UNIVERSIDADE ABERTA

Mestrado em Estatística, Matemática e Computação

(Ramo – Estatística Computacional)

Modelos de Regressão: uma aplicação em Medicina Dentária

Maria Cristina Campos dos Santos Ferreira

Dissertação apresentada na Universidade Aberta para obtenção do grau de Mestre em Matemática, Estatística e Computação

Orientadora:

Professora Doutora Teresa Paula Costa Azinheira Oliveira

Lisboa 2013

Ao meu marido

Aos meus filhos André e Carolina

À minha Mãe

À memória do meu Pai

AGRADECIMENTOS

À Professora Teresa Oliveira, minha orientadora de dissertação e docente de Estatística, a quem agradeço, em primeiro lugar, o interesse que me despertou pelo tema, o desejo de continuar a pesquisa e aprofundamento dos meus conhecimentos. Agradeço a confiança, inspiração e contribuição na transmissão dos seus conhecimentos para a elaboração desta tese.

Ao meu querido marido, pelo apoio incondicional, pela sua compreensão e contínua motivação, mesmo nos momentos mais difíceis, o meu muito obrigada.

O meu agradecimento a todos os meus amigos que tornaram possível a realização deste trabalho, sempre demonstrando o seu apoio, mesmo quando tudo parecia impossível.

RESUMO

Os fenómenos biológicos geralmente revestem-se duma elevada complexidade dada a multifatorialidade da sua etiologia. Assim, a análise estatística, como ferramenta indispensável para a determinação de associações e interações complexas entre os diferentes potenciais fatores causais e a variável dependente em estudo, adquire importância capital. Neste contexto a escolha do tipo de análise deverá ser adequadamente fundamentada numa perspectiva teórica, de modo a permitir ao investigador a percepção correta do fenómeno biológico em estudo.

Com o objetivo de clarificar a utilização em estudos na área da medicina dentária de um ajustado tratamento estatístico proponho-me a abordar diferentes análises estatísticas com os dados existentes estudando a pertinência do estudo e a sua viabilidade, tanto em função do significado biológico das variáveis como dos seus valores. São explorados modelos de regressão linear clássicos e o modelo de regressão logística procurando uma interpretação explicada do que se pode retirar de cada análise.

Na minha dissertação, proponho-me a analisar dados recolhidos no âmbito de uma investigação levada a cabo na área da Medicina Dentária. Da base de dados constam registos de observações de 158 indivíduos, sendo 79 diabéticos e 79 não diabéticos, relativos a variáveis bioquímicas, variáveis antropométricas, factores comportamentais e indicadores de saúde oral.

Ao proceder deste modo penso poder dar algum contributo para a aplicação correta da estatística em trabalhos biológicos, alertando para que a análise tem que ser adequada e os resultados devem ser devidamente interpretados.

Palavras-chave: Regressão Linear Simples, Regressão Linear Múltipla, Regressão Logística, Diabetes Mellitus, Doença Periodontal

SUMMARY

The biological phenomena are usually highly complex due to the multifactorial aetiology. Thus the statistical analysis is an indispensable tool find associations and complex interactions between different potential causal factors and the dependent variable under study. In this context the statistical analysis is a major importance tool to access the associations and interactions between dependent and independent variables. The choice of analysis type must be grounded on solide theoretical bases and take in account the experimental design and the nature of the phenomena of interest, in order give the researcher a correct perception of the phenomenon under study.

In my dissertation, I propose to apply different multivariate statistical analysis methods to a data set of 158 subjects (diabetic and nondiabetic) in order to compare their results and feasibility, and get for each one an biological interpretation of the result. The data were collected as part of a research work in the area of dentistry and includes variables on biochemical, anthropometric, behavioural, and oral health surrogated endpoints.

This work is an effort to highlight the importance of a correct application and interpretation of the statistics in biological work.

Keywords: Linear regression, Multiple Regression, Logistic Regression, Diabetes Mellitus, Periodontal disease

SIMBOLOGIA E NOTAÇÕES

ADM	Análise Discriminante Múltipla
ANOVA	Análise de Variância
IC	Intervalo de Confiança
MMQ	Método dos Mínimos Quadrados
RLS	Regressão Linear Simples
MLG	Modelo Linear Generalizado
MRLM	Modelo de Regressão Linear Múltiplo
MRLS	Modelo de Regressão Linear Simples
QME	Quadrado Médio do erro
QM Reg	Quadrado Médio devido à regressão
QM Res	Quadrado Médio dos resíduos
RM	Regressão Múltipla
ROC	Receiver Operating Characteristic
SQ Reg	Soma dos Quadrados devidos à regressão
SQ Res	Soma dos Quadrados dos resíduos
SQT	Soma Quadrática Total
SPSS	Statistical Package for the Social Sciences
A1c	Hemoglobina glicada
CT	Colesterol total
GJ	Glicemia em jejum
HDL	Lipoproteínas de alta densidade
HPS	Hemorragia pós-sondagem
IMC	Índice de Massa Corporal
IP	Índice de placa
LDL	Lipoproteínas de baixa densidade
NA	Nível de aderência clínica
PS	Profundidade de sondagem
RG	Retração gengival
RCA	Relação perímetro da cinta/perímetro da anca

T	Triglicerídeos
$\mathbf{H_0}$	Hipótese Nula
$\mathbf{H_1}$	Hipótese Alternativa
VD	Variável dependente
VI	Variável independente
$\mathbf{Y_i}$	Valor observado da variável dependente
$\mathbf{\hat{Y}_i}$	Valor estimado da variável dependente
$\mathbf{\overline{Y}}$	Média da variável dependente
$\boldsymbol{\mu_i}$	Média da observação i
$\boldsymbol{\sigma_Y}$	Desvio Padrão da variável Y
$\boldsymbol{\sigma_{X,Y}}$	Covariância entre as variáveis X e Y
$\mathbf{s_Y^2}$	Variância amostral de Y
$s_{X,Y}$	Covariância amostral entre X e Y
$\boldsymbol{\alpha}$	Nível de significância – Erro tipo I
β	Erro tipo II
$\boldsymbol{\beta_0}$ **e** $\boldsymbol{\beta_1}$	Constantes (parâmetros) desconhecidas
$\boldsymbol{b_0}$ **e** $\boldsymbol{b_1}$	Estimativas dos parâmetros $\boldsymbol{\beta_0}$ **e** $\boldsymbol{\beta_1}$
$\boldsymbol{\varepsilon}$	Erro aleatório
$\boldsymbol{\tau_i}$	Efeito do tratamento *i*
$\mathbf{r^2}$	Coeficiente de determinação
$\mathbf{R^2}$	Coeficiente de determinação múltipla
r_{aj}	Coeficiente de determinação ajustado
ρ	Coeficiente de Correlação
$\boldsymbol{\rho_P}$	Coeficiente de Correlação de Pearson
$\boldsymbol{\rho_S}$	Coeficiente de Correlação de Spearman
$\boldsymbol{CV}$	Coeficiente de variância
b	Coeficiente angular da reta de regressão
$\mathbf{\sigma_e^2}$	Variância dos erros ou resíduos e_i
or	*Odds ratio*

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

INTRODUÇÃO

O termo “regressão” foi usado, pela primeira vez, por Francis Galton num estudo sobre a relação entre a altura dos pais e dos filhos e verificou que, mesmo havendo uma tendência para os pais altos terem filhos altos e os pais baixos terem filhos baixos, a altura média de filhos tendia a deslocar-se, ou a “regredir” (daí ter começado a fazer uso deste termo) para a altura média da população.

A *lei de regressão universal* de Galton foi confirmada mais tarde por Karl Pearson, tendo este recolhido mais de mil registos das alturas de indivíduos pertencentes a grupos de famílias altas e de famílias baixas, verificou que a altura média dos filhos de um grupo de pais altos era inferior à altura de seus pais e que a altura média dos filhos de um grupo de pais baixos era superior à altura de seus pais. Assim, tanto os filhos altos como os baixos “regrediram” em direção à altura média da população. De acordo com Gujarati, “A análise de regressão ocupa-se do estudo da dependência de uma variável, a *variável dependente*, em relação a uma ou mais variáveis, as *variáveis explicativas*, com o objetivo de estimar e/ou prever a média (da população) ou o valor médio da dependente em termos dos valores conhecidos ou fixos (em amostragem repetida) das explicativas” (GUJARATI, 1996).

A análise de regressão é uma das técnicas estatísticas mais utilizadas para pesquisar e modelar a relação existente entre duas ou mais variáveis, procura avaliar a existência e o grau de dependência estatística entre as variáveis aleatórias, ou seja, aquelas que têm distribuição de probabilidade. Enquanto a análise de correlação consiste na medição do grau ou intensidade de associação entre duas variáveis. Quando se pode demonstrar que a variação de uma variável está de algum modo associada com a variação da outra, então podemos dizer que as duas variáveis estão correlacionadas.

Os modelos de regressão podem ser classificados segundo o número de variáveis independentes. Quando existe apenas uma variável independente denomina-se “Modelo de Regressão Simples”; quando se tem mais de uma variável independente denomina-se “Modelo de Regressão Múltipla”. Também se usa classificar de acordo com o tipo função que define o modelo, podendo ser representada por diferentes tipos de equações (linear, polinomial, etc.).

Existem vários métodos para construir uma equação de regressão, sendo o método dos mínimos quadrados o mais utilizado. Este método, atribuído ao matemático alemão Carl Friedrich Gauss, tem algumas propriedades estatísticas que fizeram dele um dos mais poderosos e populares métodos para construir uma equação de regressão. A reta de

regressão obtida por este método passa pela média da amostra dos valores das variáveis dependente e independente (X e Y), mostrando se elas realmente mantêm alguma relação, assim como se são positivamente ou negativamente correlacionadas e igualando a soma dos seus desvios a zero.

Quando o estudo da análise de regressão envolve situações onde existem mais de uma variável explicativa esse modelo de regressão recebe o nome de “Modelo de Regressão Múltipla”.

A regressão logística é semelhante à regressão linear e é usada quando temos uma variável dependente binária. O objetivo é saber quais as variáveis independentes que influenciam o resultado (variável dependente) e usá-las numa equação para prever o resultado de um indivíduo à custa das variáveis independentes.

Neste trabalho pretendemos analisar alguns dados recolhidos no âmbito de uma investigação levada a cabo na área da Medicina Dentária em que se avaliaram 158 indivíduos, sendo 79 diabéticos e 79 não diabéticos, relativos a variáveis bioquímicas, variáveis antropométricas, fatores comportamentais e indicadores de saúde oral.

O principal objetivo do estudo clínico de natureza epidemiológica foi avaliar a associação de diferentes fatores de risco estabelecidos e/ou potenciais na doença periodontal extensa definida pela perda de aderência clínica. Os critérios adotados na definição de caso, que serão descritos na Parte II, foram escolhidos de forma a evitar uma subavaliação da doença (nível de aderência clínica). Os dados recolhidos incluíram uma anamnese que abrangeu diversas condições sistémicas de interesse, dados analíticos referentes ao perfil lipídico e glicemia em jejum, e indicadores antropométricos, como o índice de massa corporal.

A inclusão das variáveis que nos propomos avaliar neste trabalho são aquelas consideradas como fatores de risco estabelecido para a periodontite. As variáveis analíticas relativas ao perfil lipídico e à glicemia não foram incluídas no modelo de estudo, que incluiu diabéticos e não diabéticos, pois a dislipidemia e hiperglicemia são muito mais prevalentes e graves nos doentes diabéticos do que nos não diabéticos. Porém, os dados analíticos foram utilizados em modelos que incluíram apenas diabéticos ou não diabéticos.

O principal objetivo do trabalho que no propomos a desenvolver é verificar se as associações entre fatores de risco e a variável dependente se mantêm, independentemente do método de análise multivariado usado. Os dados usados neste estudo, sendo reais, serviram apenas como “matéria-prima” para ser trabalhada por diferentes métodos estatísticos.

Para clarificar a utilização de um ajustado tratamento estatístico, abordaremos diferentes análises estatísticas com os dados existentes estudando a pertinência e a sua viabilidade, tanto em função do significado biológico das variáveis como dos seus valores, fazendo uma interpretação explicada do que se pode retirar de cada análise.

Com o intuito de se atingirem os objetivos enunciados, estruturou-se esta tese em duas partes principais: apresentação dos conceitos basilares envolvidos neste estudo e aplicação prática a uma investigação científica.

Na primeira parte abordaremos a teoria da análise de regressão, procurando clarificar este conceito, e através da revisão bibliográfica aprofundar conteúdos de interesse teórico para a fundamentação da metodologia adotada no nosso estudo. Seguidamente será revisto o tema da visualização de informação, contextualizando os gráficos que serão utilizados no decorrer do nosso exemplo de aplicação. Por fim abordaremos os aspetos patofisiológicos: diabetes mellitus e doença periodontal, tendo como objetivo uma melhor compreensão das variáveis que serão estudadas e objetos do nosso estudo clínico.

A segunda parte é dedicada à aplicação da análise de regressão a um estudo clínico, onde será explicado o desenho do estudo, as variáveis selecionadas, a caraterização da amostra e os procedimentos que serão efetuados.

Na análise dos dados, serão aplicados métodos de regressão a uma base de dados na área de Medicina Dentária. Assim, procurar-se-á promover uma discussão a nível metodológico e dos próprios resultados do estudo.

Por fim, serão apontadas as conclusões mais pertinentes do estudo, bem como algumas sugestões que se considerem adequadas. Ao proceder deste modo penso poder dar algum contributo para a aplicação correta da estatística em trabalhos biológicos, alertando para que a análise estatística tem de ser adequada e os resultados devem ser devidamente interpretados.

PARTE I

1. Análise de Variância (ANOVA) aplicada à Regressão

A comparação de médias de duas condições experimentais foi conseguida pelo t-test, descrito pela primeira vez por W.S. Gosset (1908) e publicado sob o pseudónimo de "Student". Porém, sempre que era necessário comparar mais do que duas condições numa experiência, aplicava-se mais do que um t-test, aumentando o erro tipo 1 (rejeição da $\boldsymbol{H_0}$ verdadeira). Este obstáculo foi ultrapassado por Sir Ronald Aylmer Fisher em 1925, que concebeu e descreveu o teste *Analysis Of Variance* (ANOVA) para analisar os dados de experiências agrícolas sem qualquer aumento do erro tipo 1. Em 1934, G.W. Snedecor utilizou a designação de distribuição F, como reconhecimento do trabalho de Fisher. Desde o seu aparecimento a ANOVA – como método inferencial para comparação de mais do que duas médias – tem sido aplicada por diferentes grupos de investigadores em distintas áreas do conhecimento, podendo estender-se à avaliação de modelos de regressão, nomeadamente modelos de Regressão Linear Simples e modelos de Regressão Linear Múltipla, aos quais dedicamos particular atenção neste trabalho.

A ANOVA é uma técnica poderosa que envolve a partição estatística da variância observada em diferentes componentes para realizar vários testes de significância. No nosso estudo aplicamos a ANOVA a um conjunto de dados para avaliar se existe uma relação linear entre uma variável dependente e uma variável independente e comparar médias entre grupos (diabéticos e não diabéticos). Também recorremos à ANOVA para avaliar a qualidade do ajuste dos modelos construídos.

A análise de variância (ANOVA) é uma metodologia estatística desenvolvida inicialmente com o objetivo de comparar $\boldsymbol{k} > 2$ amostras ou tratamentos, é utilizada para verificar se existem diferenças significativas entre as médias dos tratamentos, que sejam resultado dos efeitos dos tratamentos. O modelo linear subjacente a uma análise de variância a um fator é:

$$\boldsymbol{x_{ij} = \mu + \tau_i + \varepsilon_{ij}}$$

em que $\boldsymbol{x_{ij}}$ é cada uma das $\boldsymbol{j = 1, \dots, N_i}$ observações do tratamento $\boldsymbol{i}$, com $\boldsymbol{i = 1, \dots, k}$, $\boldsymbol{\mu}$ é a média global de todas as $\boldsymbol{N}$ observações, $\boldsymbol{\tau_i}$ é o efeito do tratamento *i*, isto é, a parte da variabilidade que pode ser imputada ao facto de cada uma das amostras ter sido objeto de um tratamento diferente, e $\boldsymbol{\varepsilon_{ij}}$ é a variabilidade residual ou erro experimental, isto é, a parte da variabilidade que não pode ser imputada aos tratamentos.

A ANOVA testa as hipóteses:

$$\boldsymbol{H_0}: \boldsymbol{\mu_1 = \mu_2 = \ ... \ = \mu_n = 0}$$

$$\boldsymbol{H_1}: \textit{As médias não são todas iguais}$$

Segundo o excerto do "*Studies in Crop Variation: An examination of the yield of dressed grain from Broadbalk",* Journal of Agriculture Science, 11, 107-135, 1921, a variação de qualquer quantidade (variável) que resulta de duas ou mais causas independentes é função da soma dos valores da variância produzida por cada causa separadamente. Esta propriedade da variância, pela qual cada causa independente, por si só, contribui para o total, permite-nos analisar o total, e atribuir, aproximadamente, as diferentes porções às causas apropriadas (ou grupo de causas).

Para a ANOVA ter validade têm que ser avaliados alguns pressupostos:

- O modelo deve ser aditivo, isto é, os efeitos devem-se somar (não há interação);
- Os erros (e_{ij}) devem ter distribuição normal;
- Os erros (e_{ij}) devem ser independentes;
- Os erros (e_{ij}) devem ter a mesma variância, ou seja, deve existir homocedasticidade.

Para testar a hipótese H_0, pode utilizar-se o teste F apresentado numa tabela de Análise de Variância. Convém lembrar que esse teste só é válido se os pressupostos assumidos para os erros do modelo estiverem satisfeitos.

Se $F_{calculado} > F_{tabelado}$, rejeita-se a hipótese de nulidade H_0, ou seja, existem evidências de diferença significativa entre pelo menos um par de médias de tratamentos, ao nível α de significância escolhido.

Para avaliar os pressupostos da ANOVA recorremos a métodos não paramétricos. Daí ser pertinente neste capítulo fazermos uma breve referência a alguns testes.

2. Técnicas de Inferência Estatística Não Paramétrica

As técnicas de Inferência Estatística Não Paramétrica surgem como um processo de colmatar problemas de difícil resolução no campo da Estatística Paramétrica, tendo neste trabalho constituído uma metodologia essencial, dadas as caraterísticas inerentes à amostra observada. Os métodos paramétricos, univariados e multivariados, obrigam muitas vezes a que sejam assumidos pressupostos que nem sempre são reflexo do modelo estudado e que, quando não são verificados, implicam graves erros na análise e conclusões.

Embora em 1710 se encontrem referências à utilização de métodos de estatística não paramétrica, estes surgem só bastante mais tarde, em 1942, com Wolfwitz, e só então se começa a verificar o seu desenvolvimento e um maior impacto, sendo hoje a estatística não paramétrica considerada como um ramo extremamente importante da Estatística.

A estatística não-paramétrica representa um conjunto de ferramentas de uso mais apropriado em pesquisas onde não se conhece bem a distribuição da população e seus parâmetros. Esse eventual desconhecimento da população reforça o estudo e a importância da análise de pesquisas através dos testes não-paramétricos.

A designação “Métodos ou Testes não Paramétricos” deve-se ao facto de estes não terem a pretensão de testar ou estimar parâmetros de uma dada distribuição, mas sim estudar o ajustamento de certas funções aos dados, averiguar a independência ou comparar duas ou mais distribuições». (OLIVEIRA, T., 2004).

Segundo MAROCO (2007), os testes não paramétricos são também conhecidos por “*distribution-free tests*”, ou seja, testes adistribucionais, uma vez que não necessitam do conhecimento prévio da distribuição da variável em estudo (normal), constituindo por isso uma boa alternativa aos testes paramétricos quando relativamente à variável em estudo falha a normalidade e a homogeneidade das variâncias entre os grupos.

As principais diferenças entre estes dois tipos de testes podem ser sintetizadas da seguinte forma:

Testes paramétricos (também denominados testes t):

- Exigem que as amostras tenham uma distribuição normal, especialmente se tiverem dimensão inferior a 30.
- Nas amostras de dimensão superior a 30, a distribuição aproxima-se da distribuição normal e também se aplicam testes t.

Testes não paramétricos:

- Não necessitam de requisitos tão fortes, como a normalidade, para serem usados. São também indicados quando as amostras são pequenas.
- São usados quando a amostra tem uma distribuição que não é normal ou quando, apesar da amostra ter uma dimensão superior a 30, se opta por conclusões mais conservadoras.
- A desvantagem destes testes é que não são tão potentes quanto os testes paramétricos, ou seja, com os testes não paramétricos não se encontram tantas diferenças entre os dados, quando essas diferenças realmente existem.

Ou seja, esquematicamente temos:

	Paramétricos	**Não Paramétricos**
Distribuição assumida	Normal	Qualquer uma
Variância assumida	Homogénea	Qualquer uma
Tipo de variáveis normalmente usadas	De intervalo ou rácio	Ordinal ou nominal
Relação entre os dados	Independentes	Qualquer uma
Medidas de localização central normalmente usadas	Média	Mediana

Quando precisamos de optar por um determinado tipo de teste devemos ter em conta o tipo de dados do estudo e qual o objetivo do estudo (o que pretendemos avaliar).

Tabela 1 : Teste a utilizar em função do tipo de dados e do objetivo do estudo

	TIPO DE DADOS		
OBJETIVO	Medida (de populações normais)	Ordem, resultado ou medida (de populações não normais)	Dicotómicos (dois resultados possíveis)
Descrever um grupo	Média e desvio padrão	Mediana e amplitude inter-quartil	Teste de Proporção
Comparar um grupo a um valor hipotético	Teste para uma só amostra (*one-sample t-test*)	Teste de Wilcoxon	Teste de Qui-quadrado ou Teste Binomial
Comparar 2 grupos independentes	Teste para duas amostras independentes *(Unpaired t-test)*	Teste de Mann-Whitney	Teste de Fisher ou Teste de Qui-quadrado
Comparar 2 grupos emparelhados	Teste para duas amostras emparelhadas *(Paired t-test)*	Teste de Wilcoxon	Teste de McNemar
Comparar 2 ou mais grupos independentes	*One-way ANOVA*	Teste de Kruskal-Wallis	Teste de Qui-quadrado
Comparar 2 ou mais grupos emparelhados	*Reapeted-measures ANOVA*	Teste de Friedman	Cochran Q
Quantificar a associação entre 2 variáveis	Correlação de Pearson	Correlação de Spearman	Coeficiente de contingência
Prever valores a partir de uma variável medida	Regressão linear simples ou regressão não linear	Regressão paramétrica	Regressão logística simples
Prever valores a partir de várias variáveis binomiais ou medida	Regressão linear múltipla		Regressão logística múltipla

Assim, mais especificamente, se analisarmos unicamente os testes não paramétricos vem:

Tabela 2 : Testes não paramétricos mais utilizados

		Escala de medida		
	Número de amostras	Nominal	Ordinal	Intervalo
	Uma amostra	**Teste de Qui-quadrado ou Teste Binomial**	**Teste de Kolmogorov-Smirnov para uma amostra** **Teste de iterações para uma amostra**	
Duas amostras	Amostras emparelhadas	**Teste de McNemar**	**Teste do sinal** ou **Teste de Wilcoxon**	**Teste de Walsh** **Teste de aleatoriedade para pares**
	Amostras independentes	**Teste de Fisher ou Teste de Qui-quadrado**	**Teste da mediana** **Teste de Mann-Whitney** **Teste de Kolmogorov-Smirnov para duas amostras** **Teste de Wald** **Teste de Moses para reações extremas**	**Teste de aleatoriedade para 2 amostras independentes**
K amostras	Amostras emparelhados	**Teste Q de Cochran**	**Teste de Friedman**	
	Amostras independentes	**Teste de Qui-quadrado** (para k amostra independentes)	**Teste de Kruskal-Wallis**	

Adaptado de Siegel (1975)

2.1. Teste do Qui-quadrado para a independência

O teste do Qui-quadrado permite verificar a independência entre duas variáveis, tendo por base uma disposição dos dados de acordo com uma tabela de contingência do tipo $r \times c$.

Genericamente, uma tabela de contingência resulta de uma classificação, segundo dois itens diferentes, de um mesmo grupo de indivíduos. Tem por objetivo inferir sobre a existência ou inexistência de relação entre as variáveis.

Considere-se, então, uma amostra de *n* indivíduos extraída de uma população, atendendo a dois critérios de classificação: X (variável 1) e Y (variável 2), cujos valores observados serão representados por O_{ij}, com $i = 1, \ldots, r$, e $j = 1, \ldots, c$. As frequências observadas podem apresentar-se numa "Tabela de Contingência" com r linhas e c colunas.

- ***Hipóteses a testar***

A formalização do teste de hipóteses, com a definição das hipóteses nula e alternativa, será apresentada da seguinte forma:

$\boldsymbol{H_0}$: Há independência entre as variáveis $\boldsymbol{X}$ e $\boldsymbol{Y}$
$\boldsymbol{H_1}$: Não há independência entre as variáveis $\boldsymbol{X}$ e $\boldsymbol{Y}$.

Designar-se-á, genericamente, por $\boldsymbol{X_i}$ $(\boldsymbol{i = 1, \dots, r})$ uma categoria da primeira variável e por $\boldsymbol{Y_j}$ $(\boldsymbol{j = 1, \dots, c})$ uma categoria da segunda variável, e os dados serão apresentados numa tabela de contingência, como se segue:

$\boldsymbol{X/Y}$	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	…	$\boldsymbol{Y_c}$	**Total**
$\mathbf{X_1}$	$\boldsymbol{O_{11}}$	$\boldsymbol{O_{12}}$	…	$\boldsymbol{O_{1c}}$	$\boldsymbol{O_{1*}}$
$\mathbf{X_2}$	$\boldsymbol{O_{21}}$	$\boldsymbol{O_{22}}$	…	$\boldsymbol{O_{2c}}$	$\boldsymbol{O_{2*}}$
…	…	…	…	…	…
$\mathbf{X_r}$	$\boldsymbol{O_{r1}}$	$\boldsymbol{O_{r2}}$	…	$\boldsymbol{O_{rc}}$	$\boldsymbol{O_{r*}}$
Total	$\boldsymbol{O_{*1}}$	$\boldsymbol{O_{*2}}$	…	$\boldsymbol{O_{*c}}$	$\boldsymbol{O_{**}}$

Onde:

$\boldsymbol{O_{ij}}$ representa os valores observados, $\boldsymbol{i = 1, \dots, r}$ e $\boldsymbol{j = 1, \dots, c}$,
e
$\boldsymbol{E_{ij}}$ representa os valores esperados, $\boldsymbol{i = 1, \dots, r}$ e $\boldsymbol{j = 1, \dots, c}$,

Sendo: $\boldsymbol{E_{ij} = \frac{O_{i*} \times O_{*j}}{n}}$, com $\boldsymbol{i = 1, \dots, r}$ e $\boldsymbol{j = 1, \dots, c}$

- ***Estatística de teste***

A estatística do teste é dada por: $\boldsymbol{\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$

Onde $\boldsymbol{\chi^2 \approx \chi^2_{(r-1)(c-1)(1-\alpha)}}$, sendo a aproximação tanto melhor quanto maior for a dimensão da amostra.

- ***Regra de decisão***

A regra de decisão é a seguinte:

Se $\boldsymbol{\chi^2 > \chi^2_{(r-1)(c-1)(1-\alpha)}}$, rejeita-se $\boldsymbol{H_0}$ ao nível de significância $\boldsymbol{\alpha}$.

2.2. Testes para duas ou mais amostras independentes

Teste de Mann-Whitney

Foi desenvolvido por F. Wilcoxon em 1945 para comparar as tendências centrais de duas amostras independentes de tamanhos iguais. Em 1947, H. B. Mann e D. R. Whitney generalizaram a técnica para amostras de tamanhos diferentes e passou a ser conhecido como o teste de Mann-Whitney. Este teste é pois um substituto do teste t de Student e é aplicável quando se verificam os seguintes pressupostos:

- Amostras aleatórias
- Observações independentes
- Variável de interesse tem caraterísticas contínua (mesmo que os dados não sejam contínuos)

O teste de Mann-Whitney é um teste não paramétrico que permite a comparação entre duas amostras independentes, de dimensões n_1 e n_2. É um teste alternativo ao teste t para duas amostras independentes. Enquanto que o teste t compara as médias de duas amostras independentes, o teste de Mann-Whitney compara o centro de localização das duas amostras, como forma de detetar diferenças entre as duas populações correspondentes. As vantagens do teste de Mann-Whitney são: não exigir o pressuposto da normalidade, podendo ser aplicado para amostras pequenas e em variáveis de escala ordinal.

Considerem-se duas amostras independentes:

$\boldsymbol{x_1, x_2, \dots, x_n}$, retirada da população $\boldsymbol{X}$

$\boldsymbol{y_1, y_2, \dots, y_n}$, retirada da população $\boldsymbol{Y}$

E suponha-se que $\boldsymbol{n_1 < n_2}$ com um total de $\boldsymbol{n = n_1 + n_2}$

- ***Hipóteses a testar***

$\boldsymbol{H_0}$: As duas amostras são provenientes de populações com a mesma distribuição.
$\boldsymbol{H_1}$: As duas amostras são provenientes de populações com distribuições distintas.

