
Obesidade	-0,456	0,453	1,014	0,314	0,634	0,26	1,54

No modelo de regressão logística univariada, obtive-se $\hat{\beta}_1$= 0,194 a um odds de 1,214. Analisando o intervalo de confiança, constatamos que o número 1 é incluído no intervalo de confiança para os coeficientes das categorias "Acima do peso" e "Obesidade", uma vez que a categoria Peso Normal foi considerada a categoria de referência, o que nos leva a concluir que estes previsores não têm contribuição significativa no aumento da chance do indivíduo ter câncer, o que pode ser confirmado pelo valor da significância, uma vez que não é significativo (*sig.* > 0,05).

Na tabela seguinte pode-se encontrar os resultados do Teste.

Tabela 54. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável IMC.

	Qui-quadrado	df	Sig.
Passo	2,102	2	0,350
Bloco	2,102	2	0,350
Modelo	2,102	2	0,350

O teste retornou um Qui-quadrado de 2,102 com significância de 0,350 (*sig.* > 0,05), portanto, não se rejeita a hipótese nula. Logo, o modelo de regressão logística contendo a variável IMC não contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

Pelo teste de Hosmer e Lameshow pode-se afirmar que os resultados previstos não diferem dos resultados observados, visto que *p-valor* = 1 > 0,05.

Tabela 55. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável IMC.

Qui-quadrado	df	Sig.
7,1066E-29	8	1

Pela tabela seguinte pode-se observar a classificação final dos casos.

Tabela 56. Tabela de classificação do modelo contendo somente a variável IMC.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	21	31	40,4%
	Paciente	18	46	71,9%
Percentagem global				57,8%

Com uma precisão de 57,8%, o modelo classifica corretamente 46 dos 64 casos dos indivíduos que têm câncer (71,9%), entretanto o modelo erra 18. Comparativamente com o modelo contendo apenas a variável idade, pode-se afirmar que o modelo anterior é melhor do que o modelo atual.

Glicose

Na tabela seguinte pode-se observar a estatística Wald referente à significância da variável glicose incluída no modelo de regressão logística.

Tabela 57. Regressão logística univariada para a variável desfecho com a variável glicose.

	B	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**95% C.I. para EXP(B)**	
						Inferior	**Superior**
Normal	-0,258	0,228	1,275	0,259	0,773		
Pré-diabetes	1,123	0,479	5,487	0,019	3,074	1,201	7,864
Diabetes	21,461	1,2E+4	0,000	0,999	2,09E+9	0,000	.

O modelo de regressão logística retornou um $\hat{\beta}_1$= 1,12 a um odds de 3,07, o que significa que os indivíduos pré-diabéticos tem uma maior chance de ter câncer (3 vezes) em relação aos

indivíduos pertencentes à classe normal, uma vez que o coeficiente é significativo (*sig.=0,019* < 0,05). Enquanto isso, a classe *diabetes* não é estatisticamente significativa (*sig.=0,999* > 0,05), ou seja, este previsor não tem contribuição significativa no aumento da chance do indivíduo ter câncer. A tabela seguinte ilustra os resultados do Teste Omnibus para o modelo contendo somente a variável glicose.

Tabela 58. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável glicose.

	Qui-quadrado	df	Sig.
Passo	19,906	2	0,000
Bloco	19,906	2	0,000
Modelo	19,906	2	0,000

O teste retornou um Qui-quadrado de 19,906 com significância de 0,000 (*sig.* < 0,05), portanto, rejeita-se a hipótese nula. Logo, o modelo de regressão logística contendo a variável glicose contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

Pelo teste de Hosmer e Lameshow concluiu-se que os resultados previstos não diferem dos resultados observados (*sig.* = 1 > 0,05).

Tabela 59. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável glicose.

Qui-quadrado	df	Sig.
2,5855E-07	8	1

Na tabela seguinte pode-se observar a classificação final dos casos.

Tabela 60. Tabela de classificação do modelo contendo somente a variável glicose.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	44	8	84,6%
	Paciente	34	30	46,9%
Percentagem global				63,8%

O modelo tem uma precisão de 63,8%. Comparativamente com o modelo contendo apenas a variável idade (68,1%), o modelo atual apresenta uma precisão ligeiramente inferior.

Insulina

A tabela seguinte ilustra a estatística Wald referente à significância da variável insulina incluída no modelo de regressão logística.

Tabela 61. Regressão logística univariada para a variável desfecho com a variável insulina.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I. para EXP(B)	
						Inferior	Superior
Constante	-0,543	0,316	2,965	0,085	0,581		
Insulina	0,084	0,031	7,101	0,008	1,088	1, 022	1,157

Obtive-se um $\hat{\beta}_1 = 0{,}084$ a um odds ratio de 1,088. Essa variável possui Exp(B) acima de 1, isso indica que quando este previsor aumenta, aumentam as chances do indivíduo ter câncer de mama.

A tabela seguinte apresenta os resultados do Teste Omnibus.

Tabela 62. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável insulina.

	Qui-quadrado	df	Sig.
Passo	11,006	1	0,001
Bloco	11,006	1	0,001
Modelo	11,006	1	0,001

O teste retornou um Qui-quadrado de 11,006 com significância de 0,001 ($sig. < 0{,}05$), portanto, rejeita-se a hipótese nula. Logo, o modelo de regressão logística contendo a variável insulina contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

Pelo teste de Hosmer e Lameshow concluiu-se que os resultados previstos não diferem dos resultados observados ($sig. = 0{,}794 > 0{,}05$).

Tabela 63. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável insulina.

Qui-quadrado	df	Sig.
4,657	8	0,794

Na tabela seguinte pode-se observar a classificação final dos casos.

Tabela 64. Tabela de classificação do modelo contendo somente a variável insulina.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	35	17	67,3%
	Paciente	28	36	56,3%
Percentagem global				61,2%

O modelo tem uma precisão de 61,2%. Comparativamente com os modelos contendo apenas as variáveis idade (68,1%) e glicose (63,8%), o modelo atual apresenta uma precisão ligeiramente inferior à precisão do modelo contendo apenas a variável glicose.

HOMA

A tabela seguinte apresenta a estatística Wald referente à significância da variável HOMA incluída no modelo de regressão logística.

Tabela 65. Regressão logística univariada para a variável desfecho com a variável HOMA.

	B	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**95% C.I. para EXP(B)**	
						Inferior	**Superior**
Constante	-0,624	0,315	3,920	0,048	0,536		
HOMA	0,388	0,136	8, 087	0,004	1,474	1,128	1,926

A variável HOMA possui Exp(B) acima de 1, assim quando este previsor aumenta, aumentam as chances do indivíduo ter câncer de mama.

A tabela do Teste Omnibus para o modelo contendo somente a variável HOMA ilustra os seguintes resultados:

Tabela 66. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável HOMA.

	Qui-quadrado	df	Sig.
Passo	15,081	1	0,000
Bloco	15,081	1	0,000
Modelo	15,081	1	0,000

Obteve-se um Qui-quadrado de 15,081 com significância de 0,000 (*sig.* < 0,05). Assim, concluiu-se que o modelo de regressão logística contendo somente a variável HOMA contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

O teste de Hosmer e Lameshow retornou os seguintes resultados:

Tabela 67. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável HOMA.

Qui-quadrado	df	Sig.
7,145	8	0,521

O resultado do Teste de Hosmer e Lemeshow mostra um Qui-quadrado de 7,145 com significância de 0,521 (*sig.* > 0,05). Isso indica que os valores previstos não diferem dos resultados observados.

Na tabela seguinte pode-se observar a classificação final dos casos.

Tabela 68. Tabela de classificação do modelo contendo somente a variável HOMA.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	38	14	73,1%
	Paciente	29	35	54,7%
Percentagem global				62,9%

O modelo atual retornou uma precisão de 62,9%, muito semelhante ao modelo contendo apenas a insulina (61,2%).

Leptina

A tabela seguinte apresenta a estatística Wald referente à significância da variável leptina incluída no modelo de regressão logística.

Tabela 69. Regressão logística univariada para a variável desfecho com a variável leptina.

	B	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**95% C.I. para EXP(B)**	
						Inferior	**Superior**
Constante	0,000	0,010	0,433	0,511	1, 234		
Leptina	0,211	0,320	0, 000	0,991	1,000	0,991	1,019

O coeficiente não foi estatisticamente significativo (*sig.=0,991* > 0,05), o que podemos confirmar pelo intervalo de confiança para o *odds ratio*, uma vez que o intervalo de confiança inclui o número 1. A variável é um fator de proteção.

A tabela do Teste Omnibus retornou os seguintes resultados:

Tabela 70. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável leptina.

	Qui-quadrado	df	Sig.
Passo	0,000	1	0,991
Bloco	0,000	1	0,991
Modelo	0,000	1	0,991

Analisando a significância do teste (s*ig. = 0,991* > 0,05), concluiu-se que o modelo de regressão não contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

O teste de Hosmer e Lameshow retornou os seguintes resultados:

Tabela 71. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável leptina.

Qui-quadrado	df	Sig.
8,857	8	0,355

Pela significância do teste (*sig.=0,355* > 0,05), concluiu-se que os valores previstos não diferem dos observados.

Na tabela seguinte pode-se observar a classificação final dos casos.

Tabela 72. Tabela de classificação do modelo contendo somente a variável leptina.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	0	52	0%
	Paciente	0	64	100%
Percentagem global				55,2%

O modelo contendo apenas a leptina retornou a mesma eficácia que o modelo nulo, assim a inclusão da variável leptina em nada melhora a previsão do modelo.

Adiponectina

A tabela seguinte apresenta a estatística Wald referente à significância da variável adiponectina incluída no modelo de regressão logística.

Tabela 73. Regressão logística univariada para a variável desfecho com a variável adiponectina.

	B	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**95% C.I. para EXP(B)**	
						Inferior	**Superior**
Constante	0,266	0,336	0,628	0,428	1,305		
Adiponectina	-0,006	0,027	0, 044	0,834	0,994	0,942	1,049

O coeficiente não foi estatisticamente significativo (*sig.=0,834* > 0,05), o que ainda podemos

confirmar pelo intervalo de confiança para o *odds ratio*, uma vez que o intervalo de confiança inclui o número 1.

A tabela do Teste Omnibus retornou os seguintes resultados:

Tabela 74. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável adiponectina.

	Qui-quadrado	df	Sig.
Passo	0,044	1	0,834
Bloco	0,044	1	0,834
Modelo	0,044	1	0,834

Analisando a significância do teste (s*ig. = 0,834* > 0,05), concluiu-se que o modelo de regressão não contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

O teste de Hosmer e Lameshow retornou os seguintes resultados:

Tabela 75. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável adiponectina.

Qui-quadrado	df	Sig.
8,568	8	0,380

Pela significância do teste (*sig.= 0,380* > 0,05), concluiu-se que os valores previstos não diferem dos observados.

Na tabela seguinte pode-se observar a classificação final dos casos.

Tabela 76. Tabela de classificação do modelo contendo somente a variável adiponectina.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	0	52	0%
	Paciente	0	64	100%
Percentagem global				55,2%

O modelo contendo apenas a adiponectina, à semelhança do modelo contendo apenas a leptina, em nada melhora a previsão do modelo.

Resistina

A tabela seguinte apresenta a estatística Wald referente à significância da variável resistina incluída no modelo de regressão logística.

Tabela 77. Regressão logística univariada para a variável desfecho com a variável resistina.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I. para EXP(B)	
						Inferior	Superior
Constante	-0,470	0,337	1,953	0,162	0,625		
Resistina	0,049	0,021	5,217	0,022	1,050	1,007	1,095

O coeficiente foi estatisticamente significativo (*sig.=0,022* < 0,05), o que podemos confirmar pelo intervalo de confiança para o *odds ratio* ser bem preciso. À medida que este previsor aumenta, aumentam as chances do indivíduo ter câncer de mama.

A tabela do Teste Omnibus retornou os seguintes resultados:

Tabela 78. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável resistina.

	Qui-quadrado	df	Sig.
Passo	6,962	1	0,008
Bloco	6,962	1	0,008
Modelo	6,962	1	0,008

Pela significância do teste (*sig. = 0,008* < 0,05), concluiu-se que o modelo de regressão contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

O teste de Hosmer e Lameshow retornou os seguintes resultados:

Tabela 79. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável resistina.

Qui-quadrado	df	Sig.
11,069	8	0,198

Pela significância do teste (*sig.* = *0,198* > 0,05), concluiu-se que os valores previstos não diferem dos observados.

Na tabela seguinte pode-se observar a classificação final dos casos.

Tabela 80. Tabela de classificação do modelo contendo somente a variável resistina.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	30	22	57,7%
	Paciente	18	46	71,9%
Percentagem global				65,5%

O modelo apresenta uma precisão de 65,5%, muito semelhante aos modelos contendo somente as variáveis idade (68,1%), glicose (63,8%) e HOMA (62,9%) e insulina (61,2%) isoladamente.

MCP.1

A tabela seguinte apresenta a estatística Wald referente à significância da variável MCP.1 incluída no modelo de regressão logística.

Tabela 81. Regressão logística univariada para a variável desfecho com a variável MCP.1.

	B	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**95% C.I. para EXP(B)**	
						Inferior	**Superior**
Constante	-0,083	0,349	0,057	0,811	0,920		
MCP.1	0,001	0,001	0,954	0,329	1,001	0,999	1,002

O coeficiente não foi estatisticamente significativo (*sig.=0,329* > 0,05). A variável não contribui positivamente para o desfecho.

A tabela do Teste Omnibus retornou os seguintes resultados:

Tabela 82. Teste de Omnibus do modelo de coeficientes para o modelo somente com a variável MCP.1.

	Qui-quadrado	df	Sig.
Passo	0,985	1	0,321
Bloco	0,985	1	0,321
Modelo	0,985	1	0,321

Analisando a significância do teste (s*ig. = 0,321* > 0,05), concluiu-se que o modelo de regressão não contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

O teste de Hosmer e Lameshow retornou os seguintes resultados:

Tabela 83. Teste de Hosmer & Lemeshow para o modelo contendo somente a variável MCP.1.