As hipóteses anteriores podem ser reformuladas, se $\boldsymbol{F(x)}$ definir a função distribuição da população $\boldsymbol{X}$ e $\boldsymbol{G(x)}$ a função distribuição da população $\boldsymbol{Y}$. As hipóteses podem então ser escritas do seguinte modo:

$$\boldsymbol{H_0: F(x) = G(x) \forall x}$$

$$\boldsymbol{H_1: \exists x: F(x) \neq G(x)}$$

A forma como o teste é construído torna-o particularmente sensível às diferenças de medidas de localização, especialmente às diferenças nas medianas das distribuições.

Em vez de se basear em parâmetros da distribuição normal como a média e a variância, o teste de Mann-Whitney baseia-se nas ordenações da variável.

- ***Estatística de teste***

A estatística de teste baseia-se nas ordens (*ranks*) das observações das amostras.

Combinam-se as duas amostras, o total das $\boldsymbol{n}$ observações e ordenam-se estas por ordem crescente assinalando o grupo a que pertencem. No caso das observações empatadas atribui-se a média dada pela posição sequencial das observações que lhe corresponderiam.

A estatística de teste é dada por:

$$\boldsymbol{U = \min\left(U_1, U_2\right)}$$

Em que: $\boldsymbol{U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1}$ e $\boldsymbol{U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2}$

Onde: $\boldsymbol{n_1}$ é a dimensão da amostra menor
$\boldsymbol{n_2}$ é a dimensão da amostra maior
$\boldsymbol{R_1}$ é a soma das ordenações da menor amostra
$\boldsymbol{R_2}$ é a soma das ordenações da maior amostra

● Quando $\boldsymbol{n_1}$ e $\boldsymbol{n_2}$ pequenos ($\leq \mathbf{10}$) – Compara-se o valor observado da estatística de teste com o valor tabelado (Tabela de Mann-Whitney).

- ***Regra de decisão***

Se $\boldsymbol{U < U_{tabelado}}$, então pertence à região crítica, pelo que se rejeita a hipótese nula ao nível de significância $\boldsymbol{\alpha}$.

● Quando $\boldsymbol{n_1}$ e $\boldsymbol{n_2}$ grandes (> 10) – Normalmente recorre-se à aproximação à normal.

A estatística de teste é dada por:

$$\boldsymbol{Z = \frac{U - \mu_v}{\sigma_v} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \cap N(0,1)}$$

$$\boldsymbol{\mu_v = \frac{n_1 n_2}{2}} \text{ e } \boldsymbol{\sigma_v = \sqrt{\frac{n_1 n_2 (n+1)}{12}}}$$

Se o valor observado da estatística de teste pertencer à região não crítica:

$-\boldsymbol{z}\frac{\boldsymbol{\alpha}}{\mathbf{2}} < Z < z\frac{\boldsymbol{\alpha}}{\mathbf{2}}$ para um nível de significância $\boldsymbol{\alpha}$, não se rejeita $\boldsymbol{H_0}$.

O teste Mann-Whitney pode ser aplicado em situações em que existem empates nas observações e em situações em que não ocorrem empates.

- ***Correção para empates***

No caso em que ocorrem empates entre duas ou mais observações da mesma amostra, o valor de $\boldsymbol{U}$ não é afetado. Mas se os empates envolvem elementos das duas amostras e ocorrem entre duas ou mais observações, o valor de $\boldsymbol{U}$ pode ser afetado. A correção para empates deve ser feita ao desvio padrão da distribuição amostral $\boldsymbol{U}$.

$$\boldsymbol{\sigma_v} = \sqrt{\frac{\boldsymbol{n_1 n_2}}{\boldsymbol{n(n-1)}} \times \left(\frac{\boldsymbol{n^3 - n}}{\mathbf{12}} \sum \frac{\boldsymbol{t_j^3 - t_j}}{\mathbf{12}}\right)}$$

Em que $\boldsymbol{t_j}$ corresponde ao número de observações empatadas em cada grupo $\boldsymbol{j}$.

Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é aplicado quando estão em comparação três ou mais grupos independentes e a variável deve ser de mensuração ordinal.

A aplicação da análise de variância paramétrica, normalmente abreviada para ANOVA, pressupõe a verificação de alguns pressupostos, designadamente:

1. A normalidade da sua distribuição
2. A homogeneidade da variância
3. Independência

Para além disto, a análise de variância paramétrica só pode ser aplicada quando a variável dependente admite pelo menos a escala de intervalos como nível de medida. Quando não se verificam os pressupostos da análise de variância paramétrica e/ou quando o nível de medida mais restrito da variável dependente é a escala ordinal, o teste de Kruskal-Wallis que é uma generalização, para k>2 amostras, do teste de Mann-Whitney. Pode ser considerado como a alternativa não paramétrica à ANOVA *one-way* (KRUSKAL e WALLIS,1952). Este teste destina-se a verificar se há diferenças na localização das populações (com distribuições contínuas) subjacentes aos *n* grupos.

- ***Hipóteses a testar***

$\boldsymbol{H_0}$: As *n* populações têm a mesma localização
$\boldsymbol{H_1}$: Pelo menos duas das *n* populações não têm a mesma localização

O procedimento a aplicar para efetuar o teste de Kruskal-Wallis é semelhante ao do teste de Mann-Whitney: ordenam-se as *N* observações em conjunto e atribuem-se-lhes *ranks* (posições: $\mathbf{1}$; $\mathbf{2}$; …; $\boldsymbol{N}$).

Quando há empates (observações repetidas) atribui-se o *rank* médio às observações empatadas.

A ideia base do teste é a de que, se $\boldsymbol{H_0}$ for verdadeira, os *ranks* correspondentes aos vários grupos estarão misturados de forma aleatória; caso contrário, deverão existir grupos com predominância de *ranks* reduzidos e outros grupos com predominância de *ranks* elevados.

O teste de Kruskal-Wallis baseia-se na comparação entre a média dos valores de ordem das diversas amostras e não na comparação entre as médias amostrais da variável dependente, uma vez que nem sempre é possível calcular as médias amostrais (esse cálculo só é possível quando as variáveis dependentes admitem como nível de medida mais restrito a escala de intervalos).

Considerem-se então $\boldsymbol{k}$ populações, $\boldsymbol{X_1, X_2, \dots, X_k}$, a partir das quais foram retiradas k amostras aleatórias, de dimensões $\mathbf{n_1, n_2, \dots, n_k}$.

Deste modo, tem-se:

($\boldsymbol{X_{11}, X_{12}, \dots, X_{1n}}$) da população $\boldsymbol{X_1}$

Considere-se $R(X_0)$ a ordem (*rank*) atribuída à observação e seja:

$$\boldsymbol{R_i = \sum_{j=1}^{n_i} R(X_{ij})}$$

a soma das ordens da i-ésima amostra ($\boldsymbol{i = 1, 2, \dots, k}$).

O número total de observações é $\boldsymbol{N = \sum_{1=1}^{k} n_i}$

- ***Estatística de teste***

A estatística de Kruskal-Wallis é dada por:

$$\boldsymbol{T = \frac{1}{s^2}\left[\sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{n(n+1)^2}{4}\right]}$$

Onde $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} R(X_{ij})^2 - \frac{n(n-1)^2}{4}\right]$

Para o caso de não existirem empates (ou de o seu número ser muito pequeno), esta estatística reduz-se a:

$$T = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)$$

A distribuição por amostragem da estatística de teste depende do número de amostras envolvidas na comparação, bem como do número de observações por amostra.

- ***Correção para empates***

Assim, para os casos em que não há empates, e se o número de amostras for inferior ou igual a três ou o número de observações por amostra não ultrapassar as 5, para tomar a decisão quanto à diferença nas distribuições das três populações, compara-se o valor da estatística de teste com os valores fornecido pela tabela de Kruskal-Wallis.

Para os casos em que não há empates, e se o número de amostras é superior a três ou a dimensão de uma amostra é superior a 5, a estatística de teste pode ser aproximada à distribuição do Qui-quadrado com $k - 1$ graus de liberdade (distribuição assimptótica).

Esta aproximação será tanto melhor quanto maiores forem as amostras e a dimensão de cada amostra.

- ***Regra de decisão***

Rejeita-se H_0 se o valor da estatística de teste for superior ao valor tabelado.

Quando existem observações com o mesmo valor (empates), é importante corrigir o valor do teste. Se mais de 25% das observações forem empates, a estatística de teste T´ deverá ser corrigida dividindo T´ por:

$$1 - \frac{1}{n^3 - n}\sum_{j=1}^{e} t_j(t_j^2 - 1)$$

onde e representa o número de amostras com diferentes ordens de empates e t_j representa o número de empates na amostra j.

O procedimento do teste de Kruskal-Wallis pode ser resumido da seguinte forma:

1) Ordenar as observações das k amostras num único grupo atribuindo-lhes ordens de 1 a n.
2) Calcular R_i para cada amostra (soma das ordens).
3) Calcular o valor da estatística T.

A regra de decisão será:

1) Para $\boldsymbol{k = 3}$ e $\boldsymbol{n_1, n_2, n_3 \leq 5}$, recorrer à tabela. Se $\boldsymbol{T > t_{crit,\alpha}}$ rejeitar $\boldsymbol{H_0}$ ao nível de significância $\boldsymbol{\alpha}$.

2) Se pelo menos uma das amostras tiver dimensão $\boldsymbol{n_i} > 5$, deve usar-se a distribuição do Qui-quadrado. Se $\boldsymbol{\chi^2_{k-1}}$, então rejeita-se $\boldsymbol{H_0}$ ao nível de significância $\boldsymbol{\alpha}$.

Se o valor observado da estatística de teste pertencer à região crítica então isso significa que existem diferenças significativas entre as amostras.

Para identificar onde se situa a diferença é necessário proceder a comparações dos grupos, dois a dois, o que corresponde a efetuar $\binom{k}{2}$ testes.

As hipóteses podem ser definidas, para um certo par de grupos $(\boldsymbol{i, j})$, $\boldsymbol{i \neq j}$, do seguinte modo:

$\boldsymbol{H_0}$: A distribuição da população $\boldsymbol{i}$ é idêntica à distribuição na população $\boldsymbol{j}$, $\forall_{(i,j)} \boldsymbol{i \neq j}$

$\boldsymbol{H_1}$: A distribuição da população $\boldsymbol{i}$ difere da distribuição na população $\boldsymbol{j}$, para algum $\boldsymbol{i \neq j}$

A regra de decisão é dada pela seguinte expressão:

$$\left|\frac{\boldsymbol{R_i}}{\boldsymbol{n_i}} - \frac{\boldsymbol{R_j}}{\boldsymbol{n_i}}\right| > \boldsymbol{t_{n-k;1-\frac{\alpha}{2}}} \times \sqrt{\boldsymbol{s^2} \times \frac{\boldsymbol{n-1-t}}{\boldsymbol{n-k}}} \times \sqrt{\frac{\boldsymbol{1}}{\boldsymbol{n_i}} - \frac{\boldsymbol{1}}{\boldsymbol{n_j}}}$$

Isto é, a primeira desigualdade será superior à segunda se existirem diferenças entre o grupo $\boldsymbol{i}$ e o $\boldsymbol{j}$. Assim, $\boldsymbol{t_{n-k;1-\frac{\alpha}{2}}}$ corresponde ao valor da probabilidade $\boldsymbol{1 - \frac{\alpha}{2}}$ de uma variável aleatória com distribuição *t*-Student, $\boldsymbol{n - k}$ graus de liberdade.

Teste de Kolmogorov-Smirnov

O teste paramétrico tradicional, baseado na distribuição *t*-Student, é obtido sob a hipótese de que a população tem distribuição normal. Nesse sentido, surge a necessidade de certificarmos se essa suposição pode ser assumida. Em alguns casos, assumir a normalidade dos dados é o primeiro passo que tomamos para simplificar a nossa análise. Para dar suporte a esta suposição, consideramos, dentre outros, o teste de Kolmogorov-Smirnov.

- ***Hipóteses a testar***

O teste de Kolmogorov-Smirnov pode ser utilizado para avaliar as hipóteses:

$\boldsymbol{H_0}$: Os dados seguem uma distribuição normal

$\boldsymbol{H_1}$: Os dados não seguem uma distribuição normal

Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada assumida para os dados, no caso a normal, e a função de distribuição empírica dos dados. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância.

Considere-se uma amostra aleatória simples, $\boldsymbol{x_1, x_2, \dots, x_n}$ de uma população com função de distribuição acumulada contínua $\boldsymbol{F_x}$ desconhecida. A estatística utilizada para o teste é:

$$\boldsymbol{D_n = \sup_x |F(x) - F_n(x)|}$$

Esta função corresponde a distância máxima vertical entre os gráficos de $\boldsymbol{F(x)}$ e $\boldsymbol{F_n(x)}$ sobre a amplitude dos possíveis valores de $\boldsymbol{x}$. Em $\boldsymbol{D_n}$ temos que:

- $\boldsymbol{F(x)}$ representa a função de distribuição acumulada assumida para os dados;
- $\boldsymbol{F_n(x)}$ representa a função de distribuição acumulada empírica dos dados.

Neste caso, queremos testar a hipótese $\boldsymbol{H_0: F_x = F}$ $\boldsymbol{versus}$ $\boldsymbol{H_1: F_x \neq F}$. Para isto, tomamos $\boldsymbol{X_{(1)}, X_{(2)}, \dots, X_{(n)}}$ as observações aleatórias ordenadas de forma crescente da população com função de distribuição contínua $\boldsymbol{F_x}$. No caso de análise da normalidade dos dados, assumimos $\boldsymbol{F}$ a função de distribuição da normal.

A função de distribuição acumulada assumida para os dados é definida por $\boldsymbol{F(x_{(i)}) = P(X \leq x_i)}$ e a função de distribuição acumulada empírica é definida por uma função escada, dada pela fórmula: $\boldsymbol{F_n(x) = \frac{1}{n}\sum_{1=1}^{n} I_{]-\infty,x]}(x_{(i)})}$

onde $\boldsymbol{I_A}$ é a função indicadora. A função indicadora é definida da seguinte forma:

$$\boldsymbol{I_A = \begin{cases} 1; se\ x \in A \\ 0; caso\ contrário \end{cases}}$$

Observe a função da distribuição empírica $\boldsymbol{F_n(x)}$ corresponde à proporção de valores menores ou iguais a $\boldsymbol{x}$. Tal função também pode ser descrita da seguinte forma:

$$F_n(x) = \begin{cases} 0, se\ x < x_{(1)} \\ \frac{k}{n}, se\ x_{(k)} \leq x < x_{(k+1)} \\ 1, se\ x > x_{(n)} \end{cases}$$

- ***Estatística de teste***

Sob H_0, a distribuição assintótica da estatística de kolmogorov-Smirnov é dada por:

$$\lim_{n\to\infty} P\left[\sqrt{n}D_n \leq x\right] = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}\ exp^{-2j^2x^2}$$

Esta distribuição assintótica é válida quando temos conhecimento completo sobre a distribuição de H_0, entretanto, na prática, H_0 especifica uma família de distribuições de probabilidade. Neste caso, a distribuição assintótica da estatística de Kolmogorov-Smirnov não é conhecida e foi determinada via simulação.

Como a função de distribuição empírica F_n é descontínua e a função de distribuição hipotética é contínua, vamos considerar duas outras estatísticas:

$$D^+ = \sup_{x_{(i)}} \left|F(x_{(i)}) - F_n(x_{(i)})\right|$$

$$D^- = \sup_{x_{(i)}} \left|F(x_{(i)}) - F_n(x_{(i-1)})\right|$$

para calcularmos a estatística de Kolmogorov-Smirnov. Essas estatísticas medem as distâncias (vertical) entre os gráficos das duas funções, teórica e empírica, nos pontos $x_{(i-1)}$ e $x_{(i)}$. Com isso, podemos utilizar como estatística de teste:

$$D_n = max(D^+, D^-)$$

- ***Regra de decisão***

Se D_n é maior que o valor crítico, rejeitamos a hipótese de normalidade dos dados com $(1-\alpha)100\%$ de confiança. Caso contrário, não rejeitamos a hipótese de normalidade.

Teste de Wald

O teste de Wald é obtido por comparação entre a estimativa de máxima verosimilhança do parâmetro ($\hat{\beta}_1$) e a estimativa de seu erro padrão.

- ***Hipóteses a testar***

A razão resultante, sob a hipótese H_0: $\beta_1 = 0$ tem distribuição normal padrão.

- ***Estatística de teste***

A estatística do teste Wald para a regressão logística é $\boldsymbol{W_j = \frac{\widehat{\beta}_1}{\widehat{DP}\,(\widehat{\beta}_1)}}$.

O p-valor é definido como $\boldsymbol{P(|Z| > |W_j|)}$, sendo que $\boldsymbol{Z}$ denota a variável aleatória da distribuição normal padrão.

HAUCK e DONNER (1977) examinaram o desempenho do teste de Wald e descobriram que ele se comporta de maneira estranha, em determinadas situações, frequentemente não rejeitando a hipótese nula quando o coeficiente é significativo. Eles recomendam a utilização do teste da razão de verosimilhança para testar se realmente o coeficiente não é significativo quando o teste de Wald não rejeita a hipótese nula.

Teste de Fisher

Em tabelas de contingência $\boldsymbol{2x2}$, valores esperados menores que 5 e amostras pequenas podem ter como efeito que a aproximação da distribuição Qui-quadrado para a estatística $\boldsymbol{\chi^2_{obs}}$ não seja suficientemente boa.

Neste caso é preferível usar o teste exato de Fisher, que passaremos a descrever. Neste teste baseámo-nos no cálculo da distribuição de probabilidade das frequências da tabela. Contudo isso não é possível na situação das tabelas com margens livres ou com uma margem fixa e outra livre, porque a probabilidade de uma dada distribuição das frequências é função de parâmetros de valor desconhecido.

Fisher (1934) propôs que a distribuição de probabilidade das frequências de qualquer um destes tipos de tabelas sejam substituídas pela probabilidade da distribuição das mesmas frequências considerando tabelas com duas margens fixas, ou seja uma distribuição de probabilidade hipergeométrica para a única frequência de valor livre (independente).

Tabela 3 : Tabela de Contingência $\boldsymbol{2x2}$

		Variável Coluna		**Total**
		1	**2**	
Variável linha	**1**	**A**	**B**	**A+B**
	2	**C**	**D**	**C+D**
Total		**A+C**	**B+D**	**n = A+B+C+D**

Para a tabela 3 (arranjada de modo a que ($n_{1.} \leq n_{.1} \leq n_{.2} \leq n_{2.}$), se X for a frequência de valor independente, neste caso a frequência da célula (1,1), considerando:

$$P_a = P|X = a| = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

$a-1$	$b+1$	$a+b$
$c+1$	$d-1$	$c+d$
$a+c$	$b+d$	n

$a-2$	$b+2$	$a+b$
$c+2$	$d-2$	$c+d$
$a+c$	$b+d$	n
⋮	⋮	⋮
0	$a+b$	$a+b$
$c+a$	$d-a$	$c+d$
$a+c$	$b+d$	n

o teste exato de Fisher consiste na determinação desta probabilidade e a dos arranjos possíveis que, com os mesmos totais marginais, tenham ainda mais desvio em relação à hipótese nula, isto é, as probabilidades de tabelas com as mesmas margens e com menores valores na entrada cujo valor, na tabela de contingência em questão, já foi considerado na tabela à esquerda.

Se a soma $P_a + P_{a-1} + \ldots + P_0$ for inferior ao nível de significância que escolhermos para o nosso teste, devemos rejeitar a hipótese de independência ou a hipótese de homogeneidade que estipulamos.

2.3. Testes de Correlação

Teste de Spearman

O coeficiente de correlação de Spearman é a mais antiga estatística baseada em postos e a sua utilização remonta a 1904. É utilizada para avaliar o grau de correlação entre variáveis quantitativas quando as exigências para o teste de Pearson não são satisfeitas (distribuição bivariada normal e homocedasticidade).

O teste de Spearman considera uma população da qual foi retirada uma amostra de dimensão *n* de pares ordenados de duas variáveis aleatórias x e y. Considera ainda que as mesmas variáveis são ordenadas de forma crescente e lhes é atribuído um número de ordem. Estas variáveis podem encontrar-se associadas de uma forma direta ou de uma forma inversa como se mostra na tabela seguinte:

A		B	
Variável x (Nº de ordem)	Variável y (Nº de ordem)	Variável x (Nº de ordem)	Variável y (Nº de ordem)
1	1	1	n
2	2	2	n-1
…	…	…	…
n-1	n-1	n-1	2
n	n	n	1

Considere-se d_i (com i = 1, 2, 3, …, n) a diferença entre os números de ordem de cada par de observações x_i e y_i. Em presença de uma associação direta, teremos: $\sum_{i=1}^{n} d_i^2 = 0$. No caso de uma associação inversa, teremos: $\sum_{i=1}^{n} d_i^2 = \frac{n(n^2-1)}{3}$, e o coeficiente de correlação de Spearman é dado por: $\rho_S = 1 - \frac{6\times\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$.

O coeficiente ρ_S assume o valor 1 quando entre o conjunto das observações existe uma associação direta perfeita e assume o valor -1 quando se verificar uma associação inversa perfeita. Quando não se verificar qualquer associação entre as variáveis x e y, o coeficiente ρ_S assume valores próximos de zero.

- ***Hipóteses a testar***

A partir deste coeficiente pode ser construído um teste bilateral para testar em que:

$\boldsymbol{H_0}$: As variáveis não se encontram associadas

$\boldsymbol{H_1}$: As variáveis encontram-se associadas

Como hipóteses alternativas de $\boldsymbol{H_1}$ podem ainda considerar-se:

$\boldsymbol{H_1}$: Associação direta (teste unilateral à direita)

$\boldsymbol{H_1}$: Associação inversa (teste unilateral à esquerda)

- ***Estatística do teste***

Quando a amostra for superior a 30 observações, a estatística de teste deverá ser substituída por:

$$\frac{\rho_S}{\sqrt{(1-\rho_s^2)/(n-2)}}$$

Quando $\boldsymbol{H_0}$ é verdadeira, segue uma distribuição t de Student com (n-2) graus de liberdade.

- ***Correção para empates***

Sempre que se verifiquem empates, atribui-se às observações naquela situação o número de ordem que corresponde à média dos números de ordem que as observações receberiam se não estivessem empatadas. Se existir um número pequeno de empates, o valor da estatística ρ_s deve ser calculado através da expressão:

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

e não será significativamente afetado. Caso contrário, a estatística ρ_S deverá ser calculada através da expressão:

$$\rho_S = \frac{n(n^2-1) - 6\sum_{i=1}^{n} d_i^2 - 6\left(u' + v'\right)}{\sqrt{n(n^2-1) - 12u'}.\sqrt{n(n^2-1) - 12v'}}$$

Sendo

$u' = \frac{\sum_{i=1}^{n} u_i^3 - \sum_{i=1}^{n} u_i}{12}$ e $v' = \frac{\sum_{i=1}^{n} v_i^3 - \sum_{i=1}^{n} v_i}{12}$

em que u_i e v_i representam o número de empates no i-ésimo grupo de observações iguais pertencentes, respetivamente, à variável x e à variável y.

3. Análise de Regressão

O modelo linear normal, criado no início do século XIX por Legendre e Gauss, dominou a modelação estatística até meados do século XX, embora vários modelos não lineares ou não normais tenham entretanto sido desenvolvidos para fazer face a situações que não eram adequadamente explicadas pelo modelo linear normal.

São exemplo disso, tal como referem McCULLAGH and NELDER (1989) e LINDSEY (1997), o modelo complementar log-log para ensaios de diluição (Fisher, 1922), os modelos *probit* (BLISS, 1935) e *logit* (BERKSON, 1944; DYKE and PATTERSON, 1952; RASCH, 1960) para proporções, os modelos log-lineares para dados de contagens (BIRCH,1963), os modelos de regressão para análise de sobrevivência (FEIGL and ZELEN, 1965; ZIPPIN and ARMITAGE, 1966; GLASSER, 1967).

Todos os modelos anteriormente descritos apresentam uma estrutura de regressão linear e têm em comum o facto da variável resposta seguir uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a família exponencial.

Os Modelos Lineares Generalizados introduzidos por NELDER e WEDDERBURN (1972) correspondem a uma síntese destes e de outros modelos, vindo assim unificar, tanto do ponto de vista teórico como concetual, a teoria da modelação estatística até então desenvolvida.

São pois casos particulares dos modelos lineares generalizados (MLG) os seguintes modelos:

- 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.

Neste estudo recorreremos ao modelo de regressão linear clássico e ao modelo de regressão logística.

Análise de regressão é uma técnica de modelação utilizada para analisar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes $X_1, X_2, X_3, \dots X_n$. O objetivo desta técnica é identificar e estimar uma função que descreva, o mais próximo possível, a

relação entre essas variáveis e que assim irá permitir predizer o valor que a variável dependente (Y) irá assumir para um determinado valor da variável independente X.

O modelo de regressão poderá ser escrito genericamente como:

$$Y = f(X_1, X_2, X_3, \dots X_n) + \varepsilon$$

onde o termo ε representa uma perturbação aleatória na função, ou o erro da aproximação. O número de variáveis independentes varia entre aplicações: quando se tem apenas uma variável independente, denomina-se Modelo de Regressão Simples; quando se tem mais de uma variável independente, denomina-se de Modelo de Regressão Múltipla. A forma da função também varia, podendo ser representada por uma equação linear, polinomial ou outro mesmo tipo de função (simples ou multivariada).

3.1. Regressão e Correlação Linear

Testes de Hipóteses sobre o Coeficiente de Correlação

A correlação entre duas variáveis é determinada numericamente por meio dos coeficientes de correlação que representam o grau de associação entre duas variáveis contínuas e designa-se por ρ.

O coeficiente de correlação linear, também chamado de covariância normalizada, é representado por:

$$\rho_{x,y} = \frac{\sigma_{x,Y}}{\sigma_X \sigma_Y}$$

Onde: $\sigma_{x,Y}$ é a covariância entre as variáveis X e Y

σ_x e σ_Y são os desvios padrão das variáveis X e Y

A covariância entre duas variáveis pode ser estimada pela equação:

$$s_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Onde: $s_{X,Y}$ é a covariância amostral entre as variáveis X e Y

$\bar{x}$ e $\bar{y}$ são as médias aritméticas de cada uma das variáveis

n o tamanho da amostra

x_i e y_i são as observações simultâneas das variáveis

Admitindo-se que a distribuição conjunta das variáveis é normal bivariada, torna-se conveniente utilizar, como medida da correlação, o coeficiente de correlação de Pearson cujo estimador é dado por:

$$\hat{\rho} = \frac{s_{x,Y}}{s_X s_Y}$$

Onde: $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ e $s_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$ são os desvios padrão das amostras.

Para se decidir sobre a existência de correlação e o sentido da variação da reta de regressão, calcula-se ρ e o erro de ρ, e seguidamente efetua-se um teste de *t*-Student, para as seguintes hipóteses:

H_0: $\rho = 0$, a reta de regressão em y é paralela ao eixo das abcissas.
H_1: $\rho \neq 0$, a reta de regressão em y não é paralela ao eixo das abcissas.

A estatística do teste é $t_0 = \frac{\hat{\rho}\sqrt{n-1}}{\sqrt{1-r^2}}$

Onde: t_0 é a estatística do teste
n o tamanho da amostra
$\hat{\rho}$ é a estimativa do coeficiente de correlação linear

Para encontrar o $t\ crítico\ (t_c)$ consulta-se uma tabela de *t*-Student, e é interpretado conforme o seguinte critério:

$\boldsymbol{t < t_c}$	$\boldsymbol{t_c}$	$\boldsymbol{t > t_c}$
$\boldsymbol{t}$ não é significativo		$\boldsymbol{t}$ é significativo
$\boldsymbol{\rho}$ não é significativamente diferente de $\mathbf{0}$		$\boldsymbol{\rho}$ é significativamente diferente de $\mathbf{0}$
(a reta é paralela ao eixo dos $\boldsymbol{xx}$)		(a reta não é paralela ao eixo dos $\boldsymbol{xx}$)

Quando a escala de medida é ordinal devemos utilizar o coeficiente de correlação de Spearman pois este, ao contrário do coeficiente de correlação de Pearson, não requer a suposição que a relação entre as variáveis é linear, nem requer que as variáveis sejam medidas em intervalo de classe, podendo ser usado para as variáveis medidas no nível ordinal.

É importante realçar que as correlações ordinais não podem ser interpretadas da mesma maneira que as correlações de Pearson. Inicialmente não mostram tendência linear, mas podem ser consideradas como índices de monotonia, ou seja, permitem-nos avaliar as

variações para aumentos positivos da correlação (aumentos no valor de X correspondem a aumentos no valor de Y) e para coeficientes negativos.

3.2. Modelo de Regressão Linear Simples

Um modelo de regressão linear simples (MRLS) descreve uma relação entre uma variável independente (explicativa ou regressora) X e uma variável dependente (resposta) Y, nos termos seguintes:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

Onde: β_0 e β_1 são constantes (parâmetros) desconhecidas
ε é o erro aleatório

De uma maneira geral, os problemas de regressão e de correlação podem apresentar-se sob diferentes aspetos, sobretudo dependendo da natureza dos dados e do objetivo do estudo. O caso mais simples é aquele em que uma das variáveis em estudo assume apenas certos valores, escolhidos *a priori*, de maneira arbitrária. Nesse caso, a variável independente, geralmente designada pela letra X, não é aleatória; porém, a variável independente Y, é aleatória. O objetivo final consiste em estimar o valor da variável dependente em função da variável independente.

3.2.1. Reta de Regressão

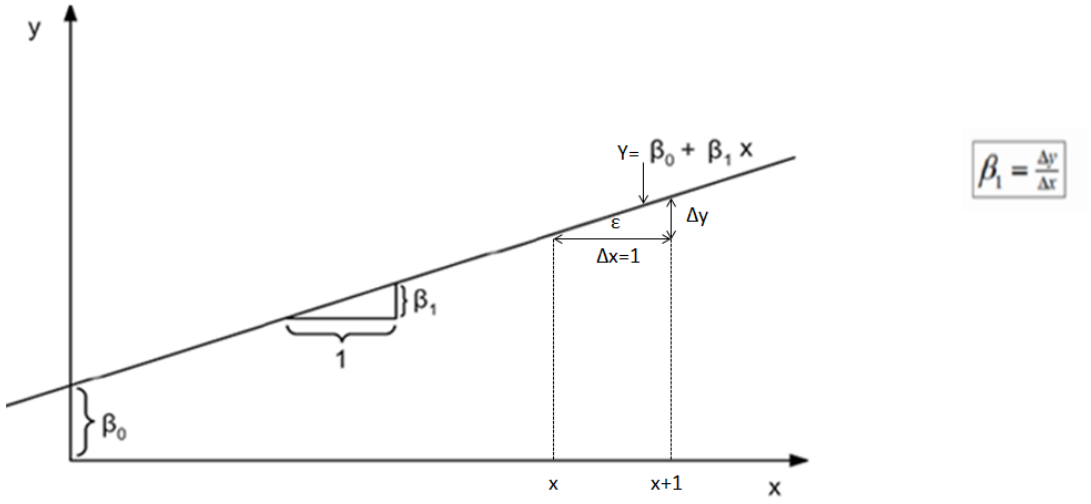


Figura 1: Interpretação geométrica dos parâmetros do modelo de regressão linear simples

O diagrama de dispersão é um gráfico constituído por pontos onde cada ponto, P_i, representa um par de valores observados, (x_i, y_i), (x_i representa o valor da variável independente observada para o indivíduo P_i e y_i representa o valor da variável dependente observada para esse mesmo indivíduo). O diagrama de dispersão é obtido pelos pontos posicionados em torno da reta de regressão.

O diagrama de dispersão tem uma função dupla: analisar se existe alguma associação entre as variáveis e permitir identificar qual o modelo matemático (equação) mais apropriado para descrever essa associação.

Quando o diagrama de dispersão indica uma tendência para uma relação linear, então os dados encontram-se bem ajustados pela reta de regressão (de equação (1)).

Ao ajustar uma reta de regressão aos dados observados anulamos os efeitos da variável residual. Verifica-se que nem todos os pontos se encontram sobre a reta e essa diferença é o erro (ε), que pode ter sido ocasionado por fatores distintos. Mas supõe-se que a média desses erros tende a anular-se, ou seja: $E\ (\varepsilon_i) = 0$.

A obtenção da reta ajustada implica o conhecimento dos parâmetros β_0 e β_1 de tal modo que o desvio entre os valores reais e os valores ajustados seja mínimo. Um método que permite minimizar estes desvios é o método dos mínimos quadrados.

3.2.2. Método dos Mínimos Quadrados

Uma vez escolhido o modelo de regressão, deve-se estimar os seus parâmetros, neste caso os coeficientes da equação da reta, β_0 e β_1. Isso pode ser feito a partir da aplicação do Método dos Mínimos Quadrados. Calculando a média sobre a equação (1), temos:

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \qquad (2)$$

uma vez que a média dos erros é zero.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i\, x_i - n.\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}} \qquad (3)$$

$$e_i = y_i - \hat{y_i} \text{ (desvios)}$$

São vantagens do método dos mínimos quadrados:

- Obter as melhores estimativas, pois elas não são enviesadas;
- Ter em conta os desvios maiores, diluindo o efeito dos maiores valores;
- Permitir realizar testes de significância na equação de regressão;
- A reta de regressão passa pelo ponto obtido pelo cálculo das médias das duas amostras.

Subtraindo as duas equações (1-2) temos:

$$Y_i - \bar{Y} = (\beta_0 - \beta_1) + \beta_1(X_i - \bar{X}) + \varepsilon_i \qquad (4)$$

Denominando de y e x as diferenças centradas nas médias, $Y_i - \bar{Y}$ e $X_i - \bar{X}$ respetivamente, temos que:

$$y_i = \beta_1 x_i + \varepsilon_i \quad \text{ou } \varepsilon_i = y_i - \beta_1 x_i \qquad (5)$$

Fazendo a soma dos quadrados dos erros (5),

$$\sum(\varepsilon_i)^2 = \sum(y_i - \beta_1 x_i)^2$$

$$\sum(\varepsilon_i)^2 = \sum y_i^2 - \sum 2\beta_1 x_i y_i + \sum \beta_1^2 x_i^2$$

como β_1 é uma constante,

$$\sum(\varepsilon_i)^2 = \sum y_i^2 - 2\beta_1 \sum x_i y_i + \beta_1^2 \sum x_i^2$$

Como o objetivo é estimar uma equação que minimize os erros, devemos então derivar a equação acima em relação a β_1 e igualar a zero. Como os verdadeiros valores são desconhecidos e apenas conhecemos os valores de uma amostra, ou seja o valor a ser determinado é um estimador do verdadeiro valor populacional, a nova nomenclatura para β_1 será $\hat{\beta}_1$. Com isso temos:

$$0 = -2\sum x_i y_i + 2\hat{\beta}_1 \sum x_i^2$$

Que pode ser reescrita como:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \qquad (6)$$

E o estimador $\hat{\beta}_0$, pode ser calculado a partir de (2):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (7)$$

Sendo que a equação de estimativa será dada por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \qquad (8)$$

$$y_i - y_i = \beta_1(x_i - \bar{x}) \quad (9)$$

Os estimadores apresentam as seguintes propriedades:

- São pontuais;
- A linha de regressão amostral é dada por: $\overline{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$;
- O valor médio do resíduo $\widehat{\varepsilon_i}$ é zero;
- Os resíduos $\widehat{\varepsilon_i}$ são não correlacionados com X_i e $\overline{Y_i}$.

Após a estimativa dos coeficientes da reta de regressão, é necessário verificar se os dados amostrais estão bem descritos pelo modelo encontrado e determinar a parcela da variabilidade amostral que se encontra explicada pela reta de regressão.

3.2.3. Qualidade do ajustamento da reta

- ***Coeficiente de determinação***

Ora: $Y_i = \left(Y_i - \widehat{Y_i}\right) + \left(\widehat{Y_i} - \overline{Y}\right) + \overline{Y}$ (10)

A partir desta equação é possível demonstrar que:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2 + \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2 \qquad (11)$$

O primeiro membro desta equação pode ser interpretado como proporcional à variância total de Y, enquanto o segundo membro reflete a soma de termos proporcionais às suas variâncias residuais e é explicada pelo modelo de regressão. Esta equação (11) pode ser escrita da seguinte forma:

$$SQT = SQ_{Res} + SQ_{Reg} \qquad (12)$$

Onde: SQT é a soma quadrática total
SQ_{Res} é a soma dos quadrados dos resíduos
SQ_{Reg} é a soma dos quadrados devidos à regressão

O coeficiente de determinação é dado pela relação entre a soma dos quadrados devidos à regressão (SQ_{Reg}) e a soma dos quadrados (SQT), ou seja

$$r^2 = \frac{Variância\ Explicada}{Variância\ Total} = \frac{SQ_{Reg}}{SQT} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \qquad (13)$$

Onde: r^2 é o coeficiente de determinação $(0 \leq r^2 \leq 1)$
Y_i é o valor observado da variável dependente
$\hat{Y}_i$ é o valor estimado da variável dependente
$\overline{Y}$ é a média da variável dependente

O coeficiente de determinação é sempre positivo e deve ser interpretado como a proporção da variância total da variável dependente Y que é explicada pelo modelo de regressão e que também pode ser estimado por:

$$r^2 = b^2 \frac{s_X^2}{s_Y^2} \qquad (14)$$

Onde: s_X^2 é a variância amostral de X

s_Y^2 é a variância amostral de Y

b é o coeficiente angular da reta de regressão

O coeficiente de correlação amostral r está relacionado ao coeficiente de determinação r^2 através da seguinte equação: $r = \pm\sqrt{r^2}$, onde o sinal de r é o mesmo do de b. Este coeficiente (r) possui as seguintes propriedades:

- Não depende de qual variável é x e qual é y;
- É independente das unidades de medida;
- Varia entre -1 e 1 (incluindo os extremos);
- Se $r = 1$ indica uma linha reta com coeficiente angular positivo;
- Se $r = -1$ indica uma linha reta com coeficiente angular negativo.

Outra medida simples de calcular é o coeficiente de variância, bastante útil para comparar modelos diferentes e é dado pela fórmula:

$$CV = \frac{\sqrt{QME}}{\bar{Y}} \times 100\% \qquad (15)$$

3.2.4. Pressupostos da Análise de Regressão Linear Simples

Os pressupostos da análise de regressão linear simples (RLS) são a linearidade, a normalidade e a homocedasticidade dos resíduos.

A teoria da regressão assenta nas seguintes suposições sobre os erros:

1. A sua média é zero e a variância desconhecida.
2. São não correlacionados, ou seja, o valor de um erro não depende de qualquer outro erro.
3. Os erros têm distribuição normal.

As verificações das suposições supracitadas são feitas através da análise dos resíduos que, segundo MAROCO (2007), consiste em avaliar os pressupostos de:

- Homogeneidade dos resíduos
- Distribuição normal dos erros
- Independência dos resíduos

- ***Erro padrão da estimativa***

O modelo de regressão linear simples seria ideal se todos os pontos da amostra estivessem sobre a reta ajustada. Porém é difícil tal acontecer e torna-se importante avaliar a medida da variabilidade dos pontos amostrais em relação à reta.

Intrinsecamente ao processo de estimação dos parâmetros da reta de regressão, foi assumida a premissa de que os erros são realizações de uma variável aleatória independente e normalmente distribuída com média zero, ou seja, $E(e_i) = 0$, e variância σ_e^2. Como $E(e_i) = 0$, a variância dos erros ou resíduos e_i será:

$$Var(e_i) = \sigma_e^2 = E\left(e_i^2\right) - E^2(e_i) = E\left(e_i^2\right) \qquad (16)$$

Uma estimativa não enviesada da variância dos resíduos em torno da reta de regressão pode ser obtida por:

$$\hat{\sigma}_e^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \qquad (17)$$

A raiz quadrada da variância dos resíduos e_i é chamada do erro padrão da estimativa, σ_e, e mede a dispersão dos resíduos em torno da reta de regressão.

O erro padrão da estimativa pode ser estimado por:

$$\hat{\sigma}_e = s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} \qquad (18)$$

3.2.5. ANOVA aplicada à RLS

É uma forma de dividir a variância total em componentes, neste caso, devido a regressão e ao resíduo ($y_i = \hat{y}_i + \hat{\varepsilon}_i$). Tem o objetivo de verificar se a parte da variação total explicada pelo modelo (regressão) é significativamente diferente de zero. Na ANOVA as variâncias são denominadas quadrados médios (QM) e obtêm-se pela divisão da soma dos quadrados pelos graus de liberdade. Seja QM o quadrado médio, QM_{REG} os quadrados médios obtidos pela regressão e QM_{RES} os quadrados médios devidos aos resíduos.

A ANOVA pode ser esquematizada no quadro seguinte:

Tabela 4 : Quadro resumo dos cálculos da ANOVA

Fonte de variação	Soma dos Quadrados	g.l.	Quadrados Médios	F
Regressão	$\sum_{i=1}^{n}(\widehat{\boldsymbol{Y}}_i - \overline{\boldsymbol{Y}})^2$	$\boldsymbol{p}$	$\frac{\boldsymbol{SQ\ Reg}}{\boldsymbol{p}}$	$\frac{\boldsymbol{QM\ Reg}}{\boldsymbol{QM\ Res}}$
Resíduos Erros	$\sum_{i=1}^{n}(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}_i)^2$	$\boldsymbol{n-p-1}$	$\frac{\boldsymbol{SQ\ Res}}{(\boldsymbol{n-p-1})}$	
Total	$\sum_{i=1}^{n}(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})^2$	$\boldsymbol{n-1}$		

Utiliza-se o teste F para validar a hipótese $H_0 : \beta_1 = 0$, pois sob H_0 a estatística

$$F = \frac{\boldsymbol{QM\ Reg}}{\boldsymbol{QM\ Res}} \sim F(1, n-2)$$

Portanto, a hipótese nula deve ser rejeitada se o valor calculado for superior ao valor crítico.

3.2.6. Teste de Hipóteses e Intervalos de Confiança para os Coeficientes do MRLS

Devido à variabilidade amostral, a reta de regressão obtida da amostra é uma das retas possíveis. Os valores calculados para b_1 e b_2 são estimativas pontuais dos parâmetros populacionais β_1 e β_2. As retas da população e da amostra são paralelas quando $b_2 = \beta_2$ e terão apenas um ponto necessariamente coincidente, a saber, a média das amostras, quando $b_2 \neq \beta_2$.

Os intervalos de confiança para os coeficientes β_1 e β_2 da reta de regressão são estimados por:

$$b_1 - t_{1-\frac{\alpha}{2},n-2} S_{b_1} \leq \alpha \leq b_1 + t_{1-\frac{\alpha}{2},n-2} S_{b_1}$$

$$b_2 - t_{1-\frac{\alpha}{2},n-2} S_{b_2} \leq \alpha \leq b_2 + t_{1-\frac{\alpha}{2},n-2} S_{b_2}$$

Onde: $t_{1-\frac{\alpha}{2},n-2}$ é o valor do t-Student para $(1-\frac{\alpha}{2})$ e $(n-2)$ graus de liberdade

b_1 e b_2 são estimadores dos parâmetros da reta de regressão

s_{b_1} é o desvio-padrão da estimativa do parâmetro β_1 e indica o quanto está afastado o parâmetro estimado do parâmetro populacional

A equação utilizada para o cálculo de s_{b_1} é dada por:

$$s_{b_1} = \sqrt{s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad (19)$$

s_{b_2} é o desvio-padrão da estimativa de b_2, calculado por:

$$s_{b_2} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (20)$$

No cálculo de s_{b_1} e s_{b_2} tem-se:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \qquad (21)$$

Onde: $e_i = y_i - \hat{y}_1$

n é o tamanho da amostra

$\bar{x}$ é a média da variável independente

x_i é o valor observado da variável independente

A construção do intervalo de confiança para a reta de equação $\beta_1 + \beta_2 x'$ pode basear-se na estimativa de $\hat{y}'$. Considerando um valor x' que não foi utilizado no cálculo dos parâmetros da reta de regressão, demonstra-se que:

$$\mu(\hat{y}') = \beta_1 + \beta_2 x' \qquad (22)$$

donde

$$\hat{\sigma}^2(\hat{y}') = \hat{\sigma}_e^2 \left[\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \qquad (23)$$

O intervalo de confiança para a reta de regressão é dado por:

$$\hat{y}' \pm t_{1-\frac{\alpha}{2}, n-2} s_e \sqrt{\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (24)$$

onde: $\hat{y}' = b_1 + b_2 x'$, $t_{1-\frac{\alpha}{2}, n-2}$ é o valor do t de Student para $\left(1 - \frac{\alpha}{2}\right)$ e $(n-2)$ graus de liberdade.

É de notar que a amplitude do intervalo de confiança será mínima quando x' for igual ao valor médio da amostra utilizada na definição da equação da reta de regressão e será tanto maior quanto mais distante x' estiver da média.

3.3. Modelo de Regressão Linear Múltipla

Muitas aplicações da análise de regressão envolvem situações com mais do que uma variável explicativa. Esse modelo de regressão recebe o nome de modelo de regressão múltipla (MRLM).

Em geral, a variável dependente ou resposta Y pode estar relacionada com k variáveis explicativas ou independentes, ou seja, a variável Y é modelada como função linear de vetores multidimensionais, onde o número de atributos preditores é variável.

O modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ \beta_k x_k + \varepsilon$ recebe o nome de regressão linear múltipla com k variáveis explicativas e os parâmetros $\beta_j, j = 0, \ldots, k$ designam-se por coeficientes de regressão.

Assim se há uma única variável preditora (X_1), a função descreve uma linha reta. Se houver duas variáveis preditoras, então a função descreve um gráfico no plano. Se existem n variáveis preditoras, então a função descreve um hiperplano n-dimensional, como se encontra na figura 2.

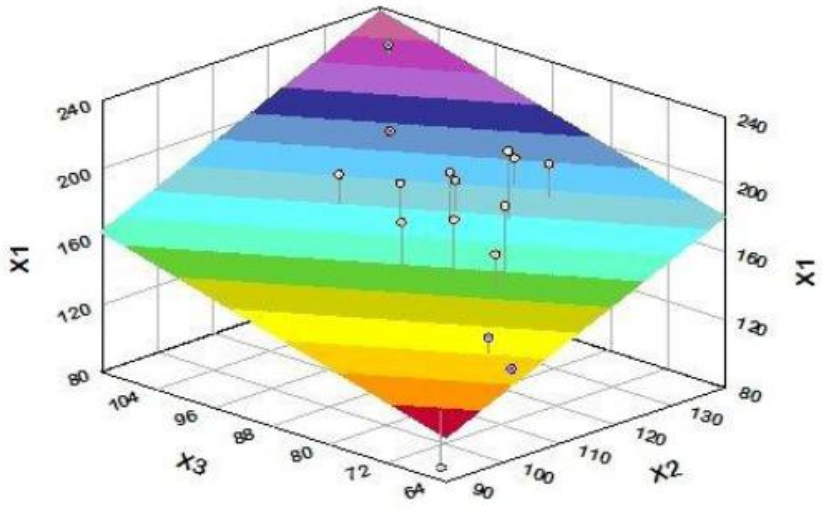


Figura 2: Representação múltipla (gráfico explicativo de uma função preditora com três variáveis)

Pressupostos sobre os erros do modelo de regressão linear múltipla:

1. Têm média zero e a mesma variância desconhecida.
2. São não correlacionados, ou seja, o valor de um erro não depende de qualquer outro erro.
3. Os erros têm distribuição normal.

As verificações das suposições supracitadas são feitas através da Análise Residual.

Em algumas situações, mais do que uma variável independente $(X_1, X_2, \dots, X_n)$ pode ser necessária para predizer o valor da variável dependente (Y). O modelo matemático para esse caso é descrito abaixo:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_{2i} + \ \dots + \beta_k X_{ki} + \varepsilon_i \qquad (25)$$

Para as *n* observações poderá ser escrito da forma:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_{21} + \ \dots + \beta_k X_{k1} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \beta_2 X_{22} + \ \dots + \beta_k X_{k2} + \varepsilon_2$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$Y_n = \beta_0 + \beta_1 X_n + \beta_2 X_{2n} + \ \dots + \beta_k X_{kn} + \varepsilon_n$$

Que na realidade é um sistema linear, que podemos escrever na forma de matriz:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 & X_{21} & X_{k1} \\ 1 & X_2 & X_{22} & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_n & X_{2n} & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_k \end{bmatrix}$$

Escrevendo ainda em outra forma mais compacta temos:

$$Y = \beta X + \varepsilon \qquad (26)$$

O estimador para β será dado por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (27)$$

Pela equação (27), há necessidade que o produto $X^T X$ tenha uma matriz inversa, o que implica a condição obrigatória que nenhuma coluna da matriz X seja combinação linear das outras.

3.3.1. Análise de Variância (ANOVA) Aplicada à Regressão Linear Múltipla

O modelo de regressão linear múltipla representa-se por:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots \ \beta_k x_k + \varepsilon \qquad (28)$$

com uma variável dependente e *k* variáveis independentes.

Segundo MAROCO (2007), após explicarmos a variabilidade total do modelo (SQT) como a soma da variabilidade explicada pelo modelo ($SQReg$) com a variabilidade não explicada pelo modelo (mas sim pelos erros) ($SQ\ Res$), vamos avaliar, a partir de estimativas amostrais, se

na população algumas das variáveis independentes (VI) podem ou não influenciar a variável dependente (VD), ou seja, se o modelo ajustado é ou não significativo.

A hipótese teórica é avaliada pelo teste que se refere de seguida.

3.3.2. Teste de significância da equação de Regressão Linear Múltipla

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: $\beta_1 = \beta_2 = \ ... \ = \beta_n = 0$ (a relação entre as variáveis é não linear)

H_1: $\exists\, \beta_i \neq 0$

Este teste é conhecido como teste do F total. A estatística do teste é a relação entre a variância decorrente da regressão linear múltipla e a variância dos resíduos:

$$F = \frac{QM\ Reg}{QMRes} \qquad (29)$$

A hipótese nula será aceite se:

$$F < F\ (\alpha, p, n - p - 1),$$

Onde: α é o nível de significância
p o número de variáveis independente
p e $n - p - 1$ são os graus de liberdade da distribuição F de Snedecor

3.3.3. Teste de Partes de um Modelo de Regressão Linear Múltipla

A contribuição de uma variável explicativa ao modelo de regressão múltipla pode ser determinada pelo critério do teste do F parcial. De acordo com este critério, avalia-se a contribuição de uma variável explicativa para a soma dos quadrados devido à regressão, após a inclusão no modelo das restantes variáveis independentes.

A verificação se a inclusão de uma variável X_k melhora significativamente o modelo de regressão é realizada por meio do seguinte teste de hipóteses:

H_0: a variável X_k não melhora significativamente o modelo
H_1: a variável X_k melhora significativamente o modelo

A estatística do teste é dada por:

$$F_p = \frac{\mathrm{SQ\ Reg(X_k)}}{\mathrm{QMRes}} \ (30)$$

A hipótese nula deve ser rejeitada se a estatística F_p for maior ou igual ao valor critico da distribuição F de Snedecor,

Onde: n é tamanho da amostra
α é o nível de significância
p o número de variáveis explicativas incluindo X_k
p e $n-p-1$ são os graus de liberdade da distribuição F de Snedecor

ou seja, rejeita-se H_0 se $Fp \geq F(\alpha, 1, n-\ p\ -1)$.

3.3.4. Coeficiente de determinação parcial

O coeficiente de determinação múltipla, R^2, avalia a proporção da variância da variável dependente Y que é explicada pelas variáveis independentes, X_i. Para um modelo de regressão múltipla com p variáveis explicativas, o coeficiente de determinação parcial para a $k-ésima$ variável é dado por:

$$R^2_{nk(p-k)} = \frac{SQ\ Reg(X_k)}{SQT - SQReg + SQRes(X_k)} \qquad (31)$$

3.3.5. Inferência sobre os coeficientes de determinação parcial

Um teste de hipótese utilizado para verificar se $\beta_i = \beta_0$, onde β_0 é um valor constante conhecido, pode ser implementado com as seguintes hipóteses nula e alternativa:

$$H_0: \beta_i = \beta_0$$

$$H_1: \beta_i \neq \beta_0$$

Para tais hipóteses, a estatística do teste é calculada pela relação: $t = \frac{\hat{\beta}_1 - \beta_0}{s_{\hat{\beta}_1}}$

A hipótese nula é rejeitada se $|t| > t_{1-\frac{\alpha}{2}, n-p-1}$,

Onde: α é o nível de significância (teste bilateral)
n é o tamanho da amostra
p é o número de variáveis independentes do modelo

Os intervalos de confiança para os coeficientes da regressão β_i são dados por:

$$\hat{\beta}_i \pm t_{1-\frac{\alpha}{2}, n-P-1} s_{\hat{\beta}i} \qquad (32)$$

3.3.6. Intervalos de Confiança da Regressão Linear Múltipla

Os limites de confiança de um valor individual previsto $\hat{Y}_k$ são estimados por:

$$[X_h]\left[\hat{\beta}\right] \pm \ t_{1-\frac{\alpha}{2},n-p-1}\sqrt{var_i\ (\ \widehat{Y_k}\ )} \qquad (33)$$

Onde: $var_i\ (\ \hat{Y}_k)$ é a variância de um valor individual previsto de Y

3.3.7. Avaliação da Regressão Linear Múltipla

Ao ajustar uma equação de regressão aos dados, na maioria das vezes o valor observado de y não corresponde exatamente ao valor predito de y. A esta diferença chamamos de resíduos ou variação residual.

A variância estimada para a $\hat{y}$ da variável independente y é dada por:

$$\hat{\sigma}\ _{\hat{\beta}}^{\ 2} = \frac{SQ\ Reg}{n-(\ k+1)} \qquad (34)$$

O coeficiente de determinação r^2 deve ser ajustado para regressão múltipla por:

$$r^2_{ajustado} = \frac{(n-1)r^2-k}{n-(k+1)} \qquad (35)$$

Quando pretendemos comparar diversos modelos com diferentes números de variáveis independentes, usamos o $r^2_{ajustado}$, e não o r². O $r^2_{ajustado}$ pondera o r^2 de acordo com o número de variáveis independentes no modelo, e o número de observações, com o intuito de estimar a redução da validade da equação de predição.

3.4. Regressão Logística

Em muitas situações práticas, no decurso da investigação de fenómenos reais, o investigador necessita de recorrer a um modelo matemático representativo, que pode ser definido como uma abstração dum sistema real que possa ser utilizada com os propósitos de predição e controle e para aplicável deve ter dois atributos, o realismo e a simplicidade (MARTINS (1988)). Se por um lado o modelo deve servir como uma aproximação razoavelmente precisa do sistema real e conter a maior parte dos aspetos importantes do mesmo, por outro não deve ser tão complexo que se torne impossível compreendê-lo e manipulá-lo.

Nas situações multifatoriais reais deparamo-nos com fatores que atuam efetivamente sobre a variável resposta influenciando-a, enquanto outros não, agindo apenas como fatores de confusão. Assim, com o objetivo de se interpretar corretamente os fenómenos, devemos utilizar modelos que considerem a ação conjunta de variáveis. Para modelar estes

fenómenos, que envolvem uma variável dependente categórica (nominal) e várias variáveis independentes métricas ou categóricas, necessitamos de selecionar um método estatístico apropriado, que nestas situações são a análise discriminante e a regressão logística.

As variáveis dependentes categóricas, como por exemplo: qualidade de vida (QOL, *Quality of Life*), indicadores da condição de saúde, gravidade da doença, etc. utilizadas em estudos epidemiológicos podem ser ordenadas na forma de score (k valores). Se as variáveis dependentes são discretas, é inadequado inclui-las no modelo como se fossem variáveis escalares, devendo-se utilizar variáveis de *design* (ou *dummy*), ou seja se uma variável discreta com k valores possíveis, então representaremos cada um deles por uma variável *dummys*, obtendo um modelo com $k-1$ variáveis *dummy*. Estes modelos, dependendo do delineamento do estudo, permitem também calcular a estatística *odds ratio* (*or*) ou a probabilidade de ocorrência de um evento A ($P(A)$).

Seja *Y* uma variável aleatória *dummy* definida como; $Y_i = \{1, 0\}$, onde cada Y_i tem distribuição de Bernoulli, cuja função de distribuição de probabilidade é dada por;

$P(y|p) = p^y(1-p)^{1-y}$

onde: y identifica o evento ocorrido

p é a probabilidade de sucesso para a ocorrência do evento

Como se trata de uma sequência de eventos com distribuição de Bernoulli, a soma do número de sucessos ou fracassos nesta experiência terá distribuição Binomial de n parâmetros (número de observações) e p (probabilidade de sucesso). A função de distribuição de probabilidade da Binomial é dada por;

$$P(y|n,p) = \binom{n}{y} p^y(1-p)^{1-y}$$

A transformação logística pode ser interpretada como sendo o logaritmo da razão de probabilidades, sucesso *versus* fracasso, daí a regressão logística nos dar uma ideia do risco de obter sucesso, dado o efeito das variáveis explicativas (que serão introduzidas mais adiante).