Qui-quadrado	df	Sig.
1,466	8	0,993

Pela significância do teste (*sig.* = *0,993* > 0,05), concluiu-se que os valores previstos não diferem dos observados.

Na tabela seguinte pode-se observar a classificação final dos casos.

Tabela 84. Tabela de classificação do modelo contendo somente a variável MCP.1.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	3	49	5,8%
	Paciente	3	61	95,3%
Percentagem global				55,2%

O modelo apresenta uma precisão de 55,2% à semelhança do modelo sem nenhum preditor, assim o modelo contendo apenas MCP.1 não melhora a previsão do modelo.

Por fim, concluiu-se que o modelo contendo apenas a variável independente idade apresenta uma melhor precisão para prever a variável desfecho.

No Gráfico 1 pode-se observar a curva ROC para os modelos simples dos nove biomarcadores.

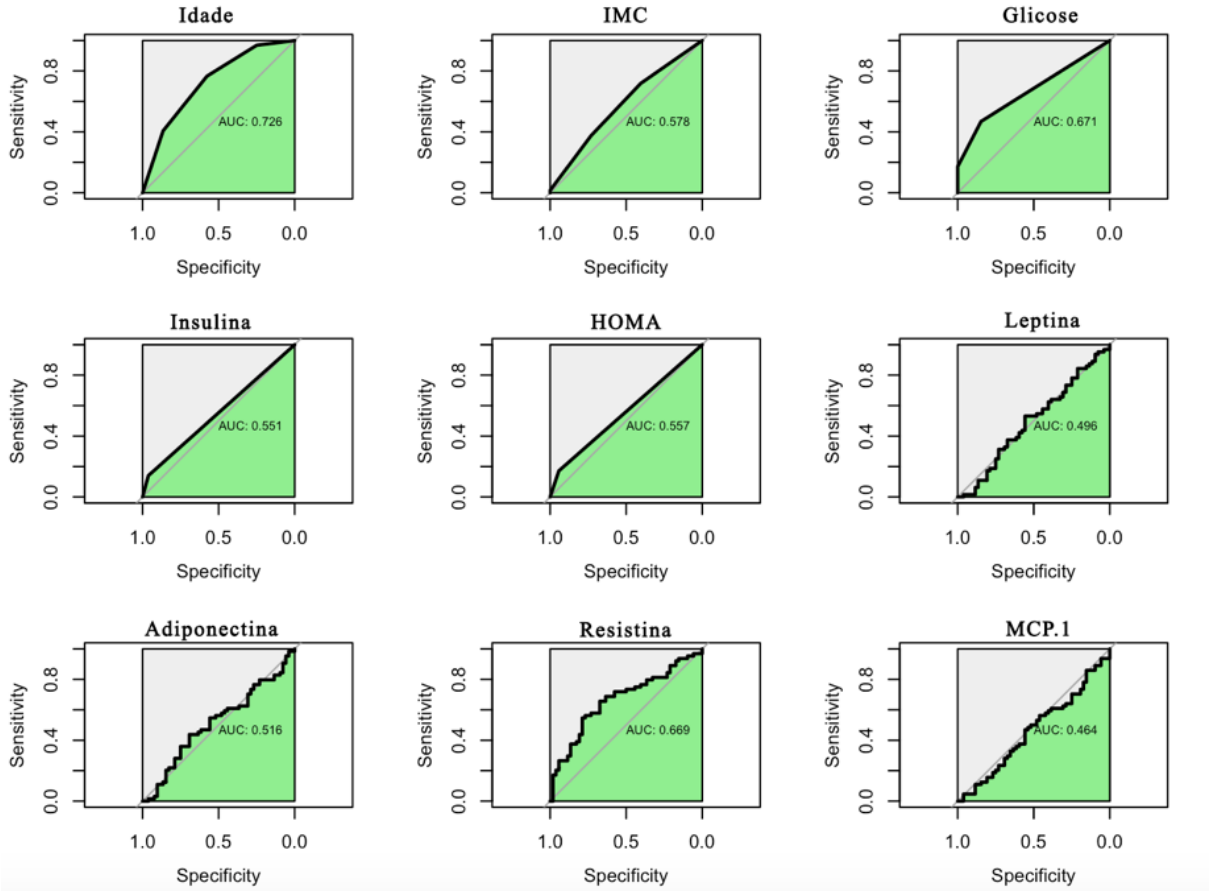

Gráfico 1. Curva ROC para o modelo simples de cada Biomarcador.

Conforme Fávero et al. (2009), quanto maior é a área abaixo da curva ROC, maior é a capacidade do modelo discriminar os grupos de interesse (que para o nosso estudo são os grupos de indivíduos que têm câncer). Em contrapartida, quanto mais próximo a curva ROC estiver da reta diagonal pior é o poder discriminatório do modelo.

Analisando a curva ROC dos biomarcardores, verificamos que por si só, as variáveis são incapazes de discriminar indivíduos doentes e não doentes. De uma forma geral, somente a variável Idade apresenta uma discriminação aceitável.

Tabela 85. Área abaixo da curva ROC dos modelos simples.

Modelo	AUC	Sensibilidade	Especificidade	Precisão
Idade	0,726	0,5769	0,7656	0,681
IMC	0,578	0,4038	0,7188	0,5776
Glicose	0,671	0,8462	0,4688	0,6379
Insulina	0,551	-	-	-
HOMA	0,557	-	-	-
Leptina	0,496	-	-	-
Adiponectina	0,516	-	-	-
Resistina	0,669	0,5769	0,7188	0,6552
MCP.1	0,464	0,05769	0,95312	0,5517

Analisando a tabela com as áreas abaixo da curva ROC, observou-se que os modelos simples não discriminam corretamente o número de resultados de testes verdadeiros positivos e nem o número de resultados negativos em pessoas que não têm a doença.

5.4. Seleção de Variáveis para o Modelo Múltiplo

Tendo em conta a lista de variáveis na análise exploratória com os modelos univariados, o próximo passo consiste em construir um modelo aplicando-se o princípio de regressão logística multivariada.

Para a seleção das variáveis utilizou-se o método de seleção *Stepwise forward* com critério de entrada de 5% e obteve-se o modelo com as variáveis Idade, IMC, Glicose e Resistina. Na tabela seguinte encontram-se os modelos nulo, saturado e o modelo final.

Tabela 86. Medidas de seleção de modelo.

Modelos	Parâmetros	deviance	Diferença de *deviance* (LRT)	AIC	BIC
Modelo Nulo	1	159,567		161,567	164,320
Modelo Saturado	13	100,130	59,437	124,130	157,173
Modelo Final	8	103,655	55,912	119,276	138,551

A estatística *G*, também conhecido como o teste da Razão de Verossimilhanças, é obtido pela diferença de duas funções desvio, ou seja, a diferença entre a *deviance* do modelo nulo em relação aos modelos com *n* variáveis. A estatística *G* é dada como uma medida de variação dos dados. O critério de Akaike (AIC) mede o grau de informação que se perde com o ajuste de um determinado modelo. Analisando a tabela com as medidas de seleção de modelo, conclui-se que o modelo final apresenta menor número de variáveis e menor valor para os critérios AIC e BIC, como esperado.