A função de ligação deste modelo linear generalizado é dada pela seguinte equação:

$$\eta_i = log\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{n} \beta_k \, x_{ik}$$

onde a probabilidade p_i é dada por:

$$p_i = \frac{exp(\sum_{k=0}^{n} \beta_\kappa \, x_{ik})}{1 + exp(\sum_{k=0}^{n} \beta_\kappa \, x_{ik})}$$

A função usada na regressão logística para estimar a probabilidade de uma determinada realização j $(j = 1, \dots, n)$ da variável independente ser "sucesso"

$$\hat{\pi}_j = \frac{exp(\sum_{k=0}^{n} \beta_\kappa \, x_{ik})}{1 + exp(\sum_{k=0}^{n} \beta_\kappa \, x_{ik})}$$

Onde $\hat{\pi}$ é o vetor das probabilidades estimadas e β é o vetor dos coeficientes de regressão logística. Este modelo pode ser ajustado recorrendo à regressão não linear, em que a solução consiste em linearizar a função através da transformação $Logit\ (\hat{\pi})$

$$Logit\ (\hat{\pi}) = ln\left(\frac{\hat{\pi}_l}{1 - \hat{\pi}_l}\right) = \sum_{k=0}^{n} \beta_\kappa \, x_{ik}$$

Um modelo de regressão deve obedecer aos seguintes pressupostos:

- Linearidade e aditividade: a escala de $Logit\ (\pi)$ é aditiva e linear (mas a de π não).

- Proporcionalidade: a contribuição para cada X_i $(i = 1, \dots, k)$ é proporcional ao seu valor com um fator β_i .

- Constância de efeito: a contribuição de uma variável independente é constante, e independente da contribuição das outras variáveis independentes.

- Os erros são independentes e apresentam distribuição binomial.

- Os preditores não são multicolineares

3.4.1.Estimação de parâmetros em regressão logística

O método de ajustamento mais utilizado para estimar os parâmetros de um modelo de regressão logística é o método da Máxima Verosimilhança. Este método estima os coeficientes de regressão que maximizam a probabilidade de encontrar as realizações da variável dependente da amostra. Como a variável tem uma distribuição de Bernoulli,

$$f(y_i \, , \pi_{\,i}) = P(Y = y_j) = \pi_i^{yi}(1 - \pi_i)^{1 - y_i}$$

As observações são independentes. Logo, a função distribuição de probabilidade conjunta de $y_1, y_2, \dots, y_n$ será:

$$\prod_{i=1}^{n} f(\, y_i \,, \pi_i) = \prod_{i=1}^{n} \pi_i^{yi} (1 - \pi_i)^{1-y_i}, \text{ com } y_i = \{0,1\}.$$

Então a função de verosimilhança será dada por: $L(\beta) = \prod_{i=1}^{n} \pi_i^{yi} (1 - \pi_i)^{1-y_i}$

O princípio da máxima verosimilhança consiste em estimar o valor de β que maximiza a função de verosimilhança. A aplicação do logaritmo natural ajuda no processo de manipulação algébrica.

$$l(\beta) = ln\,[L\,(\beta)] = ln\left[\prod_{i=1}^{n} \pi_i^{yi} (1 - \pi_i)^{1-y_i}\right]$$

donde podemos obter $l(\beta) = \sum_{i=0}^{n} [y_i\ (\beta_0 + \beta_i x_i\ ) - ln(1 + exp(\beta_0 + \beta_i x_i\ ))]$ (#)

O valor de β que maximiza $l(\beta)$ é encontrado após derivar-se em relação aos parâmetros (β_0, β_1).

Igualando as equações das funções derivadas, em relação aos parâmetros a zero obtemos $\sum_{i=1}^{n} (y_i - \pi_i) = 0$ e $\sum_{i=1}^{n} x_i (y_i - \pi_i) = 0$ que são equações não lineares nos parâmetros e requerem o emprego de processo interativo na sua solução.

Odds ratio

O *odds* é a medida de ocorrência que substitui a proporção quando esta não é aplicável. O *odds ratio* (*or*), é uma medida de efeito que compara a probabilidade de determinada condição ocorrer entre grupos. Dados dois grupos distintos pela presença/ausência de uma determinada caraterística A (A_1 e A_0), a ocorrência de uma determinada condição comparada pela razão das probabilidades em A_1 contra as probabilidades em A_0 e designando por p a presença da condição e q a ausência, o *or* é dado por

$$or = \frac{p_1/q_1}{p_0/q_0} \quad \text{ou} \quad or = \frac{p_1\, q_0}{p_0\, q_1}$$

Tratando-se de uma medida contínua independente segue um modelo logístico, onde

$$or = e^{\beta_0 + \beta_i x_i}$$

Onde: β_0 é uma constante

β_i coeficiente da i^{ésima} variável independente

x_i *valor da* i^{ésima} variável independente

x indica mudança de um atributo

Assim para uma variação de atributo Δx temos:

$$or = \frac{e^{\beta_0+\beta_i(x_i+\Delta x)}}{e^{\beta_0+\beta_i(x_i)}} = e^{\beta_i \Delta x}$$

Ou seja, o *or* é a exponencial do produto do coeficiente de regressão pela variação da variável independente.

Na área da saúde (campo do nosso estudo) o ***or***, como definida anteriormente permite-nos relacionar a ocorrência de um evento entre indivíduos expostos (A_1) e não expostos (A_0) a determinado fator de risco.

O intervalo de confiança para o ***or*** de $100(1-\alpha)\%$ é obtido calculando o intervalo de confiança de β_1 e aplicando a exponencial.

Tem-se: $exp\left[\hat{\beta}_1 \pm z_1 - \frac{\alpha}{2} SE(\hat{\beta}_1)\right]$, onde $SE(\hat{\beta}_1)$ é o erro padrão de $\hat{\beta}_1$.

- ***Inferência***

Após estimar os coeficientes de regressão, a significância da variável é o primeiro aspeto a observar antes de progredir com a análise. Tal envolve testes de hipóteses para saber se a variável é ou não significativamente correlacionada com a saída. Na regressão logística, a comparação dos valores observados com os preditos é baseada na função logaritmo da verosimilhança, apresentada na equação (#).

Segundo Hosmer e Lemeshow (1989), a estatística D é chamada *deviance*, e desempenha um papel fundamental em algumas aproximações para verificar o bom ajuste.

Seja $D = -2ln$ (Verosimilhança do modelo ajustado)

A comparação da estatística deviance do modelo com e sem variável conduz-nos a $D = -2\left[\frac{verosimilhança\ sem\ variável}{verosimilhança\ com\ a\ variável}\right]$

No caso da regressão logística simples, a verosimilhança do modelo pode ser testada se a inclusão de uma variável independente x melhoraria o ajuste do modelo sem a variável, ou seja, se o modelo apenas com a interseção β_0 descreveria melhor o comportamento dos dados observados. Isso pode ser encarado como fazer $\beta_1 = 0$, a estatística G segue uma distribuição qui-quadrado com um grau de liberdade.

E segundo Maroco (2007) a estatística do teste G^2 para testar a significância do modelo é dada por:

$$G^2 = X_0^2 - X_C^2 = -2LL_0 - (-2LL_C) = -2Ln\left[\frac{L_0}{L_C}\right] \quad \underset{\sim}{a} \chi^2_{(p)}$$

Onde: X_0^2 é o modelo nulo ou reduzido

X_C^2 é o modelo completo

Assim o valor de G^2, obtém-se a partir do rácio das verosimilhanças de dois modelos e é uma medida de incremento da qualidade do modelo nulo por adição das variáveis independentes. Logo o modelo completo é estatisticamente significativo apenas quando a adição de uma ou mais variáveis independentes ao modelo, reduz significativamente o valor de -2LL.

È de realçar que concluir que o modelo completo é significativo, permite apenas afirmar que pelo menos uma variável independente incluída no modelo influencia significativamente a variável dependente como ajustado pelo modelo.

Também podemos recorrer à formulação de um teste de hipóteses que permita afirmar se uma variável é ou não significante no modelo de regressão, além de permitir calcular o p-valor de tal variável.

O teste de Wald, compara o valor de β_1 obtido da estimação de máxima verosimilhança e o seu erro padrão $\widehat{\mathrm{SE}}\,(\hat{\beta}_1)$.

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}\,(\hat{\beta}_1)}$$

Sob a hipótese de que $\beta_1 = 0$, W segue a distribuição normal padrão.

Já o teste de Score tem como principal vantagem o uso de pequeno esforço computacional no seu cálculo. Este teste é baseado na teoria da distribuição das derivadas do log da máxima verosimilhança.

O teste de Score é dado por :

$$ST = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_{i(y_i - \bar{y})}}}{\sqrt{\bar{\mathrm{y}}(1-\bar{\mathrm{y}})\ \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})^2}}$$

Sob a hipótese de que $\beta_1 = 0$, a estatística Score tem distribuição normal padrão.

As estimativas do modelo de regressão logística são estimativas de máxima verosimilhança que se determinam através de um processo iterativo. Elas não são calculadas para minimizar a variância, logo as técnicas utilizadas para avaliar a qualidade do ajuste não se aplicam.

Para avaliar a qualidade do ajuste de modelos logísticos, foram desenvolvidos vários pseudo-r^2.

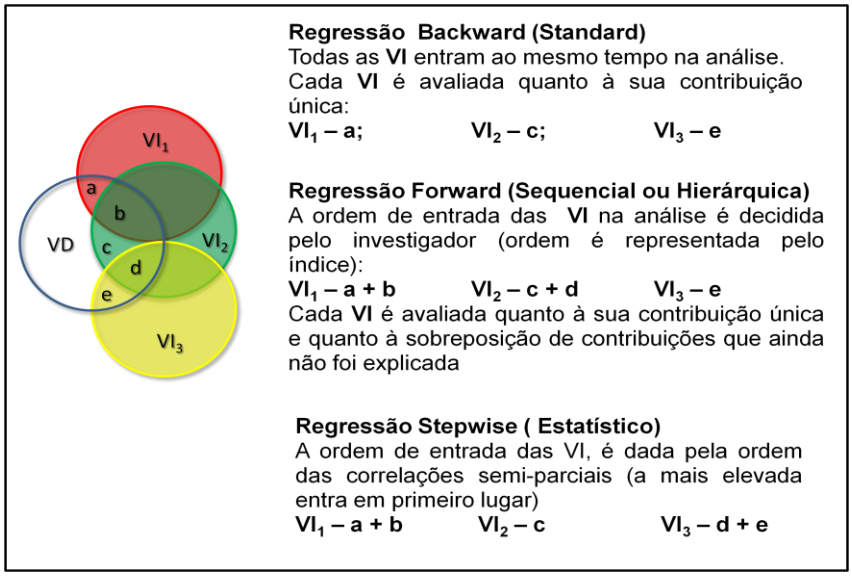
Quadro síntese dos pseudo-r^2		
Pseudo-R^2	Fórmula	Descrição
Pseudo-R^2 de McFadden R^2_{MF}	$R^2_{MF} = 1 - \frac{\ln \widehat{L}\,(M_{Completo})}{\ln \widehat{L}\,(M_{Interc})}$ $M_{completo}$ = Modelo com preditores $M_{interc.}$= Modelo sem preditores $\hat{L} = Estimador\ "likelihood"$	A verosimilhança do modelo de interceção é tratada como uma soma total de quadrados, e a verosimilhança do modelo completo é a soma dos erros quadrados. A razão das probabilidades sugere o nível de melhoria em relação ao modelo de interceção oferecida pelo modelo completo. A probabilidade de se situar entre 0 e 1, de modo que o log da probabilidade é menor do que ou igual a zero. Se um modelo tem uma probabilidade muito baixa, então o log da probabilidade terá uma magnitude maior do que o log de um modelo mais provável. Assim, uma pequena proporção das probabilidades de log indica que o modelo é um ajuste completo muito melhor do que o modelo de interseção. Se quisermos comparar dois modelos com os mesmos dados, o R^2_{CS} será maior para o modelo com a maior probabilidade.
Pseudo-R^2 de Cox& Snell R^2_{CS}	$R^2_{CS} = 1 - \left\{\frac{L(M_{Interc})}{L(M_{Completo})}\right\}^{2/N}$	O rácio das probabilidades reflete a melhoria do modelo completo sobre o modelo de interceção (a uma menor proporção corresponde uma melhoria). Definindo L (M) como a probabilidade condicional da variável dependente dadas as variáveis independentes. Se houver N observações no conjunto de dados, então L(M) é o produto de N tais probabilidades. Assim, tomando a raiz índice n do produto L (M) obtemos uma estimativa da probabilidade de cada valor de Y. R^2_{CS} obtém-se como uma transformação da -*2ln[L($M_{Interc.}$)/L($M_{Completo}$)]* da estatística que é utilizada para determinar a convergência de uma regressão logística. Note-se que R^2_{CS} tem um valor máximo que não é 1: se o modelo completo, prediz perfeitamente e tem uma probabilidade de 1, R^2_{CS} é então 1-*L($M_{Interc.}$)$^{2/N}$*, que é sempre menor que um.
Pseudo-R^2 de Nagelkerke R^2_N	$R^2_N = \frac{1 - \left\{\frac{L(M_{Interc})}{L(M_{Completo})}\right\}^{2/N}}{1 - L(M_{Interc})^{2/N}}$	Este pseudo-r^2 ajusta o R^2_{CS} de modo que varie entre 0 e 1. Para alcançar este objetivo, o R^2_{CS} é dividido pelo seu valor máximo, 1-L (M_{nterc}) 2 / N. Então, se o modelo completo prediz perfeitamente e tem uma probabilidade de 1, R^2_N = 1. Se *L($M_{Completo}$) = 1*, então R^2_N = *1*; Se *L($M_{Completo}$) = L(M_{interc})*, então R^2_N = *0*.

3.4.2. Método de seleção baseado no critério de informação

A abordagem tradicional na construção de modelos estatísticos é encontrar o modelo mais parcimonioso que explica os dados. Quantas mais variáveis no modelo, maior se torna a estimativa do erro e mais dependente o modelo fica dos dados observados.

Existem algumas técnicas para auxiliar na seleção de variáveis para um modelo de Regressão Logística, assim o critério para a adição ou remoção de covariáveis é geralmente baseado na estatística F, comparando modelos com e sem as variáveis em questão. Existem três procedimentos automáticos: o Método Forward, o Método Backward e o Método Stepwise.

Estes métodos distiguem-se pelo que aontece à variabilidade devida ao efeito comum das VI (quando estão correlacionadas entre si) e pelos critérios da ordem de entrada das VI na equação. Esquematicamente podemos visualizar estes três métodos no esquema seguinte:



Adaptado de Tabachnick & Fidell (2007)

Qualquer procedimento para seleção ou exclusão de variáveis de um modelo é baseado num algoritmo que testa a importância das variáveis, incluindo ou excluindo-as do modelo baseando-se numa regra de decisão. A importância da variável é definida em termos de uma medida de significância estatística do coeficiente associado à variável para o modelo. Essa estatística depende das suposições do modelo.

No nosso exemplo de aplicação vamos recorrer ao Método de Seleção Stepwise. Neste método, recorre-se ao teste F que é utilizado desde que os erros tenham distribuição normal.

Na regressão logística os erros seguem distribuição binomial e a significância é assegurada através do Teste da Razão de Verosimilhança.

Assim, em cada passo do procedimento a variável mais importante, em termos estatísticos, é aquela que produz a maior mudança no logaritmo da verosimilhança em relação ao modelo que não contém a variável.

4. Técnicas de visualização de informação

4.1. Fundamentos da visualização gráfica

Quando um gráfico é elaborado, um dos elementos mais importantes a ter em conta é a sua perceção, porque permite dar uma fundamentação científica à sua construção e sustentar a escolha de uma forma em detrimento de outra. Na fase da construção, a informação é codificada no gráfico através de símbolos, comprimentos, declives dos segmentos de reta, áreas, textura ou cor. Quando um gráfico é analisado, a informação nele contida é descodificada pelo analista, sendo o processo de descodificação denominado de perceção gráfica, que permite avaliar a capacidade de um gráfico transmitir informação (CLEVELAND, MCGILL, 1987). A extração de informação a partir dos gráficos envolve tarefas percetivas realizadas pelo sistema visual olho-cérebro. No quadro seguinte, estas tarefas estão ordenadas segundo a precisão na extração de informação quantitativa. Quanto menos precisa for a perceção, maior o erro de leitura, ou seja, maior a diferença entre o valor percebido e o valor correto.

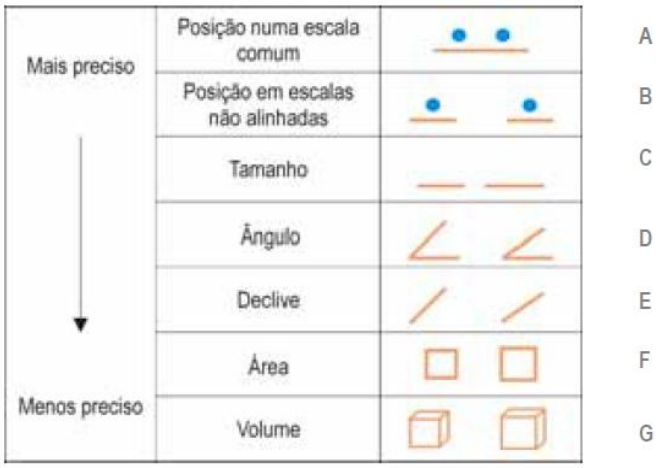


Figura 3: Avaliação de tarefas percetivas ordenadas segundo a sua precisão (adaptado de CLEVELAND, MCGILL,1987)

BERTIN (1973) foi o primeiro a sistematizar os conhecimentos sobre a representação gráfica de informação, criando uma tipologia com as seguintes variáveis visuais:

Localização – com utilização dum referencial cartesiano que atribui a um ponto determinadas coordenadas;

Tamanho – atribuição dum tamanho ao símbolo que evidencie a importância numérica da informação que ele representa (variação em comprimento, largura, área, etc.);

Valor – refere-se à variação percebida (contraste) claro/escuro da cor (ex.: preto-branco);

Textura – tamanho e espaçamento dos elementos gráficos que constituem o símbolo (pontos, linhas ou outros), expresso pelo número desses elementos que se repetem por unidade de comprimento;

Cor – sensação pela qual se diferencia entre porções particulares do espetro eletromagnético, isto é, azul, verde, vermelho, etc.;

Orientação – também designada por direção, corresponde ao ângulo do símbolo com a linha de leitura (referencial);

Forma – pode ser geométrica (como quadrados ou círculos) ou irregular.

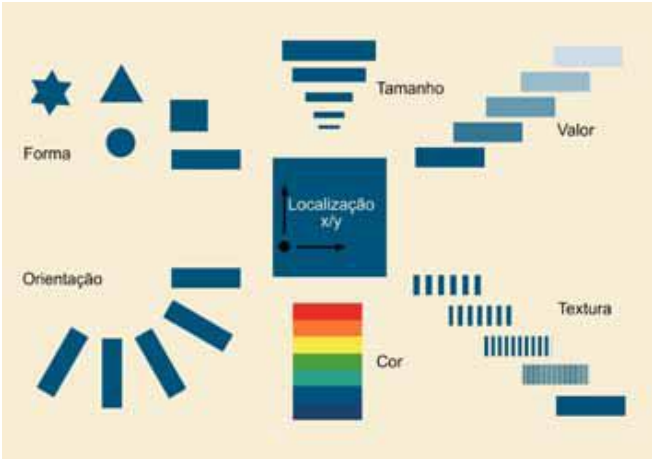


Figura 4: As variáveis visuais segundo Bertin

A representação gráfica é um conceito simples, porém poderoso, e tem causado grande impacto em diversas áreas, tais como, medicina, engenharia e ciências, por facilitar a interpretação da informação tornando-a por vezes mais acessível a indivíduos com menos conhecimentos de estatística. Porém, o seu uso deve ser sempre ponderado em função do destinatário e da mensagem a transmitir.

Citando DINIS PESTANA, “na minha investigação uso os gráficos com uma parcimónia que ronda a avareza, e creio que apenas publiquei gráficos em trabalhos de índole didática. Gráficos de construção simples podem ser inspiradores, mas de modo nenhum substituem uma análise assente em métodos decerto menos apelativos, mas mais seguros. Os gráficos são mais próprios da análise exploratória de dados do que da análise estatística confirmatória, que decerto merece um estatuto de maior relevo.”

4.2. Tipos de gráficos aplicados neste estudo

- ***Regressão Linear***

Diagrama de Dispersão é um gráfico que usaremos para observar o comportamento conjunto de duas variáveis e avaliar da existência de alguma relação entre elas. Cada ponto do gráfico representa um elemento da população e as respetivas coordenadas os valores das duas variáveis.

A este gráfico ajustamos uma reta de equação: Y = $\beta_0 + \beta_1$ X, denominada reta de regressão, que é a que melhor se ajusta aos pontos do diagrama de dispersão, onde: β_1 representa o coeficiente angular e β_0 o coeficiente linear. Esta reta será a média procurada para o Intervalo de Confiança que contém Y e os valores de β_0 e de β_1 são determinados de forma a minimizar os resíduos ou **erros** (ε_i) encontrados utilizando o **Método dos Mínimos Quadrados Ordinários**.

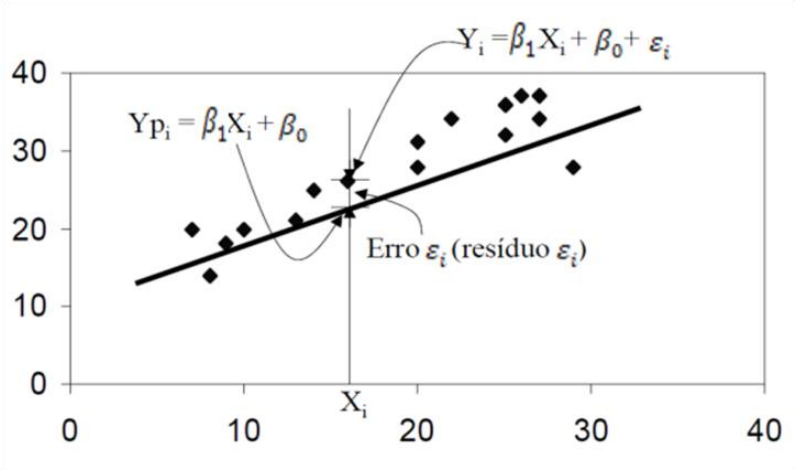


Figura 5: Digrama de dispersão com reta de regressão

O diagrama de dispersão permite visualizar o grau de associação entre as variáveis e a tendência de variação em conjunto. A figura 6, apresenta alguns exemplos de variação conjunta entre duas variáveis.

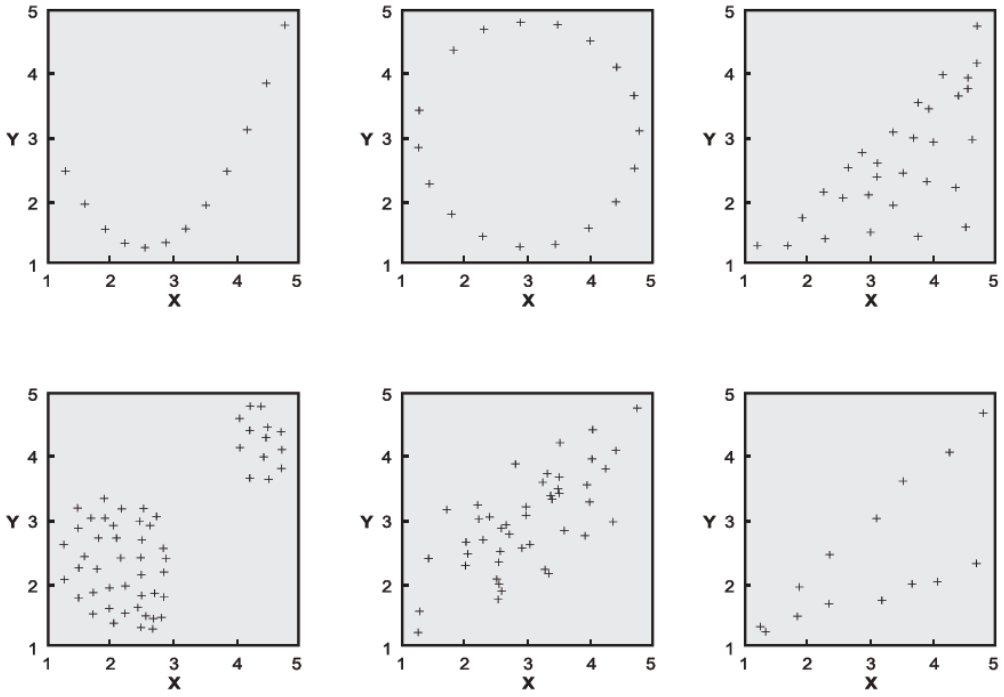


Figura 6: Exemplos de relação conjunta entre variáveis (adaptado de HELSEL e HIRSH, 1992)

A independência de resíduos pode ser verificada com gráficos em relação à variável prevista, Y.

A figura seguinte ilustra duas situações: uma onde se verifica a independência dos resíduos e a outra onde se observa a ocorrência de dependência.

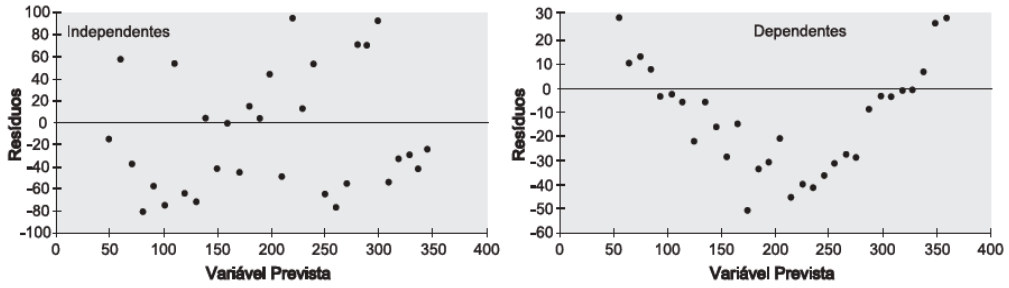


Figura 7: Verificação de independência.

Também a hipótese de variância constante no MRLS pode ser verificado por meio de análise gráfica entre os resíduos e a variável dependente X, como se ilustra na figura seguinte, que apresenta situações em que existe a violação de variância constante.

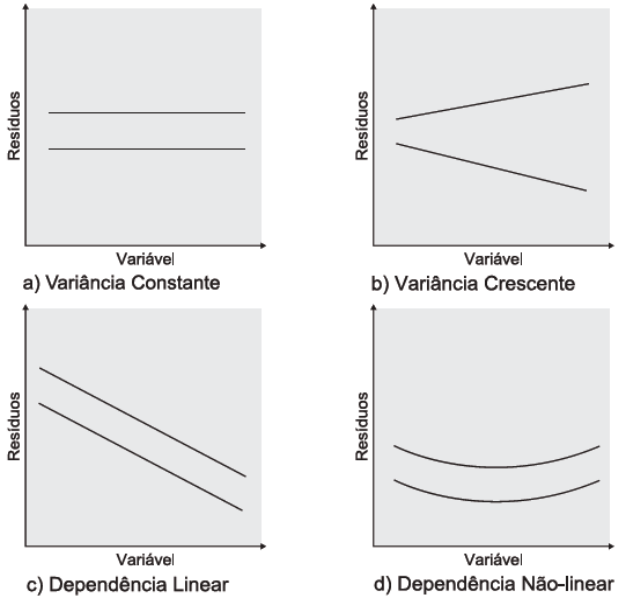


Figura 8: Verificação da variância dos resíduos.

A medida da variação conjunta das variáveis ou covariação observada num diagrama de dispersão é a correlação entre as duas variáveis. Essa medida é realizada numericamente por meio dos coeficientes de correlação que representam o grau de associação entre duas variáveis contínuas. As medidas de correlação, frequentemente designadas por ρ, são adimensionadas e variam entre -1 e 1. No caso de ρ = 0, não existe correlação entre as duas variáveis. Quando ρ > 0, a correlação é positiva e uma variável aumenta quando a outra cresce. A correlação é negativa, ρ<0, quando as variáveis variam em direções opostas.

A correlação é chamada de monotónica se uma das variáveis aumenta ou diminui sistematicamente quando a outra decresce, com associações que podem ter forma linear ou não linear. A figura 9 apresenta exemplos de correlações monotónicas não lineares e não monotónicas.

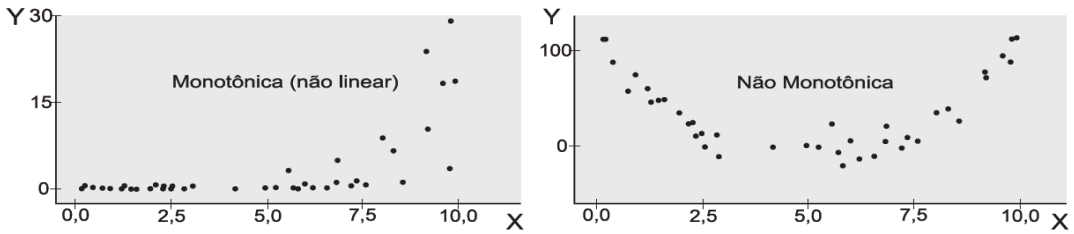


Figura 9: Exemplos de correlações (adaptado de HELSEL e HIRSH, 1992)

É importante salientar que variáveis altamente correlacionadas não apresentam necessariamente qualquer relação de causa e efeito. A correlação representa simplesmente a tendência que as variáveis apresentam quanto à sua variação conjunta. Assim, a medida de correlação não indica necessariamente que há evidências de relações causais entre duas variáveis. As evidências de relações causais devem ser obtidas a partir do conhecimento dos processos envolvidos.

- ***Coeficiente de Correlação linear de Pearson***

Duas variáveis apresentam uma correlação linear quando os pontos do diagrama de dispersão se aproximam de uma reta. Essa correlação pode ser positiva (para valores crescentes de X há uma tendência a valores também crescentes de Y) ou negativa (para valores crescentes de X a tendência é observarem-se valores decrescentes de Y). A figura seguinte ilustra correlações lineares positivas e negativas.

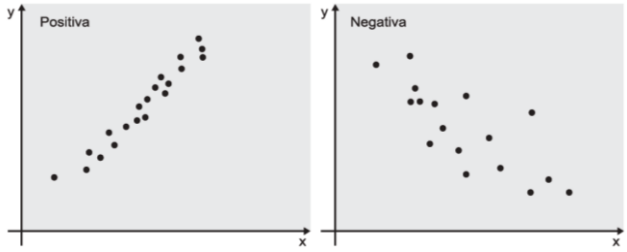


Figura 10: Correlações lineares positivas e negativas

O coeficiente de correlação linear de Pearson é adimensional e varia entre -1 e +1, o que não ocorre com a covariância. Assim, as unidades adotadas pelas variáveis não afetam o valor do coeficiente de correlação. Caso os dados se alinhem perfeitamente ao longo da reta com declive positivo teremos a correlação linear positiva perfeita com o coeficiente de Pearson igual a 1. A correlação linear negativa perfeita ocorre quando os dados se alinham perfeitamente ao longo de uma reta com declive negativo e o coeficiente de correlação de Pearson é igual a -1. A figura 11, apresenta alguns diagramas de dispersão com os respetivos valores do coeficiente de correlação.

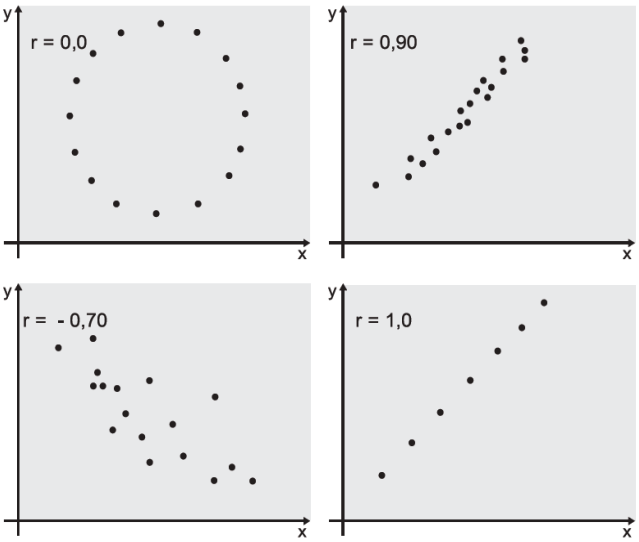


Figura 11: Exemplos de coeficientes de correlação

É de realçar que a um valor do coeficiente de correlação alto, embora estatisticamente significativo, não corresponde necessariamente uma relação de causa e efeito, mas simplesmente indica a tendência que aquelas variáveis apresentam quanto à sua variação conjunta.

Outro cuidado que se deve ter na análise de duas variáveis é com a ocorrência de correlações aparentes (em que as variáveis não estão correlacionadas). As causas mais frequentes desta ocorrência são: a distribuição não equilibrada dos dados (figura 12); a relação entre quocientes de variáveis que apresentam o mesmo denominador (figura 13); e a relação de variáveis que foram multiplicadas por uma delas (figura 14).

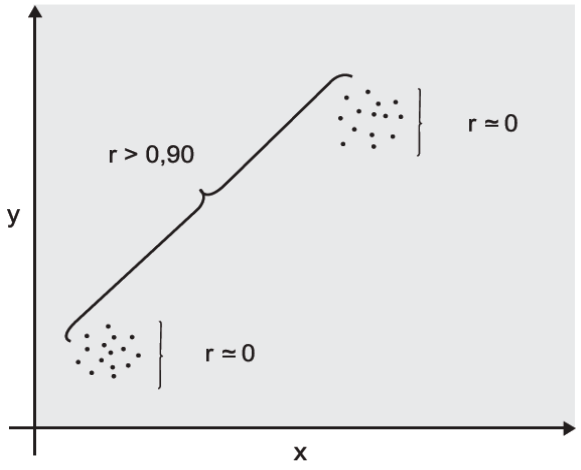


Figura 12: Distribuição não equilibrada de dados

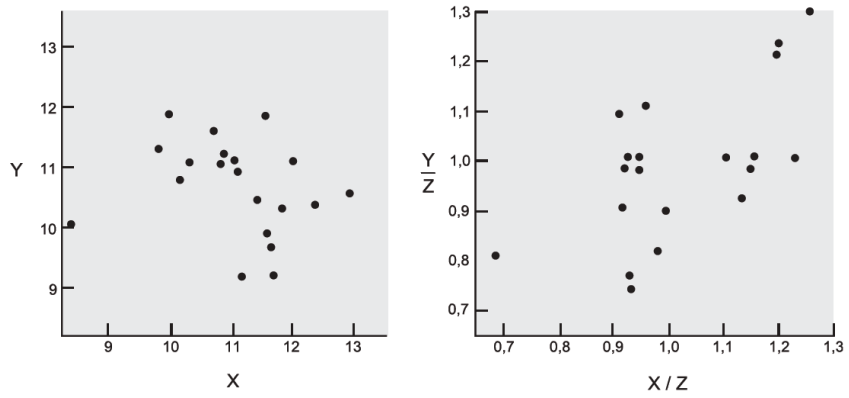


Figura 13: Correlação entre quocientes de variáveis

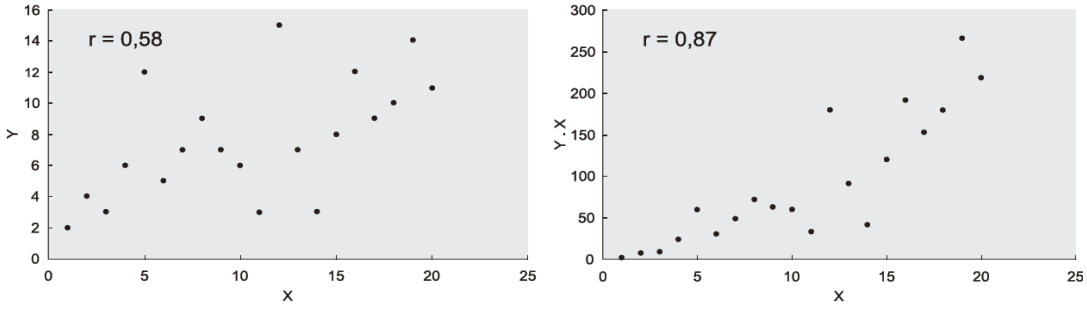


Figura 14: Correlação entre produto de variáveis

- ***Gráficos P-P plot e Q-Q plot para avaliação do ajuste do modelo***

O pressuposto de normalidade dos resíduos pode ser testado recorrendo a testes de ajustamento tais como o Teste Kolmogorov-Smirnov ou o Teste da Normalidade de Lilliefors, que já foram abordados anteriormente, porém esta condição também pode ser verificada usando um gráfico de probabilidade normal *(normal probability plot).*

O primeiro passo que deverá ser dado para verificar se os dados provêm duma população com uma determinada distribuição consiste numa comparação gráfica dos dados com a distribuição teórica. Uma das formas consiste em recorrer aos gráficos de probabilidade *P-P* Plot e *Q-Q* Plot. Existem dois tipos de gráficos de probabilidade normal: o ***Normal P-P Plot,*** que representa a probabilidade acumulada que seria de esperar se a distribuição fosse normal, em função da probabilidade observada acumulada dos erros; e o ***Normal Q-Q Plot,*** que representa o quantil de probabilidade esperado se a distribuição fosse normal em função dos resíduos.

Para elaborar estes gráficos, começa-se por estandardizar os resíduos de forma a terem um desvio padrão unitário:

$$d'_i = \frac{d_i - 0}{S} \; onde \; S^2 = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n} d_i^2}{n-k-1}$$

E ordenam-se por ordem crescente.

Para desenhar os *P-P Plot*:

O valor da função de distribuição para cada resíduo estandardizado, assumindo que tem distribuição normal, é representado no eixo das ordenadas e a probabilidade observada acumulada obtida através da fórmula $\frac{i-0,5}{n}$ vai ser o valor da abcissa.

O *P-P Plot* faz corresponder a função de distribuição teórica com a função de probabilidade acumulada observada nos dados.

Para desenhar os *Q-Q Plot*:

Os quantis de probabilidade esperados, ou seja, os z_i tais que $P(Z < z_i) = \frac{i-0,5}{n}$ serão as ordenadas dos pontos, as abcissas correspondem aos resíduos estandardizados.

O *Q-Q Plot* faz corresponder os quantis esperados com os quantis observados nos dados.

Se os erros possuírem distribuição normal, todos os pontos dos gráficos devem posicionar-se mais ou menos sobre uma reta. Logo se os dados seguirem a distribuição teórica esperada os gráficos serão aproximadamente lineares.

Exemplos de gráficos *P-P Plot* e *Q-Q Plot*:

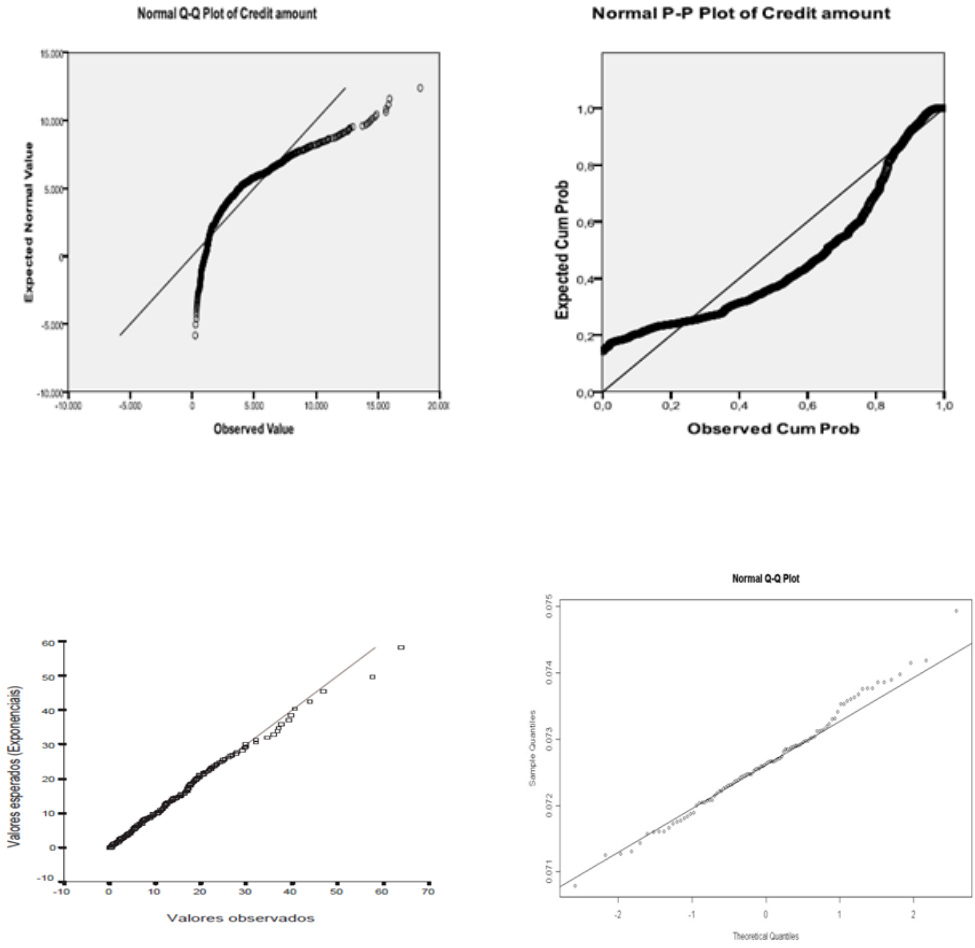


Figura 15: Gráficos P-P Plot e Q-Q Plot

Na figura 15 observa-se que o primeiro par de gráficos evidencia uma curvatura relativamente à reta, o que é indicador de enviesamento, donde somos levados a concluir que a distribuição não é normal. Já no segundo par tal não acontece, evidenciando uma distribuição normal.

- **Curva *ROC***

Uma prática comum na área relacionada com a medicina é a forma de se descrever como e quanto uma variável contínua ou categórica ordinal é capaz de classificar materiais ou indivíduos em grupos definidos.

A análise **ROC** (***R**eceiver **O**perating **C**haracteristc*) é uma ferramenta que permite medir e especificar problemas no desempenho do diagnóstico em medicina. Pode ser feita por meio de um gráfico simples e robusto, que nos permite estudar a variação da sensibilidade e especificidade, para diferentes valores de corte.

A *sensibilidade* (*Sens*.) é definida como a probabilidade do teste fornecer um resultado positivo quando o indivíduo é realmente portador da "doença", enquanto a *especificidade* (*Esp.*) é definida como a probabilidade do teste fornecer um resultado negativo quando o indivíduo não é portador da "doença" (MARGOTTO, 2002).

De outra forma, pode-se dizer que as curvas ROC foram desenvolvidas no ramo das comunicações como uma forma de demonstrar as relações entre sinal-ruído. Neste sentido, podemos interpretar o sinal como os verdadeiros positivos (*sensibilidade*) e o ruído como os falsos positivos (*1 – especificidade*)

A curva ROC é um gráfico de *Sensibilidade* (ou taxa de verdadeiros positivos) *versus* taxa de falsos positivos, ou seja, representa-nos a *Sensibilidade* (ordenadas) e *1 – Especificidade* (abcissas) resultantes da variação de um valor de corte ao longo do eixo de decisão *x* (BRAGA, 2000).

Assim, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da *Sensibilidade* em função da *Especificidade*, correspondente ao ponto que se encontra mais próximo do canto superior esquerdo do diagrama, uma vez que o indício de verdadeiro positivo é 1 e o de falso positivo 0.

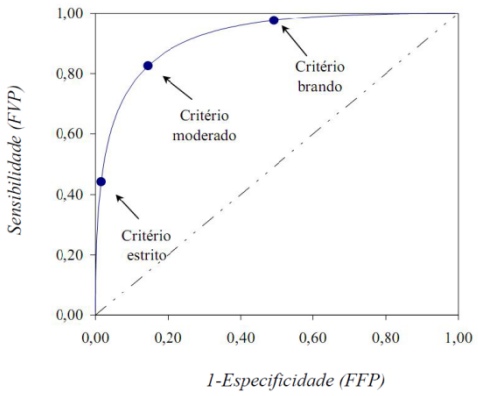


Figura 16: Curva ROC, para uma dada capacidade de discriminação, com a variação do critério de decisão (BRAGA (2000))

O valor do ponto de corte é definido como um valor que pode ser selecionado arbitrariamente pelo investigador 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.

De acordo com Braga (2000), para cada ponto de corte são calculados valores de *Sensibilidade* e *Especificidade*; estes valores podem assim ser dispostos no gráfico. Um

classificador perfeito corresponderia a uma linha horizontal no topo do gráfico, o que é bastante difícil de se obter. Na prática, curvas consideradas boas estarão entre a linha diagonal e a linha perfeita, onde quanto maior a distância da linha diagonal, melhor o sistema. A linha diagonal indica uma classificação aleatória, ou seja, um sistema que aleatoriamente seleciona saídas como positivas ou negativas. Finalmente, a partir de uma curva ROC, devemos poder selecionar o melhor limiar de corte para obtermos o melhor desempenho possível.

- ***Odds Ratio***

Dada uma tabela do tipo:

$Y \backslash X$	x_1	x_2
y_1	b	d
y_2	a	c

O *odds ratio* fornece-nos a força da associação. A figura 17 mostra um gráfico de mosaico de duas variáveis binárias, correspondentes à tabela acima assim como as escalas de medida.

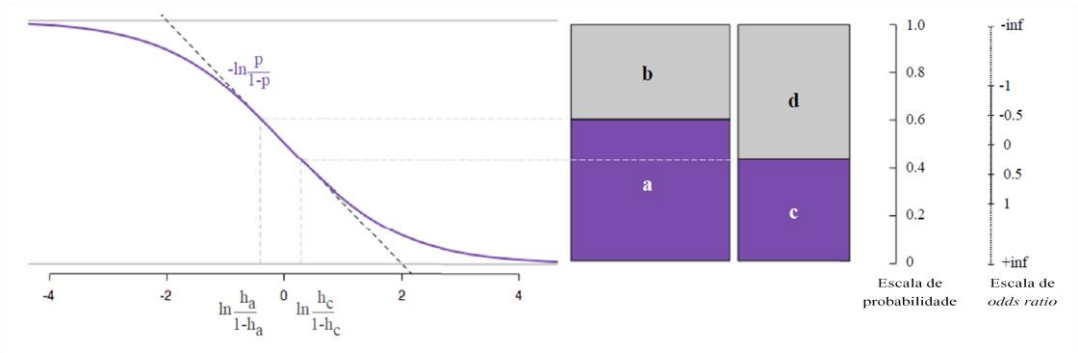


Figura 17: Gráfico de mosaico correspondente à tabela de contingência de 2×2. O gráfico da esquerda relaciona as duas variáveis binárias com o *log odds*. Os valores de *log odds* variam entre -2 e 2.

Com base na "leitura" de valores de uma representação gráfica não podemos obter números precisos, procuramos somente uma avaliação aproximada. No caso dos rácios de probabilidade *log* vamos poder observar algumas das seguintes propriedades:

- Igualdade das alturas de azulejos corresponde a valores de *odds ratio* próximos de zero (indicando independência estatística nos valores subjacentes).
- Comparações entre vários mosaicos permitem avaliar a força das associações (mais fraca e mais forte).
- A comparação do tamanho dos azulejos permite-nos afirmar que: “o *odds ratio* parece ser idêntico" em duas parcelas; ou "um *odds ratio* é de cerca de x vezes superior a outro", onde x é um múltiplo inteiro pequeno.

5. Diabetes mellitus e Periodontite

No nosso estudo, as metodologias descritas anteriormente são exploradas numa aplicação a dados reais no âmbito da Medicina Dentária, com o objetivo de avaliar o grau de relação entre as variáveis, a significância das diferenças entre diabéticos e não diabéticos e construir um modelo válido que conseguisse prever a doença (NA > 4 mm), uma vez que este assunto tem forte impacto no âmbito da Saúde Pública. Os resultados poderão ser úteis para o desenvolvimento de medidas de prevenção. Podem ser vistos como um ponto de partida para novos estudos e, ainda, dadas as implicações em Saúde Pública das doenças em estudo, permitem promover a reflexão, pois são inúmeras as suas implicações económicas que com uma prevenção adequada poderiam ser reduzidas.

Assim, entendemos ser pertinente neste ponto do trabalho, e antes de iniciarmos a parte prática da aplicação, proceder a uma breve revisão de conceitos importantes nas áreas da diabetes mellitus e da periodontite, bem como das suas relações.

A diabetes mellitus é um grupo de doenças metabólicas caraterizadas por hiperglicemia persistente que resulta principalmente de deficiente ação da insulina, secreção de insulina ou ambas. As complicações da diabetes a longo prazo incluem alterações do metabolismo dos hidratos de carbono, proteínas e gorduras; retinopatia com risco de perda de visão; nefropatia que pode levar à falência renal; neuropatia periférica com risco de úlceras nos pés (“pé diabético”); amputações; articulações de Charcot; neuropatia do sistema autónomo com disfunções do trato gastrointestinal, genito-urinário, doença cardiovascular (arterial periférica, cerebrovascular e hipertensão); disfunção sexual; e alterações no metabolismo da lipoproteínas conducentes à dislipidemia (American Diabetes Association, 2001).

A diabetes compreende duas formas de apresentação principais: a diabetes tipo 1 e a tipo 2. A diabetes tipo 1 resulta da destruição das células β do pâncreas; geralmente leva à insulinopenia absoluta e atinge 5-10% dos diabéticos. A diabetes tipo 2 resulta da combinação da resistência à insulina (ação inadequada) e da resposta inadequada de secreção compensatória de insulina, padecendo desta condição 90 a 95% da população diabética.

A diabetes mellitus é atualmente considerada um grave problema de Saúde Pública à escala mundial, tendo-se calculado que, no ano 2000, terão existido 160 milhões de diabéticos (120 milhões nos países em vias de desenvolvimento e 40 milhões nos países desenvolvidos) e que, no ano 2010, o número destes doentes foi superior a 225 milhões (DUARTE, 2002).

No âmbito da Medicina Dentária, a periodontite é uma condição infeciosa, complexa e com grande polimorfismo sintomático, caraterizando-se basicamente por perda de osso alveolar associada à perda de aderência, podendo coexistir com múltiplos sintomas e sinais como inflamação gengival, bolsas de profundidade variável, recessão gengival e mobilidade dentária patológica, culminando a sua evolução clínica na perda dentária.

Neste estudo vamos recorrer a um grupo de variáveis que melhor representam o estado de higiene oral e de saúde periodontal, sendo por isso as mais frequentemente usadas na clínica periodontal.

As variáveis que investigamos com particular atenção no nosso estudo são:

Índice de Placa (IP)

O IP pretende avaliar o grau de higiene oral em termos de presença de placa bacteriana supra gengival.

Para calcular o IP, dos seis pontos observados em torno de cada dente – três pontos vestibulares (mésio-vestibular, vestibular e disto-vestibular) e três pontos linguais (mésio-lingual, lingual e disto-lingual) – foram contabilizados apenas quatro: um mesial (vestibular ou lingual), um distal (vestibular ou lingual), o ponto intermédio vestibular e o ponto intermédio lingual. Foi determinado o número total de pontos que apresentaram placa bacteriana e dividido pelo número total de pontos considerados para este efeito, sendo o resultado apresentado como percentagem.

Os pontos considerados com placa bacteriana foram aqueles que coraram após a aplicação do revelador de placa (eritrosina) e os que, apesar de não terem corado, apresentaram pigmentação superficial e/ou cálculo dentário, ou apresentaram placa que se conseguiu destacar com a ponta da sonda.

Profundidade de Sondagem (PS)

O valor da profundidade de sondagem do sulco/bolsa periodontal obteve-se medindo a distância, em mm, entre a aderência epitelial e o bordo da gengiva livre em seis pontos por dente. Quando o valor observado não foi exato, registou-se o valor inteiro mais próximo, tendo este procedimento de aproximação sido aplicado a todos os parâmetros que foram avaliados em termos absolutos.

Retração Gengival (RG)

A retração gengival corresponde à distância entre a margem da gengiva livre e a linha amelo-cementária, tomando-se esta medida nos seis pontos correspondentes aos pontos de sondagem. Sempre que necessário procedeu-se à aproximação pelo método já referido.

Nível de Aderência Clínica (NA)

O nível de aderência clínica reflete a maior ou menor perda de aderência e corresponde à distância entre a aderência epitelial e a linha amelo-cementária, ou seja, ao valor da profundidade de sondagem adicionado do valor da retração gengival, podendo, também neste caso, estar eventualmente indicado o procedimento de aproximação anteriormente referido.

Índice de Hemorragia Pós-Sondagem (HPS)

A hemorragia pós-sondagem foi avaliada simultaneamente com a sondagem sulco/bolsa periodontal e representa a percentagem do número de sítios sondados que apresentaram hemorragia imediatamente após a sondagem ou nos dois minutos seguintes. Todos os pontos hemorrágicos foram registados independentemente da quantidade e da duração do sangramento (método qualitativo).

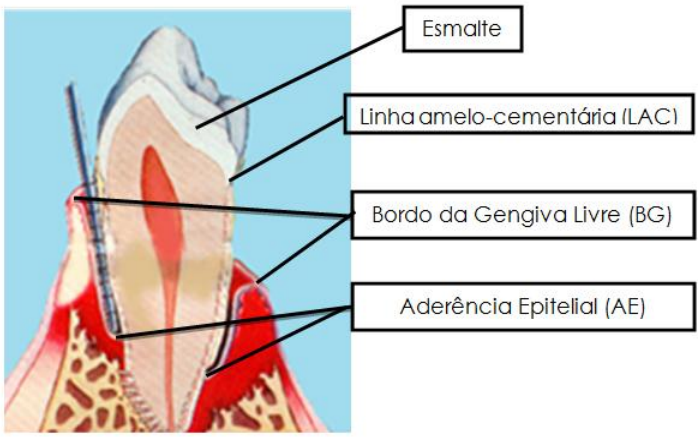


Figura 18: Figura explicativa dos conceitos utilizados nas variáveis clínicas do sistema periodontal

A associação entre a diabetes mellitus e alterações patológicas na cavidade oral, especialmente a doença periodontal, tem sido extensivamente investigada, sendo atualmente muito numerosos os estudos disponíveis na literatura médica e na literatura dentária em que a associação é demonstrada (PRESHAW et al. 2012). Diversas investigações permitiram concluir que existe uma associação entre a doença periodontal e a diabetes mellitus, quer em diabéticos tipo 1, quer em diabéticos tipo 2, como podemos constatar na revisão de LAMSTER et al. (2008). Estudos recentes permitiram ainda concluir que o tratamento da doença periodontal pode contribuir significativamente para melhorar o controlo metabólico da diabetes **(**SGOLASTRA F., 2012**).**

PARTE II

1. Introdução

Atualmente estão em foco as interações entre doenças locais e doenças sistémicas, entre as quais, pela sua importância como problema de saúde pública e pelas suas graves consequências socioeconómicas, se destaca a diabetes mellitus. A sua prevalência tem aumentado ao longo das últimas décadas acompanhando as modificações dos estilos de vida das sociedades modernas e com ela as doenças relacionadas, como por exemplo a doença cardiovascular e a periodontite. Por outro lado, a doença periodontal tem implicações a nível da inflamação sistémica, havendo dados que sugerem estar esta doença implicada em alterações no equilíbrio metabólico dos doentes diabéticos. O estado atual do conhecimento neste campo da saúde, relação entre doença sistémica e doença periodontal, fornece a plausibilidade biológica para a realização deste estudo, que aborda esta associação com ajustamento a outros fatores também relacionados com estas mesmas doenças.

Os dados que iremos utilizar neste estudo foram extraídos duma base de dados mais abrangente obtida num estudo clínico realizado no campo da medicina dentária e endocrinologia. Tal estudo foi aprovado pela Comissão de Ética da Faculdade de Medicina Dentária da Universidade do Porto e pela Comissão de Ética do Hospital de S. João, e todos os participantes assinaram uma declaração de consentimento informado, conforme a "Declaração de Helsínquia" da Associação Médica Mundial.

Alguns dos aspetos metodológicos desse estudo clínico merecem ser mostrados neste trabalho para contextualizar os dados que iremos tratar (PEREIRA, J. (2007)).

Os participantes foram selecionados aleatoriamente a partir dos doentes que frequentaram a Consulta Externa de Endocrinologia do Hospital de S. João, e nos quais foi diagnosticada diabetes tipo 1 ou diabetes tipo 2, e ainda a partir dos indivíduos que constituíam um grupo organizado, de forma aleatória, pelo Serviço de Higiene e Epidemiologia da Faculdade de Medicina do Porto para efeitos de obtenção de controlos. Foram assim selecionados, no total, 158 indivíduos classificados em 2 grupos: não diabéticos e diabéticos (tipo 1 e tipo 2):

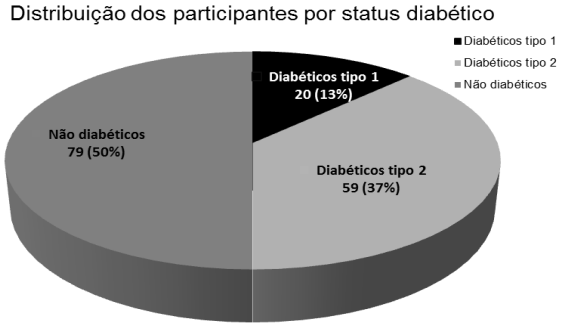


Figura 19: Distribuição dos participantes por *status diabético*

O grupo de diabéticos foi constituído por 79 indivíduos com idades compreendidas entre os 18 e os 79 anos (média = 50,57 (desvio padrão 15,2)), sendo 38 do sexo masculino (48,1%) e 41 do sexo feminino (51,9 %), e foram selecionados consecutivamente a partir dos indivíduos que frequentaram a Consulta de Endocrinologia do Hospital de S. João e aceitaram deslocar-se à Faculdade de Medicina Dentária U.Porto para serem examinados por um médico dentista.

Os participantes não diabéticos foram selecionados a partir de uma listagem fornecida pelo Serviço de Epidemiologia da Faculdade de Medicina U.Porto, e da qual constavam indivíduos que se disponibilizaram para servir de controlos em estudos epidemiológicos. A arrolação dos participantes foi feita de forma a obter pares de indivíduos diabéticos/não diabéticos do mesmo sexo e de idade aproximada a 2 anos. Este grupo foi constituído por 79 indivíduos com idades compreendidas entre os 18 e os 81 anos, tendo como média 50,81 (desvio padrão 15,38), sendo 38 do sexo masculino (48,1%) e 41 do sexo feminino (51,9%).