Na tabela seguinte, pode-se encontrar os coeficientes das variáveis que foram utilizadas no modelo de regressão logística, a respetiva estatística Wald e as estimativas para o *odds ratio*.

Tabela 87. Estimativas dos parâmetros do modelo.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para EXP(B)	
							Inferior	Superior
Resistina	0,058	0,031	3,428	1	0,064	1,06	0,997	1,127
Glicose	0,088	0,022	15,688	1	0	1,092	1,045	1,141
Classes de idades			12,49	2	0,002			
40 - 69	1,974	0,869	5,16	1	0,023	7,196	1,311	39,499
Acima de 70	0,163	0,992	0,027	1	0,869	1,177	0,168	8,228
Classes de IMC			8,339	2	0,015			
25 - 30	-0,122	0,618	0,039	1	0,844	0,886	0,264	2,975
Acima de 30	-1,675	0,621	7,265	1	0,007	0,187	0,055	0,633
Constante	-9,563	2,13	20,161	1	0	0		

Tendo em conta a estatística de Wald e o nível de significância adotado de 5%, observou-se que alguns coeficientes são significativos, uma vez que o valor de sig. é menor que 0,05, portanto contribuem para prever o risco de um individuo ter câncer. As variáveis Classe de idades I (40 – 69), Classe de idades II (Acima de 70), Glicose e Resistina possuem Exp. (B) maior que 1, o que mostra que quando os previsores aumentam, aumentam as chances do indivíduo ter câncer. As variáveis Classe de IMC I (25 – 30) e Classe de IMC II (Acima de 30) possuem Exp. (B) abaixo de 1, isso mostra que quando estes previsores aumentam, diminuem as chances do indivíduo ter câncer.

5.5. Diagnóstico e Análise de resíduo

A seguir, realizou-se algumas técnicas de diagnósticos utilizadas para avaliação da qualidade do ajuste de modelos lineares generalizados e verificação de problemas de ajuste do modelo de regressão logístico. Conforme Cordeiro e Lima Neto (2006), esses problemas são de três tipos: o primeiro é a presença de pontos mal ajustados, no caso pontos aberrantes; o segundo problema é a violação dos pressupostos para os erros e ou para as estruturas das médias; e por último, o terceiro é a presença de observações influentes.

A análise de resíduos se apresenta com uma importante etapa do ajuste Modelos Lineares Generalizados em geral.

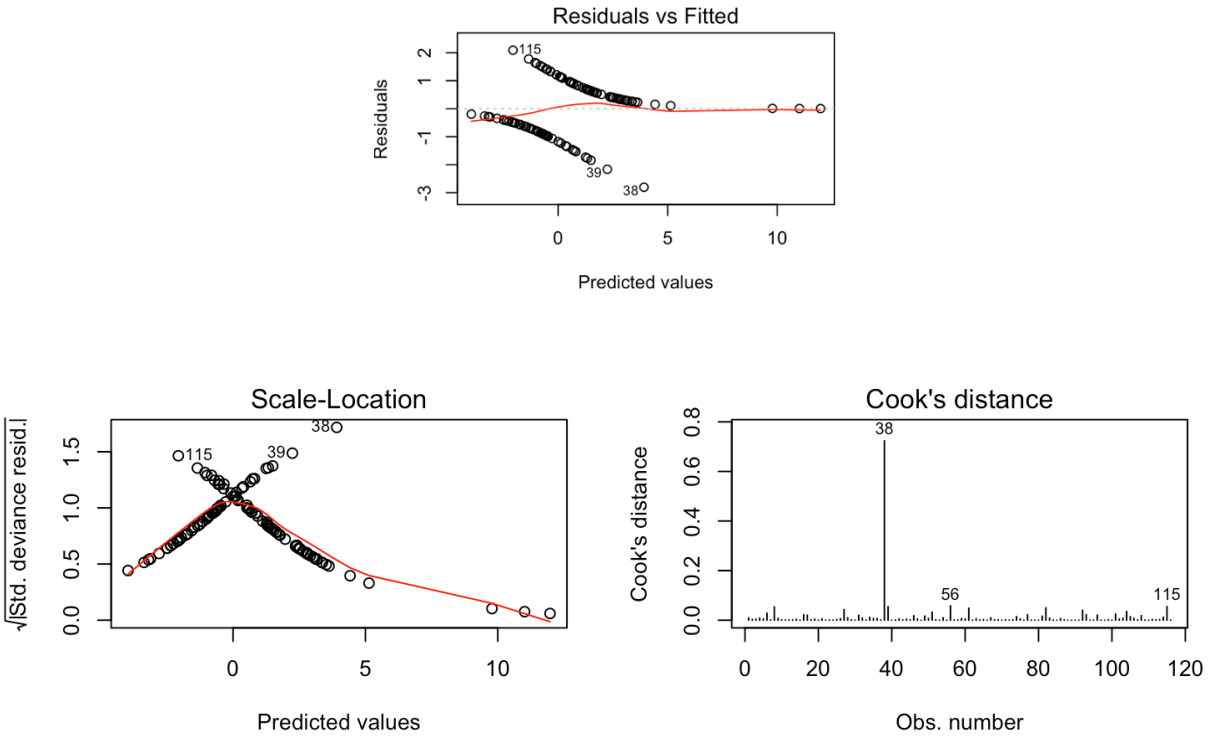

Gráfico 2. Gráficos de diagnóstico.

No gráfico Residuals vs Fitted, os resíduos são comparados com os valores ajustados. Se os resíduos estiverem distribuídos de forma homogênea e simétrica em torno da reta, indica que o

modelo está adequado. Assim, analisando o gráfico, constatou-se que o modelo é adequado. Existem alguns desvios nos extremos, onde se destacam as observações 38, 39 e 115.

O gráfico Scale-Location visualiza a raiz dos resíduos com os valores estimados, para averiguar se existem outliers em Y, onde os pontos 38, 39 e 115 estão distantes dos demais dados.

O gráfico da distância de Cook representa as observações influentes. Os pontos que se afastarem dos demais são aqueles que devem ser avaliados quanto a sua influência no ajuste do modelo. Assim, constatamos que as observações 38, 56 e 115 aparecem como possíveis pontos influentes.

Na tabela seguinte pode-se encontrar os resultados do Teste Omnibus para avaliar a validade do modelo de regressão logística.

Tabela 88. Teste de Omnibus do modelo final.

	Qui-quadrado	df	Sig.
Passo	54,291	6	0,000
Bloco	54,291	6	0,000
Modelo	54,291	6	0,000

O valor do Qui-quadrado foi de 54,291 com significância de 0,000 (*sig.* < 0,05), logo, o modelo de regressão logística contribui para melhorar a qualidade das previsões sobre o risco de ter câncer.

O teste Hosmer & Lemershow (Tabela 89) permite identificar se as variáveis se ajustam bem, ou se têm forte influência sobre a estimação dos parâmetros, ou seja, testa a hipótese de inexistência de diferenças entre os resultados previstos e os observados. Assim, pode-se dizer que o teste mede a precisão de um determinado modelo, sendo que a significância deste teste precisa ser superior ao nível de significância adotado no estudo (5%) para que se possa afirmar que os resultados previstos não diferem dos resultados observados. Sendo assim, a partir do resultado do teste de Hosmer-Lemeshow calculado para o modelo aceita-se a hipótese nula de que não existe diferença entre os resultados previstos pelo modelo e os observados na amostra.

Tabela 89. Teste de Hosmer & Lemeshow.

Modelo	***g.l.***	χ^2	***sig.***
Modelo Final	8	5,251	0,7304

Assim *sig.* = 0,7304 > nível de significância, o que implica um bom ajuste do modelo, ou seja, o modelo está prevendo os dados de forma adequada.

Na tabela seguinte pode-se encontrar os 10 grupos criados através do teste de Hosmer & Lameshow e os valores da variável resposta observados e estimados, cujo os valores probabilísticos ali apresentados se referem aos intervalos de valores que limitam cada um dos 10 grupos.

Tabela 90. Grupos de dados para teste de Hosmer-Lemeshow.

Valores Probabilísticos	Controle		Paciente	
	Observado	Esperado	Observado	Esperado
	y_0	$\hat{y}_0$	y_1	$\hat{y}_1$
[0,0186 ; 0,113]	11	11,187	1	0,813
(0,113 ; 0,205]	11	10,126	1	1,874
(0,205 ; 0,322]	8	7,945	3	3,054
(0,322 ; 0,419]	8	7,555	4	4,444
(0,419 ; 0,587]	4	5,560	7	5,439
(0,587 ; 0,716]	5	4,127	7	7,873
(0,716 ; 0,801]	3	2,809	8	8,191
(0,801 ; 0,9]	0	1,634	12	10,366
(0,9 ; 0,947]	1	0,735	10	10,265
(0,947 ; 1]	1	0,321	11	11,679

Constatou-se que à medida que a probabilidade aumenta, o número de indivíduos que não têm câncer diminui e o número de indivíduos com câncer aumenta.

O gráfico seguinte exibe os resultados obtidos para os pseudo-R^2 que servem para avaliar a qualidade do ajuste do modelo de regressão para prever o risco de um indivíduo desenvolver câncer de mama.

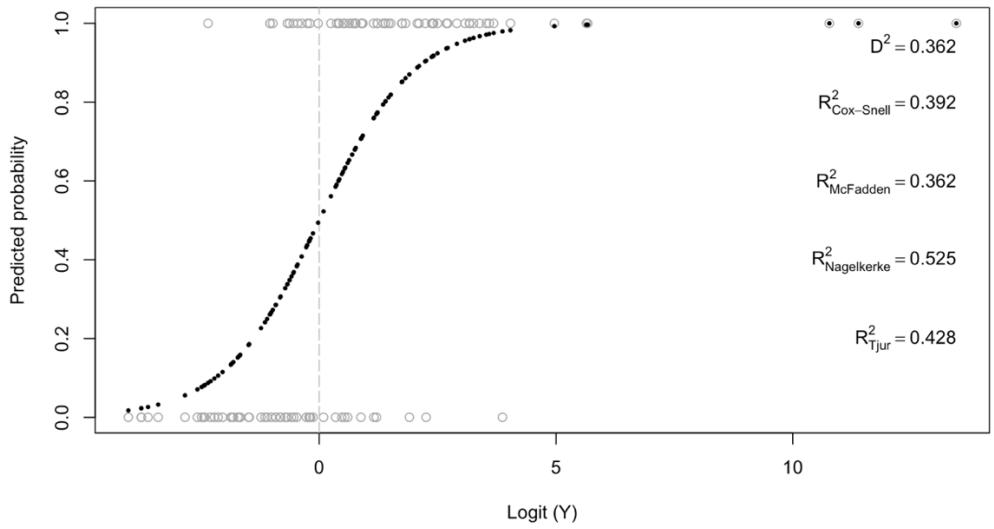

Gráfico 3. Plot do modelo final.

O pseudo *R²* de McFadden foi de 0,362, o que significa que o modelo prevê 36,2% dos fatores que interferem na probabilidade de o indivíduo ter câncer. Portanto, concluiu-se que o modelo tem um bom ajuste para a predição do câncer de mama. Dentre o pseudo R^2 de Cox & Snell e o pseudo R^2 de Nagelkerke damos preferência a de Nagelkerke, visto ser uma medida no intervalo [0;1]. Neste caso, a medida resultou em 0,525, ou seja, 52,5% da variação na variável resposta é explicada pelo modelo.

Curva ROC

Analisando a curva ROC seguinte, pode-se verificar que a curva do modelo de regressão logística está distante da reta diagonal, o que mostra que o modelo é adequado para classificar os indivíduos quanto ao risco de ter câncer, pois quanto mais distante a curva estiver da reta diagonal melhor será o poder de classificação do modelo de regressão.

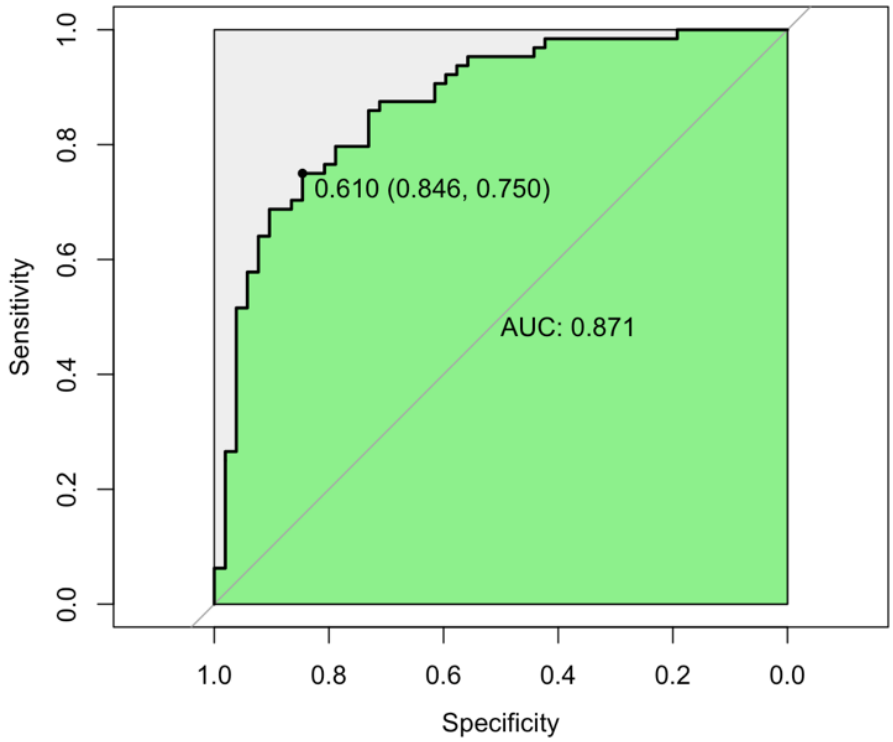

Figura 23. Curva ROC.

A tabela seguinte retorna os seguintes valores de desempenho do modelo:

Tabela 91. AUC e Acurácia para o modelo múltiplo.

Modelo	AUC	Acurácia	IC 95%	
			Limite inferior	Limite superior
Modelo Final	0,871	0,7759	0,6891	0,8481

A curva ROC traçada para o modelo confere um poder de discriminação considerado excelente, segundo Fávero et al. (2009).