Os participantes de ambos os grupos obedeceram aos critérios de inclusão definidos no início do estudo (PEREIRA, J., 2007).

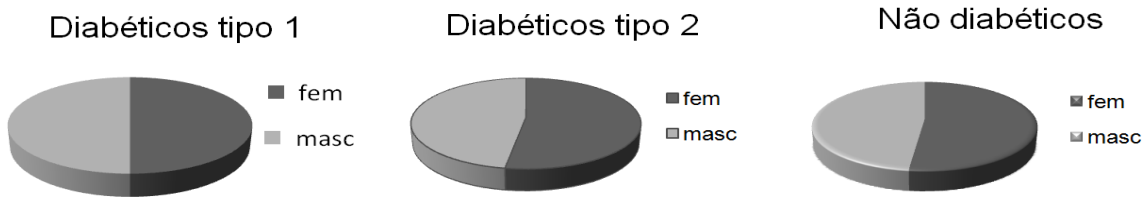


Figura 20: Distribuição dos diabéticos por sexo

Todos os indivíduos foram avaliados quanto aos seguintes parâmetros:

- **Variáveis Sociodemográficas**: Idade e Escolaridade.
- **Variáveis Antropométricas**: Peso, Estatura, Perímetro da Cinta e Perímetro da Anca. Foi calculado o índice de massa corporal [IMC = Peso (kg) / estatura^2 (m^2)], que indica a relação entre peso e altura, permitindo-nos avaliar o estado ponderal do indivíduo.
- **Variáveis Analíticas**: Lipoproteínas de baixa densidade (LDL), lipoproteínas de alta densidade (HDL), Colesterol Total (CT) e Triglicerídeos (T).
- **Hábitos tabágicos**.
- **Variáveis Periodontais:** Índice de Placa (IP), Índice de Hemorragia Pós-Sondagem (HPS), Profundidade de Sondagem (PS), Recessão Gengival (RG), Nível de Aderência (NA).

As variáveis utilizadas nos modelos foram estratificadas conforme a sua associação com a doença periodontal tendo em conta os seguintes fatores:

Género

A estratificação por sexo está fundamentada em estudos epidemiológicos que mostraram menor prevalência e gravidade das doenças periodontais nas mulheres do que nos homens (DESVARIEUX et al., 2004).

Idade >50

Os indivíduos com idades superiores a 50 anos têm maior risco para a perda de osso alveolar, passando a taxa anual de perda de aderência de 0,1 mm/ano até aos 35 anos para 0,3 mm/anos após os 50 anos de idade (AXELSSON et al., 1978). Esta alteração na taxa de progressão da doença pode ser justificada por alterações da imunidade inata e adaptativa associadas com o envelhecimento (PANDA et al., 2009) e redução dos níveis dos esteróides sexuais nos homens (ORWOL et al., 2009) e nas mulheres pós menopausa (MARKOU et al., 2009).

Escolaridade 0-4 /5-9 />=10

O nível de escolaridade é um *surrogate endpoint* do nível socioeconómico da população portuguesa, sobretudo na época em que a maioria dos nossos participantes desenvolveu os seus estudos. Outros autores já estabeleceram relações entre os níveis socioeconómicos (definidos pelos ciclos escolares) e as doenças periodontais (BOILLOT et al., 2004), daí termos adotado idêntica estratificação adaptada para os ciclos escolares portugueses.

IP >90

A placa bacteriana é o principal fator etiológico para as doenças periodontais e fundamental para a instalação da gengivite. Assim, a sua inclusão neste estudo está justificada (OFFENBACHER et al., 2007). O índice de placa superior a 90 tem em conta os valores observados na nossa amostra e a necessidade dos grupos terem dimensão viável para a aplicação das análises estatísticas.

IMC <25

A obesidade e o sobrepeso têm sido associados a maior inflamação sistémica por hiperativação macrofágica, que também desempenham um papel fundamental na regulação da inflamação local, nomeadamente da periodontite (CHAFFEE et al., 2010). O IMC <25 corresponde ao limite normoponderal, daí ter sido escolhido como valor de corte.

Tabaco

Os hábitos tabágicos foram estratificados em fumadores e não fumadores, porque está demonstrado que o tabaco é um fator de risco comportamental/sistémico para a periodontite. A não estratificação por dose de exposição resultou da impossibilidade metodológica de caraterizar com algum rigor a dose de exposição, pois os fumadores, para além do tabaco que consomem diretamente, estão expostos ao tabaco ambiental, e os não fumadores (diretos) também poderão ou não estar expostos ao tabaco ambiental (ARBES et al., 2001). Na impossibilidade de realizar a quantificação da cotinina sérica ou salivar, optamos por simplificar a estratificação dos hábitos tabágicos de forma mais simples (WALTER et al., 2012).

Os valores de corte considerados encontram-se na tabela seguinte:

Tabela 5 : Estratificação das variáveis segundo o risco para a doença periodontal

Variável	Estratificação	Risco
Idade (anos)	≤50	
	>50	++
Género	Feminino	
	Masculino	+
Status diabético	Não diabéticos	
	Diabéticos tipo 1	+++
	Diabéticos tipo 2	+++
Escolaridade (anos)	0-4	+++
	5-9	+
	≥10	
IMC (kg/m^2)	<25	
	≥25	++
Hábitos tabágicos	Não fumador	
	Fumador	+++
IP	≤90%	
	>90%	+++

No nosso estudo, recorrendo a metodologias estatísticas e ferramentas computacionais avançadas, pretendemos avaliar a associação de diferentes fatores de risco estabelecidos e/ou potenciais com a doença periodontal extensa definida pela perda de aderência clínica conforme os critérios descritos na PARTE I. Este critérios tiveram como objetivo evitar uma subavaliação da doença (Nível de Aderência clínica).

Procuramos respostas para as questões abaixo apresentadas:

1 – Comparar diabéticos com não-diabéticos, emparelhados segundo o sexo e a idade, quanto aos indicadores socioeconómicos, dados antropométricos, valores analíticos e indicadores de saúde periodontal.

2 – Avaliar a relação dos indicadores de saúde periodontal tais como Profundidade de Sondagem, Nível de Aderência, Recessão Gengival e Hemorragia Pós-Sondagem, com a Idade, Sexo, GPJ, Colesterol Total, HDL, Triglicerídeos, IMC e Índice de Placa.

3 – Avaliar a associação entre a extensão e gravidade da doença periodontal e o *status* diabético, ajustada para as variáveis independentes de interesse.

Para organização da nossa análise de modo a responder às questões colocadas anteriormente, e de acordo com Tabachnick e Fidell (2007), apresentamos a estrutura da Figura 21.

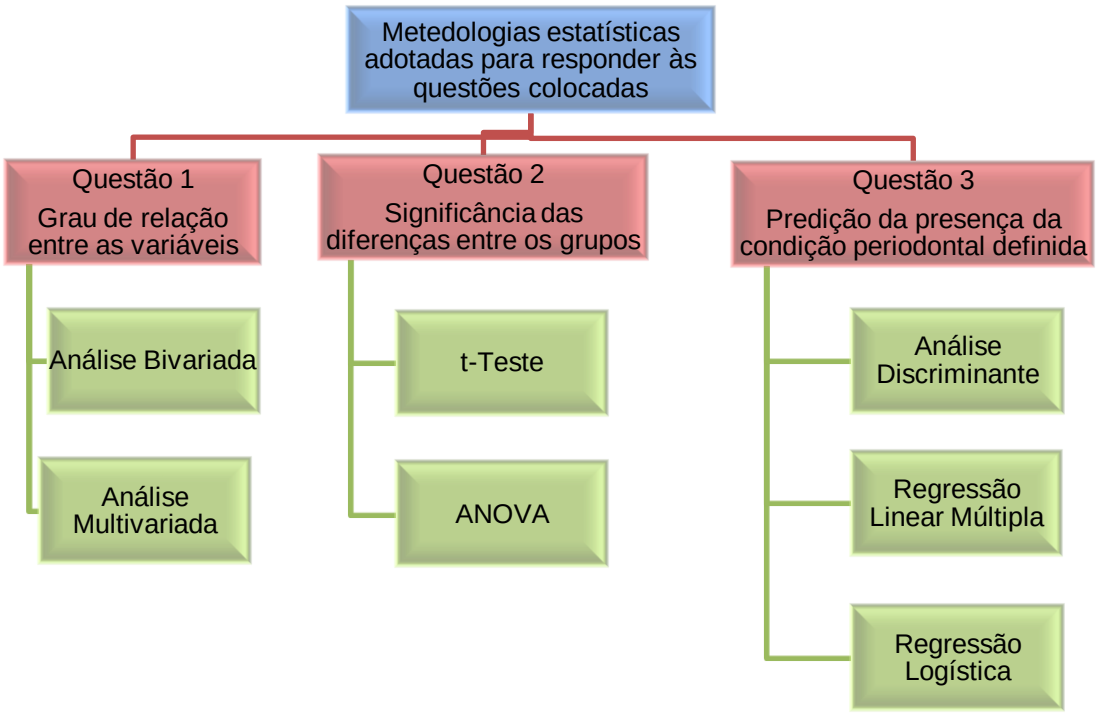


Figura 21: Esquema com a estrutura do estudo

Considerando a relevância do tema do estudo clínico, entendemos que seria justificável explorar outras metodologias estatísticas adequadas e ferramentas computacionais que permitissem explorar os dados de forma a enriquecer a informação científica extraída no estudo inicial.

Passamos à descrição e justificação de seleção das metodologias adotadas. Assim, para comparar as diferenças entre os diabéticos e não diabéticos quanto às variáveis quantitativas recorreu-se aos testes de *t*-Student para observações não emparelhadas e U de Mann-Whitney. Sempre que as variâncias das distribuições a comparar pelo teste *t*-Student foram significativamente diferentes, efetuou-se a correção de Welch. Para comparar variáveis nominais dicotómicas utilizamos o teste do Qui-quadrado e o teste exato de Fisher (sempre que o número de observações foi inferior a 5).

As correlações entre duas variáveis foram avaliadas pelo coeficiente de Pearson ou de Spearman. Esta análise de correlações fez parte dos critérios de seleção das variáveis a incluir no modelo.

No nosso estudo recorremos à análise multivariada para determinar a contribuição de diversas variáveis (Idade, Sexo, Escolaridade, *Status Diabético* e Índice de Placa) para o Nível de Aderência e também para escolher a combinação de variáveis que melhor permite prever o valor do Nível de Aderência (VD). Desta forma pretendemos determinar, numa perspetiva matemática, um modelo linear que melhor estime o valor da VD.

Foi construído um modelo de Regressão Linear para determinar a influência das variáveis *Status Diabético*, Sexo, Idade, Escolaridade e Índice de Placa no Nível de Aderência. Para avaliar o ajustamento do modelo recorremos à ANOVA, cujos pressupostos normalidade, multicolinearidade e homocedasticidade foram avaliados graficamente pelos testes do SPSS: Kolmogorov-Smirnov, de VIF e de tolerância. Também recorremos ao package *Design* no R.

A investigação dos fatores associados à gravidade da saúde periodontal, conforme definida, foi feita utilizando a regressão logística múltipla, pois o que se pretendia era descrever a relação entre o Nível de Aderência (variável dependente ou resposta) e um conjunto simultâneo de variáveis explicativas (preditoras ou independentes) mediante um modelo que tivesse bom ajuste e fosse biologicamente plausível. A análise logística controla grande número de variáveis simultaneamente, permitindo que os dados sejam utilizados mais eficientemente; o teste de homogeneidade pode ser feito em conjunto, bastando introduzir no modelo o termo produto entre os fatores. O ajuste do modelo foi avaliado pelo método da máxima verosimilhança (*maximum likelihood*), que é o método de ajustamento utilizado na regressão logística. Este método estima os parâmetros do modelo de forma a maximizar a probabilidade de encontrar as realizações da variável dependente.

Avaliou-se a significância de cada um dos coeficientes das covariáveis selecionadas no processo anterior através do teste de Wald, considerando o nível de significância de 5%. A

escolha deste teste deve-se ao facto de nos permitir avaliar em simultâneo hipóteses sobre várias combinações lineares dos parâmetros.

Selecionaram-se as covariáveis mais importantes pelo método de *Stepwise*, o qual permite selecionar variáveis a partir de um conjunto inicial de variáveis explicativas. A escolha das variáveis baseia-se num procedimento heurístico, mas não garante, do ponto de vista prático, que o modelo seja o melhor. A qualidade do ajuste do modelo foi realizada com a estatística χ^2 de Pearson, com o teste de Hosmer-Lemeshow e o teste de Deviance. O *teste de Hosmer-Lemeshow* é um teste que avalia o modelo ajustado comparando as frequências observadas e as esperadas, associando aos dados as suas probabilidades estimadas de forma crescente; seguidamente realiza um teste Qui-quadrado para determinar se as frequências observadas estão próximas das frequências esperadas. O *teste de Pearson* fornece-nos uma medida útil para avaliar o quão bem o modelo selecionado ajustou-se aos dados. O *teste de Deviance* do modelo é uma estatística de bondade que se baseia nas funções de log-verosimilhanças maximizadas para verificar se um subconjunto das variáveis pode ser retirado do modelo de regressão logística múltiplo, testando se os coeficientes de regressão são iguais a zero.

A fundamentação teórica das metodologias estatísticas utilizadas, referidas ao longo dos últimos parágrafos, foi apresentada na primeira parte desta dissertação.

As ferramentas computacionais usadas neste trabalho foram o SPSS (*Statistical Package for the Social Sciences*) versão 18, e o R.

A utilização destes dois *softwares* foi justificada pelas razões que passamos a descrever:

O SPSS é um dos *softwares* mais utilizados em diferentes áreas do saber pois possui um ambiente gráfico muito apelativo e de utilização intuitiva, bastando para a maioria das análises efetuar a seleção das respetivas opções em menus e caixas de diálogos. Além disso permite tratar variáveis de diferentes tipos e permite desenvolver todo o processo da investigação, desde o planeamento do estudo até ao tratamento de dados para a análise, possibilitando a elaboração de relatórios, quer pelo próprio programa, quer por uma articulação com um processador de texto (LAUREANO e BOTELHO, 2010).

O R é uma ferramenta poderosa, com boas capacidades ao nível da programação, e possui um vasto número de *packages* (e em constante crescimento), que têm vindo a acrescentar bastantes potencialidades estatísticas e gráficas, o que lhe confere uma crescente importância no contexto atual e internacional. Além disso, é disponibilizado online gratuitamente.

2. Visualização gráfica e análise exploratória dos dados

Analisando a amostra obtivemos, **com recurso ao SPSS**, os gráficos seguintes:

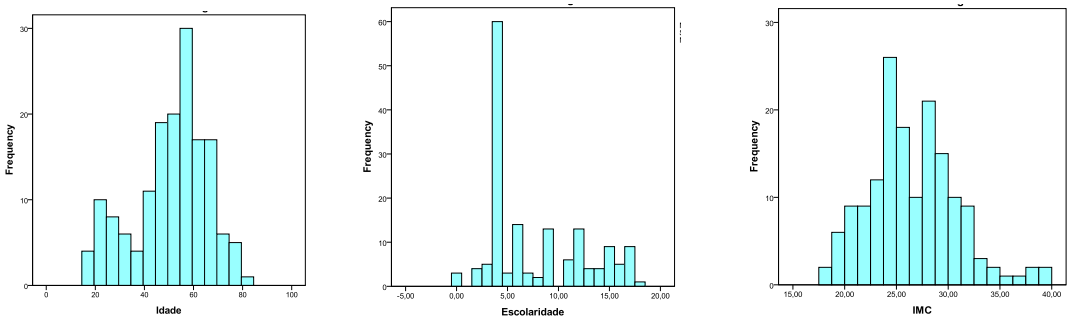


Figura 22: Histogramas das variáveis Idade, Escolaridade e IMC

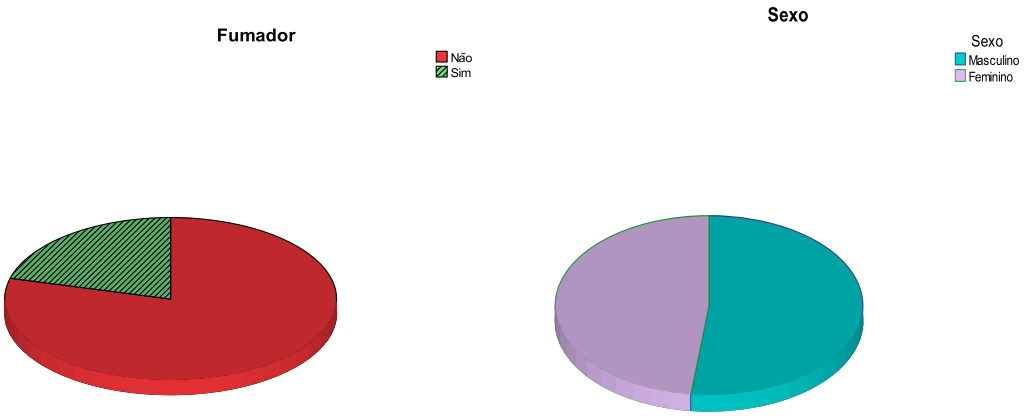


Figura 23: Gráfico circular das variáveis Sexo e Hábitos Tabágicos

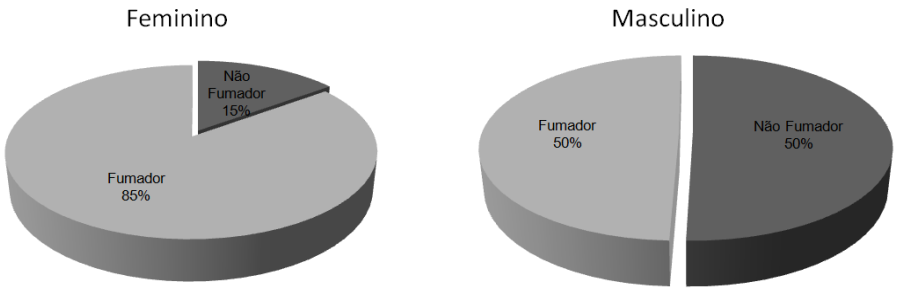


Figura 24: Gráficos circulares comparativo dos Hábitos Tabágicos por sexo

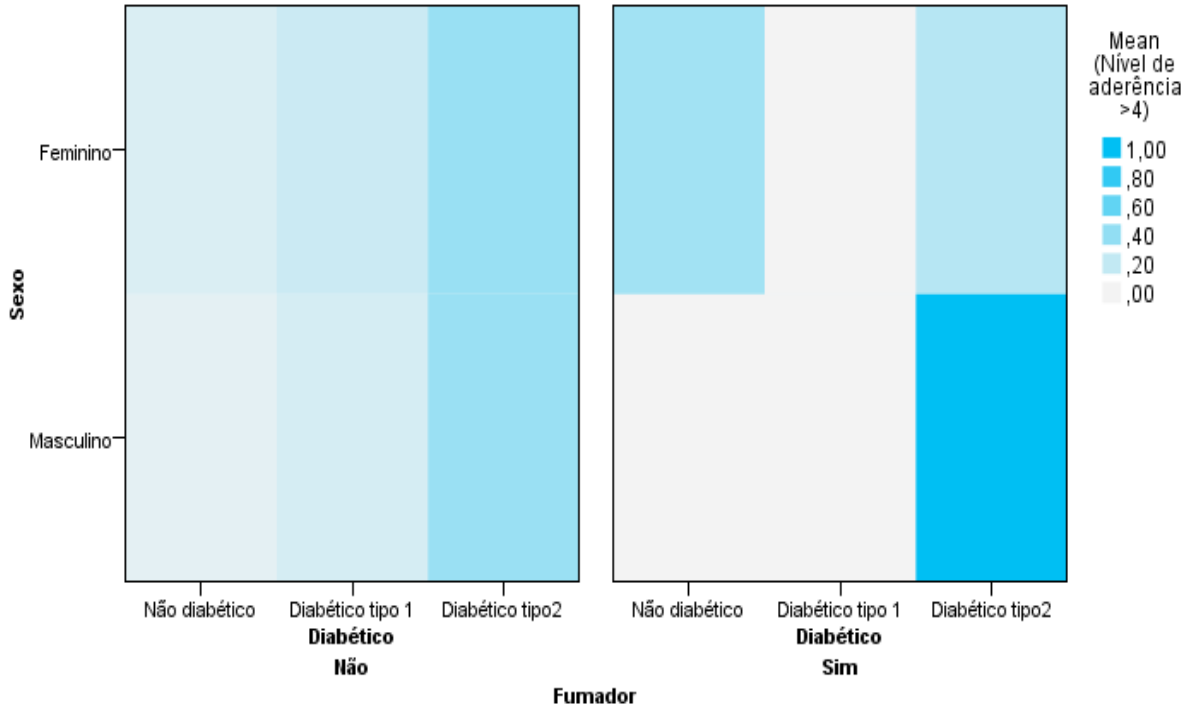


Figura 25: Distribuição do Nível de Aderência em fumadores por sexo e *status diabético*

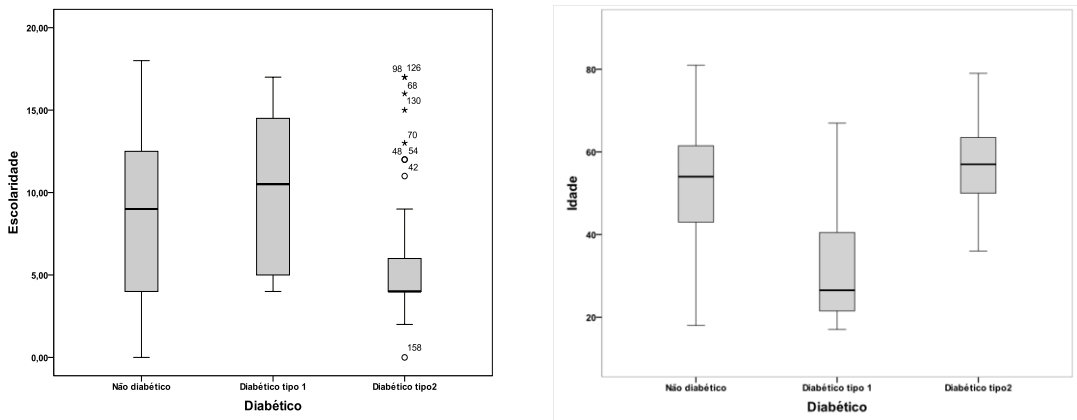


Figura 26: Diagrama de extremos das variáveis Escolaridade e Idade em função do Status Diabético

Com recurso ao **R**, analisámos graficamente a distribuição da variável Nível de Aderência.

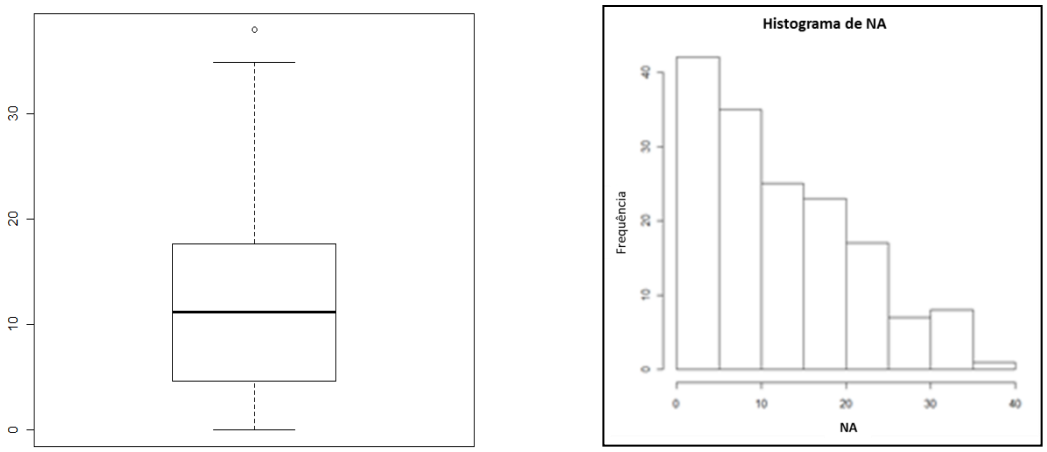


Figura 27: Gráficos da variável Nível de Aderência

Foram realizados os diagramas de dispersão que nos permitem visualizar o grau de associação entre as variáveis e a tendência de variação em conjunto que apresentam. De seguida selecionamos alguns dos gráficos mais elucidativos do estudo da relação entre as variáveis.

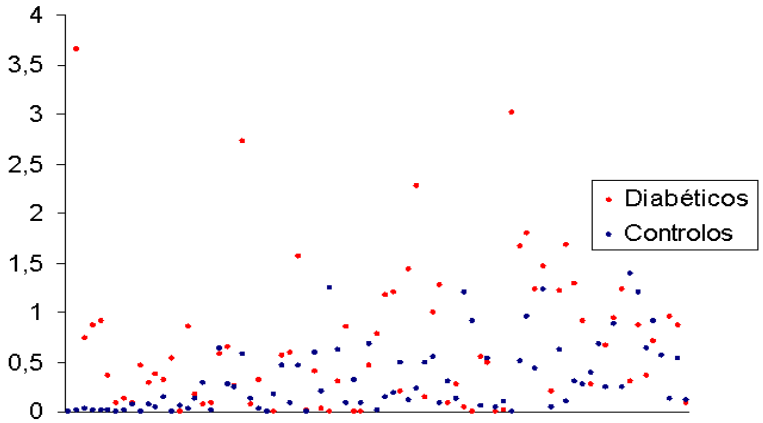


Figura 28: Diagrama de dispersão da variável Nível de Aderência em diabéticos e não diabéticos

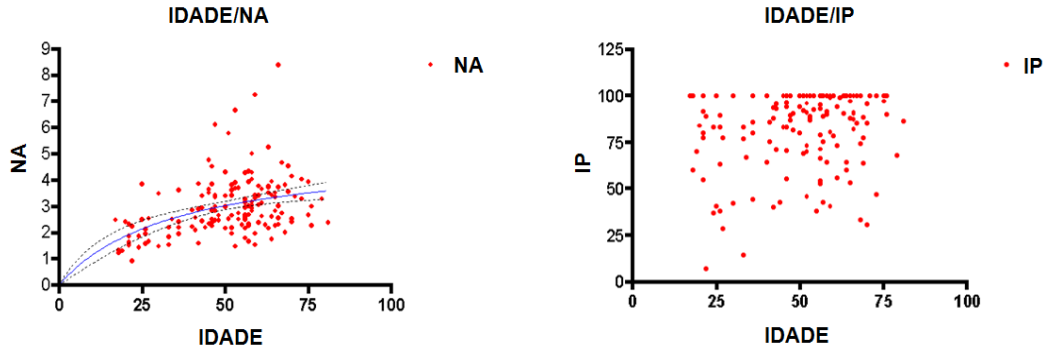


Figura 29: Nuvens de pontos

As nuvens de pontos evidenciam a possível relação entre Idade e Nível de Aderência e a inexistência de relação entre Idade e Índice de Placa.

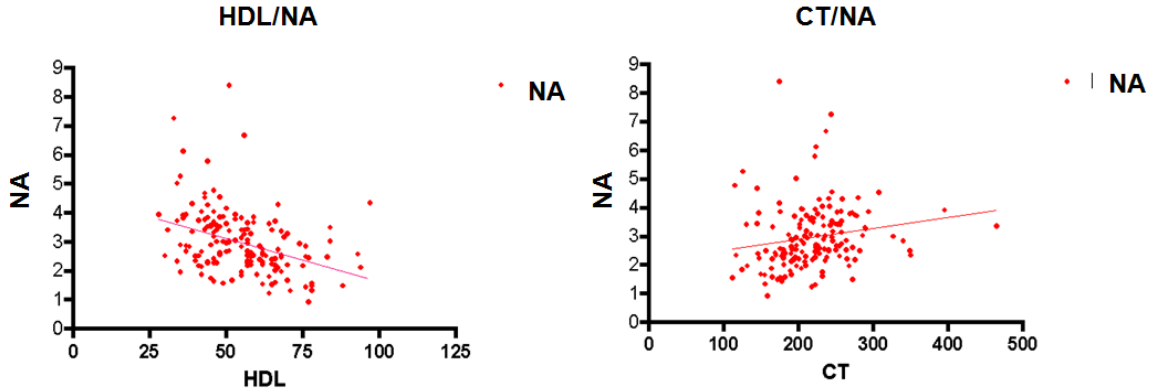


Figura 30: Diagramas de dispersão

Pela observação do diagrama de dispersão podemos inferir que a correlação linear entre HDL e Nível de Aderência é negativa e que entre Nível de Aderência e Colesterol Total é positiva.

3. Testes não paramétricos

- **Comparação entre diabéticos e não diabéticos**

Procedeu-se à comparação de diferentes variáveis no grupo dos diabéticos e não diabéticos, com recurso ao SPSS, efetuando o teste não paramétrico de Mann-Whitney, com o objetivo de avaliar se existiam diferenças entre eles.

3.1. Teste de Mann-Whitney

Para comparação entre diabéticos e não diabéticos, no que respeita às variáveis Idade e Escolaridade, foram consideradas as seguintes hipótese, nula e alternativa:

> $\boldsymbol{H_0}$: As variáveis socioeconómicas têm a mesma distribuição nos diabéticos e não diabéticos.
> $\boldsymbol{H_1}$: As variáveis socioeconómicas não têm a mesma distribuição nos diabéticos e não diabéticos.

Com recurso ao SPSS:

Tabela 6 : Indicadores socioeconómicos dos diabéticos e não diabéticos

	Idade	**Escolaridade** ǂ	♀/♂
Diabéticos	50,57(1,71)	6,71	41/38
Não Diabéticos	50,81(1,73)	8,89	41/38
p	0,921	**0,003**	

A escolaridade média dos não diabéticos (8,89) foi superior à dos diabéticos (6,71). Quando comparados os dois grupos usando o teste de Mann-Whitney observamos uma diferença estatisticamente significativa (p=0,0026). Podemos portanto concluir que existe diferença na escolaridade entre os diabéticos e não diabéticos.

Considerando agora a comparação entre diabéticos e não diabéticos, no que respeita à medida da cinta, anca e Índice de Massa Corporal, foram consideradas as seguintes hipóteses, nula e alternativa:

> $\boldsymbol{H_0}$: Os dados antropométricos têm a mesma distribuição nos diabéticos e não diabéticos.
> $\boldsymbol{H_1}$: Os dados antropométricos não têm a mesma distribuição nos diabéticos e não diabéticos.

Tabela 7 : Dados antropométricos dos diabéticos e não diabéticos			
	Cinta	**Anca**	**Índice de Massa Corporal**
Diabéticos	90,62±11,21	101,60±10,35	27,51±4,46
Não diabéticos	86,68±11,79	98,81±8,32	25,67±3,90
p	**0,0329**	0,0603	**0,0063**

A comparação dos dados antropométricos revelou que o grupo dos diabéticos tinha valores médios superiores nas variáveis observadas, sendo significativas apenas as diferenças do perímetro médio da cinta (C) (p=0,0329) e do IMC médio (p=0,0063) (Tabela 7).

Quanto à comparação entre diabéticos e não diabéticos no que respeita aos valores analíticos e considerando as hipóteses nula

> $\boldsymbol{H_0}$: Os valores analíticos têm a mesma distribuição nos diabéticos e não diabéticos *versus* a hipótese alternativa
>
> $\boldsymbol{H_1}$: Os valores analíticos não têm a mesma distribuição nos diabéticos e não diabéticos

Foram obtidos os resultados seguintes:

Tabela 8 : Valores analíticos dos diabéticos e não diabéticos				
	Colesterol Total	**LDL**	**HDL**	**Triglicerídeos**
Diabéticos	224,91±56,35	138,44±47,34	52,00±14,40	175,13±188,09
Não Diabéticos	212,53±45,70	130,18±41,89	58,01±12,75	119,82±69,45
p	0,1314	0,2738	**0,0061**	**0,0053** [H]

H- Teste de Mann-Whitney

As variáveis analíticas Colesterol Total, LDL e TG apresentaram valores médios superiores nos doentes diabéticos e o HDL inferiores. As diferenças das médias foram estatisticamente significativas para a variável HDL (-6,01 (2,16); IC: -10,25 ; -1,77) e não significativas para Colesterol Total (12,38 (8,16); IC: -3,62 ; 28,38) e LDL (7,81 (7,11); IC: -6,13 ; 2,75). Sendo a distribuição dos valores de T não normal, foram comparadas as medianas dos grupos pelo teste de Mann-Whitney e a diferença encontrada foi significativa (p = 0,0053) (tabela 8).

Conclui-se portanto que há evidência estatística acerca das diferenças no HDL e Triglicerídeos entre os não diabéticos e diabéticos.

Quanto à avaliação da saúde periodontal, foram obtidos os respetivos indicadores que constam na tabela 9:

Tabela 9 : Indicadores da saúde periodontal dos diabéticos e não diabéticos

	PS	NA	RG	HPS	IP
Diabéticos	2,76 (0,73)	3,40 (1,18)	0,68 (0,72)	49,58 (24,27)	85,45 (16,86)
Não Diabéticos	2,19 (0,63)	2,53 (0,85)	0,33 (0,35)	40,86 (27,56)	78,85 (23,72)
p	**< 0,0001**	**< 0,0001** ǂ	**0,0017** ǂ	**0,0366**	0,1111 ǂ

ǂ- Teste de Mann-Whitney

A comparação da situação periodontal entre os doentes diabéticos e não diabéticos, emparelhados para o sexo e idade, evidencia pior saúde periodontal nos diabéticos, avaliada em termos de valores médios de Nível de Aderência, Profundidade de Sondagem, Recessão Gengival, Hemorragia Pós Sondagem e IP (ver tabela 9). A observação dos valores médios dos indicadores da doença periodontal dos dois grupos permitiu-nos constatar que os doentes diabéticos tinham maior Profundidade de Sondagem, Nível de Aderência, Recessão Gengival e Hemorragia Pós Sondagem.

3.2. Interpretação gráfica

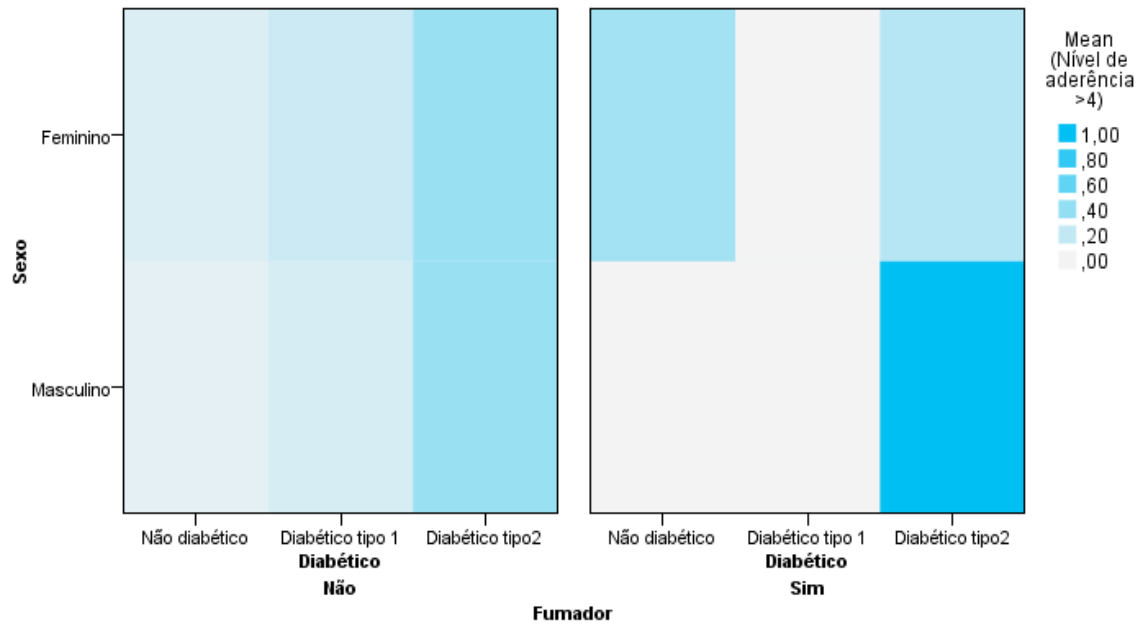


Figura 31: Comparação de grupos relativamente aos valores do Nível de Aderência

Como podemos observar, o maior Nível de Aderência é verificado na figura no canto inferior direito, correspondente a Fumadores masculinos com diabetes tipo 2. Podemos ainda

observar que os indivíduos que apresentam os valores inferiores do Nível de Aderência são masculinos, não diabéticos ou diabéticos tipo 1 e fumadores. Globalmente, os não fumadores têm os menores valores do Nível de Aderência (no gráfico do lado esquerdo não encontramos a cor correspondente ao valor máximo).

3.3. Teste de t de Student

Recorreu-se ao teste t de Student para comparar as diferenças entre os diabéticos e não diabéticos para observações não emparelhadas, tendo sido testadas as hipóteses nulas H_0: $\mu_X-\mu_Y=0$ ($\mu_X=\mu_Y$) *versus* H_1: $\mu_X-\mu_Y>0$ ($\mu_X>\mu_Y$) para a média da Profundidade de Sondagem e da Hemorragia pós sondagem e as diferenças observadas foram estatisticamente significativas, sendo os respetivos intervalos de confiança (IC: 0,36 ; 0,79) e (IC: 0,62 ;16,81). Um intervalo de confiança dá-nos uma estimativa da amplitude dos valores que provavelmente incluirão o parâmetro populacional desconhecido. A estimativa dessa amplitude (intervalo) é calculada a partir de um conjunto de dados de uma amostra.

Se muitos pares de amostras (no caso dependentes) são extraídas repetidamente da mesma população, e um intervalo de confiança é calculado para cada par de amostras, então uma certa percentagem (nível de confiança) destes intervalos incluirão o parâmetro populacional desconhecido. No nosso exemplo podemos afirmar que, com um nível de confiança de 95%, o valor médio da Profundidade de Sondagem se encontra entre 0,36 e 0,79, e analogamente a Hemorragia Pós Sondagem se encontra entre 0,62 e 16,81.

As variáveis Nível de Aderência e Recessão Gengival foram comparados pelo teste de Mann-Whitney, uma vez que este teste não exige o pressuposto da normalidade, podendo ser aplicado para amostras pequenas, e em variáveis de escala ordinal. Tendo-se testado as hipóteses:

> $\boldsymbol{H_0}$: As duas amostras são provenientes de populações com a mesma distribuição
> *versus*
> $\boldsymbol{H_1}$: As duas amostras são provenientes de populações com distribuições distintas

verificou-se serem as diferenças das suas medianas estatisticamente significativas com $p<0,0001$ para o Nível de Aderência e $p=0,0017$ para a Recessão Gengival (tabela 9).

Conclusão global: Comparando os diabéticos com os não diabéticos, verificou-se que só não existem diferenças quanto ao Índice de Placa. Para as restantes variáveis (Nível de Aderência, Hemorragia Pós Sondagem e Recessão Gengival) os dois grupos apresentam diferenças significativas, independentemente do método utilizado na avaliação.

4. Estudo das Correlações

Foram determinadas as correlações entre os indicadores da saúde periodontal Profundidade de Sondagem, Nível de Aderência, Recessão Gengival e Hemorragia Pós Sondagem, e as variáveis independentes: Idade, GPJ, Colesterol Total, HDL, LDL, Triglicerídeos, IMC, Índice de Placa, pelo coeficiente de Pearson ou de Spearman quando bivariada e recorrendo ao SPSS, sendo que as opções no grupo *Correlation Matrix* apresentam a matriz de correlações entre variáveis (ver anexo I). Recorremos ao cálculo do coeficiente de Pearson com o R para avaliar a correlação das variáveis Nível de Aderência e *Status Diabético*.

Dos valores encontrados podemos destacar as seguintes relações:

4.1. Relação entre o Nível de Aderência (NA) e as variáveis independentes

Os valores médios do Nível de Aderência dos não diabéticos e dos diabéticos estão diretamente correlacionados com a Idade (r=0,44; p<0,0001: r=0,51; p<0,0001), o HDL (r=-0,45; p<0,0001: r=-0,28; p=0,0119), Triglicerídeos (r=0,26; p=0,023: r=0,28; p=0,0124) e Índice de Placa (r=0,51; p<0,0001: r=0,30; p<0,0001).

Apenas nos não diabéticos as variáveis Colesterol Total (r=0,33; p=0,004), LDL (r=0,41; p=0,0002) e Índice de Massa Corporal (r=0,34; p=0,002) apresentam diferenças entre os coeficientes de correlação significativas para LDL (z=2,04) e não significativas as restantes variáveis.

É de referir que para a variável HDL o valor do coeficiente de correlação é sempre negativo, pelo que podemos afirmar que um agravamento dos indicadores de saúde periodontal estão associados a uma diminuição de HDL. Tal associação é estatisticamente significativa em geral para o grupo dos não diabéticos, enquanto que para os diabéticos só é estatisticamente significativa quanto ao Nível de Aderência e Receção Gengival.

4.2. Relação entre o Nível de Aderência e o *status diabético* com recurso ao R

Para avaliarmos a relação entre o Nível de Aderência e o *status diabético* recorremos ao **R**, uma vez que este programa é, neste caso, de utilização muito intuitiva e fornece-nos resultados de fácil leitura.

Utilizamos o comando *lm* usado para a regressão linear simples. A indicação de qual a variável resposta *y* e quais as variáveis preditoras x_1, ...,x_p faz-se através do argumento *function*.

Objetivos:

1. Determinar os coeficientes da reta de regressão utilizando o método dos mínimos quadrados. Escrever a equação teórica do modelo e descrever os parâmetros.

2. Determinar o coeficiente de correlação linear de Pearson e verificar se é significativamente diferente de zero.

3. Criar uma tabela de ANOVA e com recurso ao teste F testar a hipótese nula β =0.

4. Calcular o coeficiente de determinação.

Rotinas:

```
> #Método dos minimos quadrados
> mmq=lm(diabetes~Na)
> mmq

Call:
lm(formula = diabetes ~ Na)

Coefficients:
(Intercept)           Na
    0.31789      0.01501

> #Coeficiente de correlação
> cor.test(Na,diabetes)

        Pearson's product-moment correlation

data:  Na and diabetes
t = 3.5986, df = 156, p-value = 0.0004292
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1261672 0.4150520
sample estimates:
      cor
0.2768536

> #ANOVA
> anova(mmq)
Analysis of Variance Table

Response: diabetes
           Df Sum Sq Mean Sq F value    Pr(>F)
Na          1  3.028  3.0276  12.950 0.0004292 ***
Residuals 156 36.472  0.2338
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Intervalo de Confiança a 95%
> confint(mmq,level=0.95)
                  2.5 %     97.5 %
(Intercept) 0.192324225 0.44345140
Na          0.006770855 0.02324915
```

Interpretação dos resultados:

1. Como a variável diabetes é dicotómica, o diagrama de dispersão permitiu-nos comparar os grupos quanto à gravidade da doença periodontal avaliada pelo Nível de Aderência. Assim podemos concluir que nos diabéticos a doença (Nível de Aderência) é mais evidente que nos não diabéticos.
2. O coeficiente de variação é inferior a 0,50, logo o grau de dispersão é pequeno e podemos concluir que a média é representativa.
3. Como estamos a determinar a 95% de confiança o α = 0,05 e como o *p-value* = 0,0004 é menor que α, logo rejeitamos a hipótese nula $\boldsymbol{H_0}$ de que as variáveis não se encontram associadas. Há evidências significativas de que o Nível de Aderência e *Status Diabético* se encontram relacionados. Quanto à ANOVA, os resultados mostram que o modelo é altamente significativo pois o p-value é próximo de 0,000. Considera-se portanto que o parâmetro β é significativamente diferente de zero.
4. O coeficiente de determinação obtido igual a 0,077 afasta a hipótese de linearidade dos dados, uma vez que este deve situar-se entre 0 e 0,1, sugerindo uma correlação ínfima positiva.

5. Análise de Regressão Múltipla

O objetivo da análise de regressão múltipla é determinar a força de cada uma das Variáveis Independentes que, em conjunto, melhor explicam o comportamento da Variável Dependente. No nosso estudo pretendemos prever mudanças da variável Nível de Aderência associadas a mudanças das variáveis independentes (Idade, Sexo, Escolaridade, *Status Diabético*, Fumador e Índice de Placa).

As variáveis com correlações mais fortes serão aquelas que conduzirão a um modelo que melhor irá predizer a Variável Dependente. Porém, como o nosso estudo foi feito no âmbito da Medicina Dentária, a escolha das Variáveis Independentes a incluir tem relevância na interpretação biológica do fenómeno. Não se trata da escolha cega de variáveis mas sim da seleção de variáveis nas condições anteriormente indicadas que tenham significado no contexto do problema.

5.1. Regressão linear múltipla

Foi construído um modelo de Regressão Linear, para conhecer quanto e se as variáveis *Status Diabético*, Sexo, Idade, Escolaridade e Índice de Placa influenciam o Nível de Aderência:

$$Y = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 . X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon$$

Onde em todos os modelos desenhados temos:

Y – representa o Nível de Aderência, com i a variar entre 1 e 158

X_{1i} – representa a variável Diabetes, do tipo dicotómico

X_{2i} – representa a variável Sexo, do tipo dicotómico

X_{3i} – representa a variável Idade, do tipo contínuo

X_{4i} – representa a variável Escolaridade, do tipo contínuo

X_{5i} – representa a variável Fumador, do tipo dicotómico

X_{6i} – representa a variável Índice de Placa, do tipo contínuo

β_0 – é o interceto do plano de regressão (coeficiente linear).

β_1 , β_2 , ... β_6 – são coeficientes de regressão (coeficientes angulares)

ε – erro experimental ou variabilidade residual

Recorrendo ao SPSS obteve-se:

Tabela 10 : Variáveis incluídas no Modelo de regressão linear simples			
Modelo	**Variáveis incluídas**	**Variáveis excluídas**	**Método**
1	IP, Sexo, Diabético, Fumador, Idade, Escolaridade[a]	___	Enter

a. Test distribution is Normal.

Tabela 11 : Resumo do Modelo de regressão linear simples				
Modelo	**r**	**Quadrado de r**	**Quadrado de r ajustado**	**Erro Padrão da Estimativa**
1	0,619	0,384	0,359	7,43606

A tabela 10, apresenta o sumário do modelo, com as variáveis (Variáveis preditoras: Constante), IP, Sexo, *Status Diabético*, Idade, Escolaridade, Fumador, cuja construção será feita posteriormente.

Neste modelo encontramos r^2_{aju}=0,384, donde podemos afirmar que 38,4% da variabilidade do Nível de Aderência é explicada pelas variáveis independentes do modelo ajustado. O valor do coeficiente de correlação é r=0,619. Ou seja, 61,9% da variabilidade encontrada para o Nível de Aderência pode ser explicada pelas variáveis independentes, ficando neste caso por explicar cerca de 39%, que se devem a outros fatores. Não podemos considerar que este modelo seja um bom ajuste pois o valor de r^2 não se encontra próximo de 1. Não existe pois uma forte relação entre as variáveis.

5.2. ANOVA para testar a significância do modelo

Quando colocamos a questão "Será ρ significativamente diferente de 0?", ou seja, será o modelo ajustado significativo?, temos que observar a tabela ANOVA (tabela 12).

Tabela 12 : ANOVA (Variáveis preditoras: Constante), IP, Sexo, Diab., Idade, Escol, Fumador					
Modelo	**Soma dos Quadrados**	**Graus de liberdade**	**Quadrado Médio**	**F**	**Sinal**
Regressão	5196,977	6	866,163	15,644	0,000
Residual	8349,543	151	55,295		
Total	13546,520	157			

A ANOVA apresentou um *p-value* de 0,000, ou seja, **o modelo é altamente significativo**, para qualquer nível de significância. Logo este modelo podia ser generalizado a outras amostras. O modelo ajustado (tabela 13) é dado por:

$$\widehat{NA} = -1{,}976 + 3{,}117\ \boldsymbol{St.Diab} - 0{,}379\ \boldsymbol{Sexo} + 0{,}165\ \boldsymbol{Idade} - 0{,}346\boldsymbol{Escol} + 2{,}667\ \boldsymbol{Fum.} + 0{,}064\ \boldsymbol{IP}$$

Tabela 13 : Coeficientes (Variáveis dependente: NA)

Modelo	Coeficientes não padronizados		Coeficientes padronizados	t	Sinal
	B	Erro Std.	Beta		
(Constante)	-1,976	4,377		-0,451	0,652
Diabético	3,117	0,677	0,312	**4,604**	**0,000**
Sexo	-0,379	1,203	-0,020	-0,315	0,754
Idade	0,165	0,047	0,271	3,529	0,001
Escolaridade	-0,346	0,159	-0,178	-2,173	0,031
Fumador	2,667	1,532	0,117	1,741	0,084
IP	0,064	0,031	0,144	2,068	0,040

A coluna t dá-nos os valores observados das estatísticas dos testes de t de Student aplicados aos coeficientes de regressão, que têm como finalidade testar a significância dos parâmetros estimados do modelo. Com um nível de confiança de 95%, apenas a variável *Status Diabético* é significante, sendo aquela que tem maior contribuição individual (4,604).

5.3. Validação dos pressupostos do modelo

A validação dos pressupostos do modelo assenta na análise de resíduos. Com esse objetivo desenhámos o diagrama de dispersão. Da sua observação é razoável afirmar que existe uma relação linear entre as variáveis, uma vez que os resíduos se distribuem de forma mais ou menos aleatória em torno do zero, sugerindo a validade do pressuposto de independência e da homocedasticidade. Porém, não nos dá qualquer informação sobre o pressuposto da normalidade dos resíduos.

Da interpretação do Normal P-P dos resíduos podemos concluir que o pressuposto da normalidade também é válido, pois a maioria dos pontos está sobre a diagonal principal.

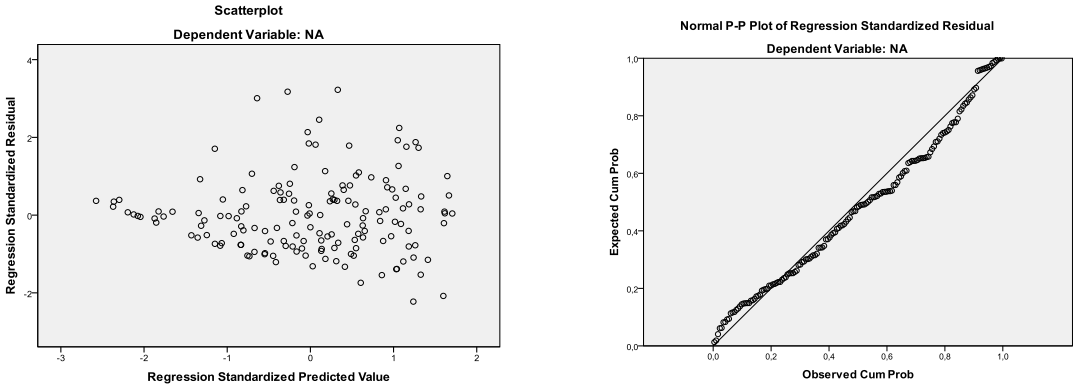


Figura 32: Gráficos dos Resíduos *versus* preditos; resíduos padronizados e da probabilidade normal dos resíduos

Procedamos pois à realização de testes exatos da ocorrência de *outliers* e ao pressuposto da distribuição normal. Assim, para testar a normalidade optou-se por usar o teste Kolmogorov-Smirnov que é um teste paramétrico tradicional, baseado na distribuição *t*-Student e é obtido sob a hipótese de que a população tem distribuição normal, e o teste de Shapiro-Wilk, por se tratar de um teste de ajustamento específico para a distribuição normal que tem uma melhor performance que o teste anterior em amostras reduzidas (n < 30).

Tabela 14 : Teste One-Sample Kolmogorov-Smirnov		
N		**158**
Parâmetros da distribuição normal	Média	0,000
	Desvio padrão	7,365
Diferenças mais Extremas	Absoluto	0,069
	Positivo	0,069
	Negativo	-0,049
	Kolmogorov-Smirnov Z	0,869
	Asymp. Sig. (2-caudas)	0,436
	Exact Sig. (2-caudas)	0,418
	Point Probability	0,000

O *p-value* (exato) é 0,418, logo não rejeitamos a hipótese de que a variável em estudo segue uma distribuição normal para o nível de significância de $\alpha = 0,05$.

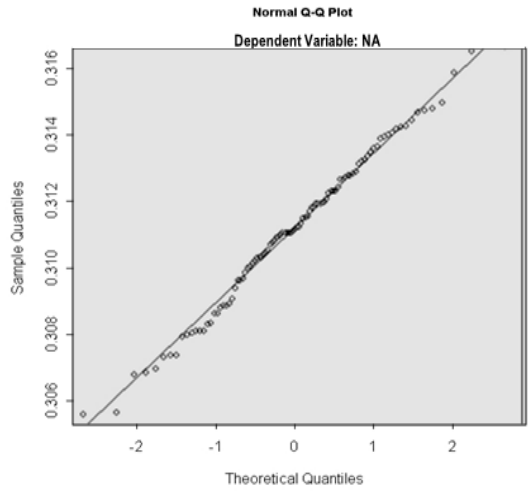


Figura 33: Q-Q plot

Usando a variável RES (os resíduos guardados) e fazendo um *Q–Q plot* (figura 33) e os testes de ajustamento de Kolmogorov-Smirnov e de Shapiro-Wilk podemos concluir que os resíduos têm uma distribuição normal (o *Q–Q plot* identifica um ajuste entre os quantis amostrais e os quantis de distribuição normal) e os testes de ajustamentos fornecem os *p-values* superiores aos níveis de significância usual (0,062 > 0,05).

Tabela 15 : Teste de normalidade

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	**Graus de liberdade**	**Sinal**	**Estatística**	**Graus de liberdade**	**Sinal**
Erros Padronizados	0,069	158	0, 062	0,959	158	0,000

a. Correção de significância de Lilliefors

Tabela 16 : Verificação da multicolinearidade (Variáveis dependente: NA)

	Modelo	Coeficientes não padronizados		Coeficientes padronizados	t	Sinal	Estatísticas de colinearidade	
		β	Erro Padrão	Beta			Tolerância	VIF
1	**(Constante)**	19,241	1,250		15,395	0,000		
	Escolaridade	-0,916	0,137	-0,473	-6,706	0,000	1,000	1,000
2	**(Constante)**	15,146	1,517		9,983	0,000		
	Escolaridade	-0,732	0,136	-0,378	-5,366	0,000	0,902	1,109
	Diabético	3,041	0,704	0,304	4,319	0,000	0,902	1,109
3	**(Constante)**	5,700	3,285		1,735	0,085		
	Escolaridade	-0,488	0,153	-0,252	-3,194	0,002	0,679	1,473
	Diabético	3,066	0,684	0,307	4,484	0,000	0,902	1,109
	Idade	0,148	0,046	0,244	3,217	0,002	0,735	1,360
4	**(Constante)**	-0,627	4,308		-0,146	0,884		
	Escolaridade	-0,381	0,158	-0,196	-2,407	0,017	0,617	1,622
	Diabético	2,998	0,676	0,300	4,436	0,000	0,900	1,111
	Idade	0,145	0,046	0,239	3,192	0,002	0,735	1,361
	IP	0,069	0,031	0,155	2,232	0,027	0,849	1,178

Para avaliar a multicolinearidade, o SPSS utiliza a Tolerância de cada variável que é a medida da proporção da variância da variável que não é explicada pelas restantes variáveis independentes e que é calculada aquando da aplicação do método Stepwise, onde se vão selecionar as variáveis que vão entrar na análise. As variáveis do nosso estudo que se encontram nestas condições (Tolerância > 0,8) são a Escolaridade e Diabetes, assim revelando poder discriminante, pelo que não há a recear a violação do pressuposto de multicolinearidade.

5.4. Ajustamento do Modelo pelo Método Stepwise

A partir da tabela 17 também podemos verificar quais as variáveis que são significantes para o modelo, ou seja, aquelas cujo coeficiente apresenta um valor significativamente diferente de zero.

Tabela 17 : Diagnóstico de colinearidade (Variável Dependente: NA)

Modelo	Dimensão	Valores próprios	Condition Index	Variance Proportions						
				Constante	Idade	Sexo	Diab.	Escol.	Fumador	IP
1	**1**	4,922	1,000	0,00	0,00	0,01	0,01	0,00	0,01	0,00
	2	0,803	2,476	0,00	0,00	0,01	0,08	0,00	0,71	0,00
	3	0,540	3,020	0,00	0,00	0,06	0,48	0,08	0,15	0,00
	4	0,438	3,353	0,00	0,00	0,89	0,01	0,05	0,04	0,00
	5	0,234	4,585	0,00	0,06	0,01	0,38	0,30	0,00	0,02
	6	0,050	9,903	0,00	0,49	0,00	0,01	0,02	0,06	0,56
	7	0,014	19,080	1,00	0,45	0,00	0,03	0,55	0,03	0,41

De seguida procedemos ao ajuste do nosso modelo recorrendo ao Método de *Stepwise*.