Outra forma complementar para qualificar o ajuste do modelo de regressão logística é pela projeção do modelo na Matriz de confusão. Em primeiro lugar determina-se o ponto de corte para classificar o estimado como 0 ou 1. De acordo com Hosmer & Lameshow (2000), se utiliza o valor de 0,5 mas dependendo do estudo proposto pode não ser limitado a este nível.

Tabela 92.Matriz de classificação.

		Previsto		Percentagem correta
		Controle	Paciente	
Observado	Controle	39	13	75%
	Paciente	13	51	79,7%
Percentagem global				77,6%

O modelo retorna uma eficácia de 77,6%. Portanto, é possível perceber que o modelo classificou corretamente 39 dos 52 casos de indivíduos que não tinham câncer, ou seja, uma sensibilidade de 75%. Além disso, o modelo classificou corretamente 51 dos 64 casos dos indivíduos que têm câncer, ou seja, o modelo retorna uma especificidade de 79,7%. Não obstante, 13 observações foram classificadas pelo modelo como sendo de indivíduos que não tinham câncer quando na verdade eram de indivíduos que tinham câncer. Por fim, 13 casos foram classificados como sendo de indivíduos com câncer quando na verdade eram de indivíduos que não tinham câncer.

Conforme Fávero et al. (2009), os modelos de regressão logística precisam cumprir tanto o pressuposto de ausência de multicolinearidade e de heterocedasticidade. Para que não exista multicolinearidade os valores de tolerância devem ser maiores do que 0,1 e os valores de VIF devem ser menores que 10. Algumas bibliografias afirmam que o valor VIF limite para estabelecer que uma variável não seja colinear é igual a 4, sendo que, se este valor for superior a 10, a variável pode ser considerada altamente colinear.

Tabela 93. VIF e tolerância para o modelo.

Variáveis	**Idade (40 - 69)**	**Idade (Acima de 70)**	**IMC (25 - 30)**	**IMC (30 - 40)**	**Glicose**	**Resistina**
VIF	2,367	2,632	1,488	1,405	1,200	1,165
Tolerância	0,422	0,380	0,672	0,712	0,833	0,858

Analisando a Tabela 93, é possível constatar que o modelo não é multicolinear, uma vez que o teste de VIF para as variáveis do modelo indicaram inexistência de colinearidade.

Pelo teste de Breusch-Pagan, analisamos a homocedasticidade dos erros, assim o teste retornou os seguintes resultados:

Tabela 94. Estatística de Breusch-Pagan.

Estatística de BP	***g.l.***	***p-value***
12,660	6	0,0488

Considerando o nível de significância de 5%, concluiu-se que existem evidências para afirmarmos que os erros são heterocedásticos.

Para verificar independência dos resíduos aplicou-se o teste de Durbin Watson que possui a hipótese nula de ausência de correlação serial. Assim, o teste retornou os seguintes resultados:

Tabela 95. Estatística de Durbin-Watson.

Estatística de DW	***p-value***
0,506	0,000

A estatística do teste de Durbin-Watson é construída somente com os resíduos. Ao nível de 5% de significância rejeitamos a hipótese nula de que a correlação entre resíduos sucessivos é nula, e portanto, podemos afirmar há evidências suficientes para afirmar que os resíduos são dependentes.

CAPÍTULO 8

6. Conclusão e Considerações Finais

Esta dissertação assumiu como objetivo explorar biomarcadores antropométricos que facilitam o diagnóstico de câncer de mama utilizando modelos lineares generalizados, mais especificamente a Regressão Logística. Para tal, a análise apoiou-se em nove biomarcadores que podem ser facilmente obtidos a partir de análises de sangue de rotina ou medições diretas (p. ex. idade, IMC).

O software R-Studio mostrou-se consistente no ajuste do modelo de regressão logística, com a função de ligação logit, não obstante ao uso de outros softwares como o Microsoft Excel e o SPSS ao longo da análise, uma vez que as técnicas computacionais são necessárias para a criação de gráficos, estimação dos parâmetros e para verificar pontos atípicos.

Da análise descritiva inicial univariada das variáveis independentes em estudo concluiu-se que as mulheres com idades compreendidas entre 40 e 69 anos apresentam uma maior chance de ter câncer de mama, assim como as mulheres pré-diabéticos ou com resistência à insulina. Concluiu-se ainda que algumas variáveis, isoladamente, aparentam não ter uma relação evidente com o câncer de mama.

No passo seguinte realizou-se uma análise multivariada utilizando o método *stepwise* para a seleção de variáveis explicativas para prever o câncer de mama, onde procedeu-se a algumas técnicas de diagnóstico para o modelo. Dessa forma, pelo método *stepwise*, o modelo selecionou as variáveis Idade, IMC, Glicose e Resistina. A análise do *odds ratio* mostrou que as mulheres na faixa de idades entre 40 e 69 anos tem maior impacto sobre a chance de desenvolver câncer de mama, no entanto, ao analisar a amplitude do intervalo de confiança observou-se que por si só não tem grande relevância.

Baseado nas variáveis Idade, IMC, Glicose, Insulina e Resistina, observou-se que a presença de câncer de mama pode ser predito com 77,6% de chance de acertos. O que sugere que a Idade,

IMC, Glicose e Resistina podem ser considerados bons biomarcadores para o fácil diagnóstico de câncer de mama. Assim, pela análise da curva ROC, verificou-se também que o modelo é adequado para classificar os indivíduos quanto ao risco de ter câncer.

Os resultados obtidos neste trabalho convergem em parte com os resultados de Patrício et al. (2018), que também selecionaram para o seu modelo final as variáveis Idade, IMC, Glicose e Resistina obtidas pelo método SVM, obtendo uma sensibilidade entre 82 e 88% e especificidade entre 85 e 90%. Comparativamente, no presente trabalho obteve-se uma sensibilidade de 75% e uma especificidade de 79,7% com um valor de AUC de 87,1% para o modelo logístico com as mesmas variáveis. Patrício et al. (2018) obtiveram um valor de AUC entre 87 e 91%.

Maulaz et al. (2018) realizaram um estudo onde os parâmetros numéricos de sensibilidade e especificidade para as lesões benignas foram de 91% e 63% para mamografia/ecografia e 77% e 100% para ressonância magnética. Assim, não sendo o método aqui utilizado invasivo por exigir apenas análise de rotina ao sangue, conclui-se que o teste que considera as variáveis Idade, IMC, Glicose e Resistina pode ser um exame complementar à mamografia, devido à sua alta sensibilidade e especificidade, indo ao encontro da conclusão de Patrício et al. (2018).

Guimarães (1985) afirma que é praticamente impossível ter um teste com sensibilidade e especificidade próximos de 100%, por isso geralmente usa-se primeiro um exame sensível (afim de não "deixar escapar" qualquer chance de doença) e posteriormente lança-se mão de um teste específico (afim afastar outras hipóteses e determinar o diagnóstico preciso). Comparativamente, neste trabalho, obteve-se um alto poder discriminante para, dentre os suspeitos, discriminar aqueles que efetivamente possuem a patologia, e um alto poder de discriminar um teste negativo, em face de uma amostra de indivíduos que não têm a doença em questão.

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, entretanto, este modelo de regressão logística não pode ser a única fonte para diagnosticar uma pessoa com câncer, uma vez que o modelo proposto não considera outras variáveis que podem ser relevantes na análise.

Futuras investigações poderiam considerar outras variáveis potencialmente relevantes e novos modelos, bem como recorrer a técnicas de aprendizagem automática (ML) para prever o diagnóstico do câncer de mama como forma de comparação com os anteriores.

Uma outra vertente de pesquisa seria considerar a elaboração de rotinas computacionais para auxiliar no monitoramento da evolução do crescimento do tumor antes e após a utilização de tratamentos de interesse, e avaliar quantitativamente a eficácia desses tratamentos.

Bibliografia

Turkman, A., & Silva, G. (2000). *Modelos Lineares Generalizados - da teoria à prática -.* Lisboa: Universidade Técnica de Lisboa. Obtido em 22 de 07 de 2018, de https://docs.ufpr.br/~taconeli/CE225/tp.pdf

Rossi, A., & Portela, C. (2018). *Modelos Lineares Generalizados.* UnB, Brasilia. Obtido em 12 de Novembro de 2018, de https://lamfo-unb.github.io/2018/09/29/MLG/

Pousinho, A. P. (2013). *Modelos Lineares Generalizados Tweedie Aplicados ao Cálculo de Previsões para Sinistros.* Dissertação de Mestrado, Instituto Superior de Economia e Gestão - Universidade Técnica de Lisboa, Lisboa.

Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135*(3), 370-384. Obtido em 13 de 07 de 2018, de https://docs.ufpr.br/~taconeli/CE225/Artigo.pdf

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Second ed.). London: Chapman and Hall.

Maxwell. (1992). *Modelos Lineares Generalizados* . Obtido em 2 de Novembro de 2018, de maxwell.vrac.puc-rio.br: https://www.maxwell.vrac.puc-rio.br/18163/18163_5.PDF

Laureano, R. M. (2011). *Testes de Hipóteses com o SPSS - O Meu Manual de Consulta Rápida.* Lisboa: Edições Sílabo, Lda.

Laureano, R. M., & Botelho, M. D. (2012). *SPSS - O Meu Manual de Consulta Rápida.* Lisboa: Edições Sílabo, Lda.

Costa, M. A. (2000). *Modelos Lineares Generalizados: uma aplicação à tarifação automóvel.* Dissertação de Mestrado em Probabilidades e Estatística, Universidade de Lisboa, Faculdade de Ciências, Lisboa.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression.* New York, USA: Springer.

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces* (1ª ed.). New York: Wiley.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2ª ed.). Florida: Wiley.

Paula, G. A. (2013). *Modelos de Regressão com apoio computacional.* Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo. Obtido em 23 de 11 de 2018

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society B 49*, 127-162.

Buse, A. (1982). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *The American Statistician, 36*, 153-157.

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922. *Statistical Science*, 162-176.

Hauck Jr, W. W., & Donner, A. (1977). Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, 851-853.

keppel, G. (1991). *Design and analysis: A researcher's handbook* (Terceira Edição ed.). Englewood Cliffs: Prentice-Hall.

Lassance, R. F. (2015). *Comparação dos Modelos Lineares Generalizados Logístico e Log-Binomial.* Trabalho para a obtenção do título de Bacharel em Estatística, UNB, Departamento de Estatística, Brasília.

Schmidt, C. C. (2003). *Modelo de regressão de Poisson aplicado à área da saúde.* Mestrado em Modelagem Matemática, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijúi.

Cordeiro, G. M., & Neto, E. d. (2006). *Modelos Parametricos.* Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, Recife.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second Edition ed.). New York: Wiley.

Cabral, C. S. (2013). *Aplicação do Modelo de Regressão Logística num Estudo de Mercado.* Dissertação de Mestrado, Universidade de Lisboa, Departamento de Estatística e Investigação Operacional, Lisboa.

Ross, S. (2010). *PROBABILIDADE: Um curso moderno com aplicações* (8 ed.). (A. R. De Conti, Trad.) Porto Alegre: Bookman.

Martins, S. A. (2012). *Modelo De Avaliação De Risco Em Acidentes No Ramo Automóvel.* Dissertação de Mestrado, Universidade Nova de Lisboa, Instituto Superior de Estatística e Gestão de Informação, Lisboa.

Díaz, M. d., & do Couto, H. Z. (Dezembro de 1999). Modelos generalizados para a mortalidade de árvores de Eucalyptus grandis no Estado de São Paulo, Brasil. *SCIENTIA FORESTALIS*, 101-111.

Cordeiro, D. B. (2016). *Ajuste De Modelos Lineares Generalizados Para Dados Positivos Assimétricos.* Bacharel em Estatística , Universidade Estadual Da Paraíba, Campina Grande - PB.

Tadano, Y. D., Cassia Maria, U. L., & Franco, A. T. (Julho-Dezembro de 2009). Método de regressão de poisson: metodologia para avaliação do impacto da poluição atmosférica na saúde populacional. *Ambiente & Sociedade, XII*, 241-255.

Schwarz, G. (1978). Estimating The Dimension Of A Model. *The Annals of Statistics, 6*, 461-464.

Cordeiro, G. M., & Demétrio, C. G. (2013). *Modelos Lineares Generalizados e Extensões.* Piracicaba: ESALQ/USP. Obtido em 09 de 07 de 2018, de http://www.ufjf.br/clecio_ferreira/files/2013/05/Livro-Gauss-e-Clarice.pdf

Hair, J., Black, B., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariated Data Analysis* (6th Edition ed.). Prentice Hall.

Nakamura, K. G. (2013). *Multicolinearidade em Modelos de Regressão Logística.* Dissertação de Mestrado, Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo.

Montgomery, D., Peck, E., & Vining, G. (2001). *Introduction to Linear Regression Analysis* (terceira ed.). Wiley Series in Probability and Statistics.

World Health Organization . (2019). *Cancer: Early diagnosis and Screening*. Obtido em 22 de Junho de 2019, de WHO: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

Pinheiro, A. B., Lauter, D. S., Medeiros, G. C., Cardozo, I. R., Menezes, L. M., Souza, R. B., . . . Thuler, L. S. (2013). Breast Cancer in Young Women: Analysis of 12.689 Cases. *Revista Brasileira de Cancerologia*.

Field, A. (2009). *Descobrindo a Estatística usando o SPSS* (Segunda ed.). (L. Viali, Trad.) São Paulo: Artmed.

Silva, A. C. (2011). *Análise Estatística de Inquéritos online.* Dissertação de Mestrado, Universidade do Minho, Braga.

Oliveira, S. (2013). *Inferência E Análise De Resíduos E De Diagnóstico Em Modelos Lineares Generalizados.* Bacharel em Estatística, Universidade Federal De Juiz De Fora, Juiz de Fora.

Stern, S. E., Williams, K., Ferrannini, E., De Fronzo, R. A., Bogardus, C., & Stern, M. P. (2005). *Identification of Individuals With Insulin Resistance Using Routine Clinical Measurements*. Obtido em 2019, de American Diabetes Association: https://diabetes.diabetesjournals.org/content/54/2/333.full-text.pdf

Akaike, H. (December de 1975). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Fávero, P. L., Belfiore, P., Silva, F. L., & Chan, B. L. (2009). *Análise de dados: modelagem multivariada para tomada de decisões.* Rio de Janeiro: Elsevier.

Abbad, G., & Torres, C. V. (2001). Regressão múltipla stepwise e hierárquica em Psicologia Organizacional: aplicações, problemas e soluções. *Scielo*, 23.

Alves, M., Lotufo, A., & Lopes, M. (2013). *Seleção de variáveis stepwiseaplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas.* São Paulo.

Maulaz, C. M., Valentini, B. B., Silva, A. M., & Papaléo, R. M. (2018). Estudo Comparativo do Desempenho de Imagens por Ressonância Magnética, Mamografia e Ecografia na Avaliação de Lesões Mamárias Benignas e Malignas. *Revista Brasileira de Física Médica, 12*, 23-29.

Guimarães, M. (1985). Exames de laboratório: sensibilidade, especificidade, valor preditivo positivo. *Revista da Sociedade Brasileira de Medicina Tropical, 18*, 117-120.

Tabachnick, B., & Fidell, L. S. (1996). *Using multivariate statistics* (Third ed.). New York: Harper Collins.

Patricio, M., Pereira, J., Crisostomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*.

Louviere, J., Hensher, A., & Swait, D. (2000). Stated choice methods. *Cambridge University Press*.

ADA. (1995 - 2020). *American Diabetes Association*. Obtido de Web site da American Diabetes Association: https://www.diabetes.org/a1c/diagnosis

Previato, H. R. (2013). *Qualidade da dieta de mulheres com câncer de mama.* Dissertação de mestrado, Universidade Federal de Ouro Preto, Escola de Nutrição, Ouro Preto - MG.

Boughorbel, S., Al-Ali, R., & Elk, N. (15 de Jan de 2016). Model Comparison for Breast Cancer Prognosis Based on Clinical Data. (M. Ebrahimi, Ed.) *PLoS ONE, 11*(1).

INCA. (2011). *ABC do câncer : abordagens básicas para o controle do câncer.* Instituto Nacional do Câncer, Rio de Janeiro - Brasil.

Chhatwal, J., Alagoz, O., Lindstrom, M. J., Kahn, C. E., Shaffer, K. A., & Burnside, E. S. (1 de Apr de 2009). A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis. *AJR Am J Roentgenol, 4*(192), 1117–1127.

Zhou, X., Liu, K.-Y., & Wong, S. T. (August de 2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics, 37*, 249-259.

Abdolmaleki, P., Dizagi, M. M., & Gity, M. (2004). Logistic discriminant anlysis of breast cancer using ultrasound measurements. *International Journal of Radiation Research, 2*(1), 1-8.

Sultana, J., & Jilani, A. K. (Nov de 2018). Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers. *International Journal of Engineering & Technology (UAE), 7*(4.20), 22-26.

Anene, U. A. (2019). *Classification of Breast Cancer using Logistic Regression.* Thesis, African University of Science and Technology , Department of Computer Science, Abuja, Nigeria.

Oliveira, H. M. (2010). *Modelação do cancro da Mama na região do Alentejo.* Dissertação de Mestrado, Universidade de Évora, Évora.

Cordeiro, G. (1986). Simpósio Nacional de Probabilidade e Estatística. *Livro Texto do VII SINAPE.* Campinas, SP: UFPE/ABE.

Belfiore, P. (2015). *Estatística: aplicada a administração, contabilidade e economia com Excel e SPSS.* Rio de Janeiro: Elsevier.

Cramer, J. (2003). *The origins and development of the logit model.* University of Amsterdam and Tinbergen Institute, Amsterdam.

Corrar, L., Paulo, E., & Filho , J. (2009). *Análise multivariada: para os cursos de Administração, Ciências Contábeis e Economia.* São Paulo: Atlas.

Casella, G., & Berger , R. (2010). *Inferência estatística.* São Paulo: Cengage Learning.

Portal do Médico. (2016). *Mamografia: Um exame doloroso, ineficaz e que aumenta o risco de ter câncer em mais de 200%.* Obtido em 20 de Jul de 2020, de Portal do Médico: https://www.portaldomedico.com/noticia/leia/6cb4f5d2-90f6-4da8-9207-c097c55477ff/mamografia-um-exame-doloroso-ineficaz-e-que-aumenta-o-risco-de-ter-cancer-em-mais-de-200

American Cancer Society. (10 de Sept de 2019). *Breast Cancer Risk and Prevention*. Obtido em 22 de Jun de 2020, de American Cancer Society: https://www.cancer.org/cancer/breast-cancer/risk-and-prevention.html

Coelho, A. S., Santos, M. d., Caetano, R., Piovesan, C., Fiuza, L., Machado, R. D., & Furini, A. d. (12 de Abril de 2018). Predisposição hereditária ao câncer de mama e sua relação com os genes BRCA1 e BRCA2: revisão da literatura. *Revista Brasileira de Análises Clínicas*.

Anexos

Anexo 1 - Descrição das variáveis

Tabela 96. Descrição das variáveis.

Atributos	Informação	Unidade de medida
Idade	Idade de cada indivíduo	Anos
IMC	Índice de Massa Corporal que serve para avaliar o peso do indivíduo em relação à sua altura	kg/m^2
Glicose	Monossacarídeo (açúcar simples) usado pelo organismo como principal fonte de energia para o corpo	mg/dL
Insulina	Hormônio produzido pelo pâncreas que auxilia na passagem do açúcar (glucose) presente no sangue para o interior dos tecidos	µU/mL
HOMA	Homeostatic model assessment - uma medida que serve para avaliar a resistência à insulina e atividade do pâncreas	
Leptina	Hormônio regulador do apetite, responsável pelo controlo da homeostase energética	ng/mL
Adiponectina	Hormônio associado a obesidade e diabetes tipo 2, diretamente associado a quantidade de tecido adiposo	µg/mL
Resistina	A resistina é uma proteína que contém 108 aminoácidos e pertence à família das moléculas resistin-like. Apesar de não se conhecer bem a função potencial da resistina, tem sido sugerido que tem efeitos na regulação positiva da adesão molecular, tendo uma ação pró-inflamatória.	ng/mL
MCP.1	Proteína quimiotática de monócitos-1- produzida por células do sistema renal em resposta ao processo de isquemia-reperfusão	pg/dL
Classificação		0=Ausência / 1=Presença

Anexo 2

Teste U de Mann–Whitney

Tabela 97. Amplitude interquartil.

	Control.median	Control.IQR	Pacient.median	Pacient.IQR	*p-value*
Idade	65,0	33,2	53,0	23,0	0,4789
IMC	28,3	5,4	27,0	4,6	0,2017
Glicose	88,2	10,2	105,6	26,6	0,0000
Insulina	6,9	4,9	12,5	12,3	0,0266
HOMA	1,6	1,2	3,6	4,6	0,0029
Leptina	26,6	19,3	26,6	19,2	0,9491
Adiponectina	10,3	7,6	10,1	6,2	0,7665
Resistina	11,6	11,4	17,3	12,6	0,0019
MCP.1	499,7	292,2	563,0	384,0	0,5035