Tabela 18: Variáveis incluídas e excluídas do modelo (Variáveis dependente: NA)

Modelo	Variáveis incluídas	Variáveis removidas	Método
1	Escolaridade		Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100)
2	Diabético		Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100)
3	Idade		Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100)
4	IP		Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100)

Tabela 19 : Sumário do modelo (Variáveis dependente: NA)

Modelo	r	Quadrado de r	Quadrado de r ajustado	Erro Padrão do estimador	Durbin-Watson
1	0,473[a]	0,224	0,219	8,21011	
2	0,554[b]	0,307	0,298	7,78158	
3	0,592[c]	0,351	0,338	7,55694	
4	0,609[d]	0,371	0,355	7,46112	1,919

a. Predictors: (Constant), Escolaridade; **b.** Predictors: (Constant), Escolaridade, Diabético

c. Predictors: (Constant), Escolaridade, Diabético, Idade **d.** Predictors: (Constant), Escolaridade, Diabético, Idade, IP

e. Dependent Variable: NA_4

Pela leitura da tabela 19, podemos concluir que o teste de significância da equação de Regressão Linear Múltipla indicou que todos os modelos construídos podem ser considerados significativos para um nível de significância de 5%, uma vez que o F calculado é maior que o F crítico. Assim, rejeitamos a hipótese H_0, o que quer dizer que as variâncias são iguais e consequentemente os modelos de regressão são válidos. Como os *p-value* são

todos inferiores a 0,05, podemos assegurar que qualquer um dos quatro modelos de regressão (descritos a seguir) são melhores que a média para predizer os valores do Nível de Aderência.

Tabela 20 : ANOVA (Variáveis dependente: NA)

Modelo		Soma dos Quadrado	Graus de liberdade	Média dos Quadrados	F	Sinal
1	Regressão	3031,203	1	3031,203	44,969	,000[a]
	Resíduo	10515,317	156	67,406		
	Total	13546,520	157			
2	Regressão	4160,805	2	2080,402	34,357	,000[b]
	Resíduo	9385,715	155	60,553		
	Total	13546,520	157			
3	Regressão	4751,983	3	1583,994	27,737	,000[c]
	Resíduo	8794,537	154	57,107		
	Total	13546,520	157			
4	Regressão	5029,260	4	1257,315	22,586	,000[d]
	Resíduo	8517,260	153	55,668		
	Total	13546,520	157			

Tabela 21 : Coeficientes (Variáveis dependente: NA)

Modelo		Coeficientes não padronizados		Coeficientes padronizados	t	Sinal	Estatísticas de colinearidade	
		B	Erro Padrão	Beta			Tolerância	VIF
1	**(Constante)**	19,241	1,250		15,395	0,000		
	Escolaridade	-0,916	0,137	-0,473	-6,706	0,000	1,000	1,000
2	**(Constante)**	15,146	1,517		9,983	0,000		
	Escolaridade	-0,732	0,136	-0,378	-5,366	0,000	0,902	1,109
	Diabético	3,041	0,704	0,304	4,319	0,000	0,902	1,109
3	**(Constante)**	5,700	3,285		1,735	0,085		
	Escolaridade	-0,488	0,153	-0,252	-3,194	0,002	0,679	1,473
	Diabético	3,066	0,684	0,307	4,484	0,000	0,902	1,109
	Idade	0,148	0,046	0,244	3,217	0,002	0,735	1,360
4	**(Constante)**	-0,627	4,308		-,146	0,884		
	Escolaridade	-0,381	0,158	-0,196	-2,407	0,017	0,617	1,622
	Diabético	2,998	0,676	0,300	4,436	0,000	0,900	1,111
	Idade	0,145	0,046	0,239	3,192	0,002	0,735	1,361
	IP	0,069	0,031	0,155	2,232	0,027	0,849	1,178

a. Predictors: (Constant), Escolaridade; b. Predictors: (Constant), Escolaridade, Diabético; c. Predictors: (Constant), Escolaridade, Diabético, Idade; d. Predictors: (Constant), Escolaridade, Diabético, Idade, IP; e. Dependent Variable: NA_4

A tabela 21 permite-me escrever a equação que nos dá uma estimativa do Nível de Aderência em cada um dos modelos.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_{1i} + \hat{\beta}_2 \cdot X_{2i}$$

No Modelo 1:

$\hat{Y}_i = 19{,}241 - 0{,}916 \cdot X_{4i}$, onde Y_i é o valor do Nível de Aderência e X_{4i} é o valor da Escolaridade

No Modelo 2:

$$\hat{Y}_i = 15{,}146 - 0{,}732 \cdot X_{4i} + 3{,}041 \cdot X_{1i}$$

onde Y_i o valor do Nível de Aderência e X_{4i} e o valor da Escolaridade e X_{1i} o valor do Status Diabético.

No Modelo 3:

$$\hat{Y}_i = 5{,}700 - 0{,}488 \cdot X_{4i} + 3{,}066 \cdot X_{1i} + 0{,}148.X_{3i}$$

onde Y_i o valor do Nível de Aderência e X_{4i} e o valor da Escolaridade, X_{1i} o valor do Status Diabético e X_{3i} é a Idade.

No Modelo 4:

$$\hat{Y}_i = -0{,}627 - 0{,}381 \cdot X_{4i} + 2{,}998 \cdot X_{1i} + 0{,}145 X_{3i} + 0{,}069 X_{6i}$$

onde Y_i o valor do Nível de Aderência e X_{4i} e o valor da Escolaridade, X_{1i} o valor do Status Diabético , X_{3i} é a Idade e X_{6i} é o Índice de placa.

Tabela 22 : Variáveis excluídas (Variáveis dependente: NA)

Modelo		Beta in	t	Sinal	Parcial	Estatísticas de colinearidade		
						Tolerância	VIF	Tolerância mínima
1	**Diabético**	0,304[a]	4,319	0,000	0,328	0,902	1,109	0,902
	Sexo	0,011[a]	0,161	0,872	0,013	1,000	1,000	1,000
	Idade	0,240[a]	2,988	0,003	0,233	0,735	1,360	0,735
	Fumador	0,045[a]	0,637	0,525	0,051	1,000	1,000	1,000
	IP	0,175[a]	2,327	0,021	0,184	0,851	1,175	0,851
2	**Sexo**	0,012[b]	0,173	0,863	0,014	1,000	1,000	0,902
	Idade	0,244[b]	3,217	0,002	0,251	0,735	1,360	0,679
	Fumador	0,072[b]	1,067	0,288	0,086	0,992	1,008	0,895
	IP	0,162[b]	2,260	0,025	0,179	0,850	1,177	0,786
3	**Sexo**	0,004[c]	0,068	0,946	0,006	0,999	1,001	0,678
	Fumador	0,127[c]	1,909	0,058	0,153	0,939	1,065	0,663
	IP	0,155[c]	2,232	,027	0,178	0,849	1,178	0,617
4	**Sexo**	-0,001[d]	-0,018	,986	-0,001	0,997	1,003	0,616
	Fumador	0,113[d]	1,718	0,088	0,138	0,930	1,076	0,608

a. Predictors in the Model: (Constant), Escolaridade

b. Predictors in the Model: (Constant), Escolaridade, Diabético

c. Predictors in the Model: (Constant), Escolaridade, Diabético, Idade

d. Predictors in the Model: (Constant), Escolaridade, Diabético, Idade, IP

Da análise da tabela 23 verificamos que a proporção de variância atribuível à colinearidade caraterizada por cada valor próprio associada a cada coeficiente (*Variance proportions*) é ainda um indicador da existência ou não de problemas na estimação dos parâmetros. Um valor elevado de *k* (*Condition Index*) associado a uma proporção de variância (β_1 elevada (>50%) – tabela 22) revela uma situação problemática por causa da multicolinearidade na estimação dos parâmetros.

Tabela 23 : Diagnóstico de colinearidade (Variável Dependente: NA)

Modelo	Dimensão	Valores próprios	Variance Proportions					
			Condition Index	(Constante)	Escolaridade		Diabético	Idade	IP
1	1	1,853		1,000	0,07	0,07			
	2	0,147		3,545	0,93	0,93			
2	1	2,349		1,000	0,03	0,03	0,05		
	2	0,550		2,066	0,01	0,15	0,56		
	3	0,101		4,827	0,97	0,82	0,38		
3	1	3,219		1,000	0,00	0,01	0,03	0,01	
	2	0,550		2,418	0,00	0,11	0,56	0,00	
	3	0,209		3,920	0,01	0,33	0,36	0,12	
	4	0,021		12,289	0,99	0,54	0,05	0,88	
4	1	4,145		1,000	0,00	0,01	0,02	0,00	0,00
	2	0,551		2,742	0,00	0,10	0,58	0,00	0,00
	3	0,237		4,185	0,00	0,33	0,38	0,06	0,02
	4	0,053		8,812	0,00	0,03	0,01	0,52	0,51
	5	0,014		17,261	1,00	0,54	0,02	0,42	0,47

Para avaliarmos a qualidade do modelo podemos comparar a variação de Nível de Aderência que é explicada pelo modelo, com a variação do Nível de Aderência que não é explicada pelo modelo e o modelo será tanto melhor quanto maior for este quociente (r^2). Pela leitura da tabela 19 o melhor modelo é que inclui todas as variáveis e se apresenta como modelo 4 (r^2=0,371).

Tabela 24 : Estatísticas Residuais (Variáveis dependente: NA)

	Mínimo	Máximo	Média	Desvio Padrão	N
Predicted Value	-2,273	21,910	12,095	5,753	158
Residual	-16,547	23,971	0,000	7,293	158
Std. Predicted Value	-2,576	1,706	0,000	1,000	158
Std. Residual	-2,283	3,224	0,000	0,981	158

5.5. Ajustamento do Modelo de RLM com recurso ao R

1. Construir um modelo de Regressão linear múltipla

VI: Na

VD: Diabetes, Idade, Sexo, Escolaridade, Fumador e índice de placa

Rotinas:

```
#Determinar os coef da reta regressão método min quadrados
modelo=lm(Na~diabetes+Idade+Sexo+Escol+Fumador+Ip)
 #imprimir
 modelo
```

Resultados:

```
>
> lm(formula = Na~diabetes+Idade+Sexo+Escol+Fumador+Ip)

Call:
lm(formula = Na ~ diabetes + Idade + Sexo + Escol + Fumador +
    Ip)

Coefficients:
(Intercept)     diabetes        Idade         Sexo        Escol      Fumador
    0.40553      3.99244      0.17774     -0.29207     -0.43564      2.44080
         Ip
    0.04563

>
```

Interpretação dos resultados:

O modelo encontrado no *output* do R é idêntico ao que se obteve recorrendo ao SPSS, logo as conclusões a tirar são as mesmas.

Assim: $\hat{Y}_i = 0{,}406 + 3{,}992 \cdot X_{1i} - 0{,}292 X_{3i} + 0{,}178 X_{2i} - 0{,}436 X_{4i} + 2{,}441 X_{5i} + 0{,}046 X_{6i}$

onde Y_i o valor do Nível de Aderência e X_{1i} o valor do *Status Diabético*, X_{3i} é o sexo, X_{2i} é a Idade, X_{4i} e o valor da Escolaridade, X_{5i} é ser Fumador e X_{6i} é o Índice de Placa.

6. Regressão logística

6.1. Introdução e Estratificação dos dados

Procedeu-se à estratificação dos dados conforme tabela 5.

O valor de corte (4 mm) para a variável dependente (Nível de Aderência) foi feito com recurso à metodologia adotada em Medicina Dentária (PEREIRA, J. (2007)), que consiste num gráfico de percentagem acumulada do Nível de Aderência em que o ponto de inflexão das curvas relativas aos diabéticos e não diabéticos vai corresponder ao valor de corte, neste caso será de aproximadamente 4 mm, conforme figura 34.

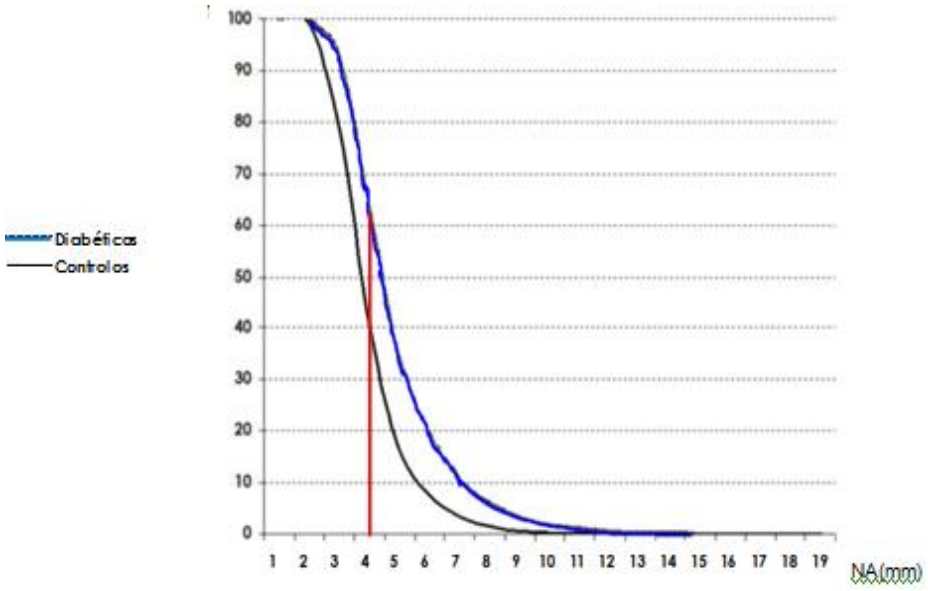


Figura 34: Periodontal *fingerprint*

Em análise estatística, o valor 4 mm seria determinado com recurso à curva ROC, que representa a sensibilidade em função da proporção de falsos positivos (1 – Especificidade) para um conjunto de valores de "*cutoff point*". Por outro lado, as curvas ROC permitem quantificar a exatidão de um teste diagnóstico, já que esta é proporcional à área sob a curva ROC, isto é, tanto maior quanto mais a curva se aproxima do canto superior esquerdo do diagrama. Sabendo isto, a curva será útil, também, na comparação de testes diagnósticos, tendo o teste uma exatidão tanto maior quanto maior for a área sob a curva ROC.

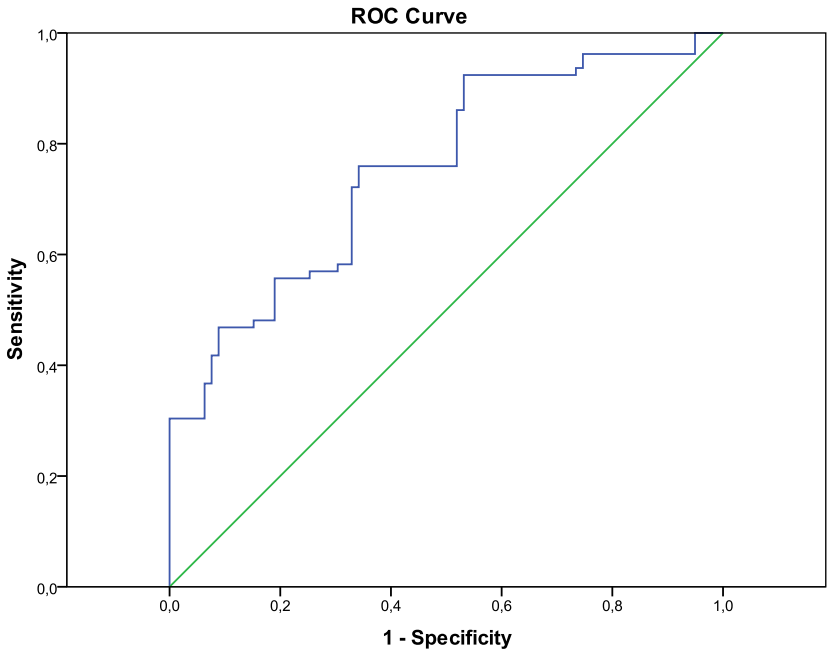


Figura 35: Curva ROC

Tabela 25 : Cálculo da área sob a curva ROC

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,755	,038	,000	,681	,829

a.Under the nonparametric assumption b.Null hypothesis: true area = 0.5

A área sob a curva representa a probabilidade de que o resultado do ensaio de um caso positivo escolhido aleatoriamente irá exceder o resultado de um processo escolhido aleatoriamente negativo.

O conhecimento da área sob a curva possibilita quantificar a exatidão de um teste diagnóstico (proporcional à área sob a curva), além de possibilitar a comparação de testes diagnósticos. A área sob a curva ROC constitui um dos índices mais usados para sumariar a qualidade da curva.

A área sob a curva ROC é uma medida do desempenho de um teste (índice de exatidão do teste). Um teste totalmente incapaz de discriminar indivíduos doentes e não doentes teria uma área sob a curva de 0,5 (seria a hipótese nula). Acima de 0,70 é considerado desempenho satisfatório. Por observação da tabela 25 podemos concluir que no nosso exemplo a área é de 0,755 (IC a 95% de 0,681 a 0,829), logo a significância é inferior a 0,05, o que significa que a utilização do ensaio é melhor do que a probabilidade do acaso.

Tabela 26 : Cálculo da área sob a curva ROC - **Coordenadas da Curva**

Test Result Variable(s):Predicted probability

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity	Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
,0000000	1,000	1,000	,0987914	,671	,291
,0120818	1,000	,962	,1002904	,671	,278
,0151700	1,000	,873	,1017453	,658	,278
,0183215	1,000	,823	,1030482	,658	,266
,0189955	1,000	,810	,1041776	,658	,253
,0197070	1,000	,722	,1081524	,658	,241
,0227586	1,000	,709	,1121348	,633	,241
,0258442	1,000	,696	,1136288	,620	,241
,0265755	,949	,696	,1158933	,582	,241
,0271112	,937	,696	,1212101	,544	,241
,0281013	,937	,671	,1258524	,532	,241
,0314572	,937	,633	,1281231	,506	,241
,0362993	,924	,633	,1313628	,481	,241
,0386131	,924	,620	,1340601	,481	,228
,0396809	,924	,608	,1384628	,481	,215
,0416825	,924	,582	,1418235	,481	,152
,0431778	,911	,582	,1435435	,481	,139
,0447365	,911	,570	,1470820	,468	,139
,0462146	,911	,557	,1496545	,468	,127
,0468769	,911	,544	,1507594	,468	,114
,0474568	,911	,532	,1604783	,468	,076
,0520042	,873	,532	,1706842	,380	,076
,0566896	,861	,532	,1728060	,316	,076
,0577958	,848	,532	,1773360	,291	,076
,0593667	,848	,519	,1823956	,278	,076
,0605891	,823	,519	,1885982	,266	,076
,0614011	,810	,519	,1966153	,266	,063
,0639968	,810	,494	,2025346	,253	,063
,0662411	,810	,481	,2095441	,253	,051
,0692917	,810	,443	,2213457	,253	,038
,0738447	,797	,443	,2351092	,241	,038
,0806669	,772	,443	,2417898	,101	,038
,0860077	,747	,443	,2482103	,089	,038
,0873823	,747	,430	,2551515	,076	,038
,0891976	,734	,430	,2611699	,063	,038
,0897044	,722	,430	,2723517	,051	,038
,0903673	,709	,430	,2792284	,051	,025
,0915395	,696	,430	,2868084	,051	,013
,0932573	,696	,418	,3044728	,051	,000
,0944404	,684	,418	,3205961	,038	,000
,0951521	,671	,418	,3864775	,013	,000
,0963793	,671	,392	1,0000000	,000	,000
,0974435	,671	,304			

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

O nosso exemplo encontra-se ilustrado na figura 35. A tabela 26 relata a sensibilidade e 1– especificidade para cada ponto de corte (no nosso caso será 4 mm) possível para a classificação positiva. O ponto de corte 0 é equivalente a assumir que todos são positivos para a doença periodontal. O ponto de corte 1 é equivalente a assumir que todos são negativos para a doença periodontal. Note-se que ambos os extremos são insatisfatórios, pelo que propomo-nos escolher um corte que permita equilibrar as necessidades de sensibilidade e especificidade, vamos analisar o que se passa no 0,04 pois na figura 34 (periodontal *fingerprint*) é no valor 4 que encontramos uma maior diferença no Nível de Aderência entre diabéticos e não diabéticos (que destacámos a vermelho).

Por exemplo, considere-se o ponto de corte 0,04 (correspondente a NA=4). Calculemos as médias das coordenadas dos pontos nestas condições:

$$(0{,}924 + 5x\ 0{,}911)\ /5 = 0{,}913$$

$$(2x0{,}582 + 0{,}570 + 0{,}557 + 0{,}544 + 0{,}532)\ /5 = 0{,}561$$

Usando este ponto de corte vamos ter uma sensibilidade de 0,913 e 1 – especificidade de 0,561. Assim, aproximadamente 91,3% de todas as amostras do NA > 4 seriam corretamente identificadas como tal, e 56,1% de todas as amostras do NA ≤ 4 poderiam ser incorretamente identificadas como positivas.

A escolha do ponto de corte será decidida pela necessidade de aumentar a sensibilidade ou a especificidade e vai depender do investigador. Devemos notar que os valores da tabela representam as melhores orientações para os quais devemos considerar os pontos de corte. Esta tabela não inclui as estimativas de erro, portanto, não há garantia da exatidão da sensibilidade ou especificidade para um dado ponto de corte na tabela.

O uso da Curva ROC possibilita-nos avaliar a acurácia deste teste.

A área sob a curva (figura 35) mostrou que a utilização do teste é melhor do que a probabilidade de ocorrência, mas, além disto, as coordenadas da curva (destacadas na tabela 26) são úteis porque fornecem algumas orientações para determinar qual o melhor ponto de corte para a determinação os resultados do teste positivo e negativo.

A probabilidade de ocorrer NA ≥ 4 mm em 25% dos casos refere-se à extensão da doença periodontal. Enquanto que o número 4 mm nos dá a gravidade da doença, o valor 25% refere-se à extensão da doença (superfície afetada).

Após se definir o conjunto de variáveis independentes a serem incluídas no modelo, precisamos de identificar as variáveis mais importantes para explicar a probabilidade de

sucesso. Assim, interessa-nos identificar, entre todas as variáveis independentes (Idade, Sexo, Diabetes, Escolaridade, Fumador e Índice de Placa) o conjunto de variáveis que melhor contribui para a explicação da variabilidade.

6.2. Codificação de fatores

Na regressão logística utiliza-se uma codificação de fatores das variáveis independentes em que se determina um fator de referência.

Os efeitos dos restantes fatores da respetiva variável independente estarão em relação ao fator de referência. Por exemplo, a variável sexo está codificada segundo o seguinte critério indicador: 0 = Feminino e 1 = Masculino, sendo a categoria de referência a que tem o código zero, conforme consta na tabela 28.

Tabela 27 : Codificação da variável dependente

Valor Original	Código
Inferior a 50%	0
Igual ou Superior a 50%	1

Tabela 28 : Codificação das variáveis independentes

		Frequência	Código do Parâmetro	
			(1)	(2)
Escolaridade	Menos que 5	72	1	0
	Entre 5 e 9	35	0	1
	Mais que 9	51	0	0
Idade_classe	<= 50	63	1	
	>50	95	0	

		Frequência	Código do Parâmetro	
			(1)	(2)
Diabético	Não diabético	79	1	0
	Diabético tipo 1	20	0	1
	Diabético tipo2	59	0	0
Fumador	Não	125	1	
	Sim	33	0	
Sexo	Feminino	82	1	
	Masculino	76	0	

O processo de seleção de variáveis pode ser feito de várias formas.

A seleção de *Forward* baseada no teste de Wald começa por considerar um modelo apenas com a constante (tabela 29).

A estatística de Wald é usado para avaliar a significância dos coeficientes da regressão logística. As hipóteses são:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, i.e., o modelo não é estatisticamente significativo

H_1: $\exists\, \beta_{i \neq 0,\ (i = 1,\ldots,K)}$, i.e., o modelo é estatisticamente significativo

No nosso exemplo, como o *p-value* = 0,000, não se rejeita a hipótese nula. Concluindo-se que o modelo linear assim obtido não é estatisticamente significativo.

Tabela 29 : Cálculo da Estatística de Wald

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-2,051	0,250	67,110	1	**0,000**	0,129

Tabela 30 : Estudo das variáveis não incluídas

			Score	g.l	Sig.
Step 0	Variables	Idade_classe	0,595	1	0,441
		Sexo(1)	0,109	1	0,741
		Diabetes	6,147	2	**0,046**
		Diabetes(1)	1,003	1	0,317
		Diabetes(2)	2,944	1	0,086
		Escol_classes	4,769	1	**0,029**
		Fumador(1)	1,905	1	0,168
		IP_classe	1,377	1	0,241
	Overall Statistics		11,681	7	0,112

A tabela 30, apresenta as estatísticas *Score* das variáveis não incluídas no modelo e os seus *p-value*. Para $\alpha = 0{,}05$, as variáveis escolaridade e diabetes são estatisticamente significativas.

Método *Stepwise (Forward)*

Tabela 31 : Quadro inicial das iterações

Iteração			Coeficientes	
		-2 Log likelihood	Constante	Escol_classes
Step 1	1	113,985	-1,325	-0,253
	2	107,444	-1,563	-0,527
	3	106,926	-1,573	-0,692
	4	106,917	-1,571	-0,720
	5	106,917	-1,571	-0,720

a. Method: Forward Stepwise (Wald)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 112,067

d. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

O valor de -2 Log Likelihood é um indicador da qualidade do ajustamento do modelo aos dados. Quanto maior for o seu valor, pior é o ajustamento.

A tabela 32 mostra que o modelo encontrado tem uma má qualidade de ajustamento do modelo aos dados, pois os seus valores são muito elevados (>106).

Tabela 32 : Teste do rácio das verosimilhanças entre modelos

		Chi-square	df	Sig.
Step 1	Step	5,150	1	0,023
	Block	5,150	1	0,023
	Model	5,150	1	0,023

No nosso exemplo, sendo $\chi^2(1) = 5{,}150$ e $p > 0{,}001$, nada podemos concluir sobre o valor preditivo da nossa variável dependente sobre o Nível de Aderência.

Tabela 33 : Qualidade do ajustamento do modelo

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	106,917[a]	0,032	0,063

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Também é importante detetarmos pontos influentes, ou seja, pontos que afetam de forma significativa o ajuste do modelo.

Recorremos à estatística do teste $\boldsymbol{G^2}$ para testar a significância do modelo de regressão logística.

Os valores do $\boldsymbol{pseudo - R^2}$ podem ser determinados por:

$G^2 = 5{,}150 \quad -2LLC = 106{,}917, \quad logo \ -2LLo = 5{,}150 + 106{,}917 = 112{,}067$

$Pseudo - R^2 \ de \ Cox\& \ Snell \ = R^2_{CS} = 0{,}032$

$Pseudo - R^2 \ de \ Nagelkerke \ = R^2_N \ = 0{,}063$

$$Pseudo - R^2 \ de \ McFadden = R^2_{MF} = 1 - \{-2Ln(L_C)/-2Ln(L_o)\} = 1 - \frac{106{,}917}{112{,}067} = 0{,}046$$

Estes valores dos pseudo-r^2 revelam o modelo em que a variável dependente é a Escolaridade e terá uma qualidade fraca (tal modelo será adiante escrito) pois o seu valor varia entre 0 e 1 e o valor 1 indica o melhor ajuste do modelo.

6.3. Qualidade do ajuste do modelo

Para avaliar a qualidade do ajuste recorremos ao teste de Hosmer and Lemeshow.

Tabela 34 : Teste de Hosmer and Lemeshow

Step	Chi-square	df	Sig.
1	0,293	1	0,588

Tabela 35 : Tabela de contingência do teste de Hosmer and Lemeshow

		Nível de aderência >4 = Inferior a 50%		Nível de aderência >4 = Igual ou Superior a 50%		Total
		Observed	Expected	Observed	Expected	
Step 1	1	49	48,608	2	2,392	51
	2	31	31,785	4	3,215	35
	3	60	59,608	12	12,392	72

As tabelas 34 e 35 apresentam o teste de ajustamento de Hosmer and Lemeshow que compara, após divisão da amostra em aproximadamente 10 classes iguais, a correspondência entre os números reais e os previstos em cada classe, utilizando a estatística Qui-quadrado. Um bom ajuste de modelo é indicado por um valor Qui-quadrado não-significante (também se pode recorrer aos valores esperados e observados para calcular a estatística do teste). Como *p-value* é 0,588, podemos concluir que os valores estimados pelo modelo estão próximos dos valores esperados, ou seja, o modelo ajusta-se aos dados.

Tabela 36 : Classificação observada e prevista no modelo ajustado

Observado			Predito		
			Nível de aderência >4		Percentagem
			< 50%	≥ 50%	
Step 1	Nível de aderência >4	< 50%	140	0	100,0
		≥ 50%	18	0	0,0
	Overall Percentage				**88,6**

a. O valor de corte é 0,500

De acordo com os dados da tabela 36, apenas 18 indivíduos positivos para a condição estão classificados como negativos (falso-negativos); os restantes encontram-se bem classificados. Podemos assim concluir que a sensibilidade do modelo é de 100% (ou seja, classifica corretamente os doentes) e a especificidade é 0% (que é a probabilidade condicionada de prever um diagnóstico negativo sabendo que o indivíduo não tem doença). Globalmente o modelo classifica corretamente 88,6% dos indivíduos que apresentam NA <4.

O estimador do *logit* e seu intervalo de confiança fornece o estimador dos valores ajustados. O intervalo de confiança dos valores ajustados é apresentado na tabela 37.

Tabela 37 : Informações sobre variáveis independentes no modelo completo

		β	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Inferior	Superior
Step 1[a]	Escol_classes	-0,720	0,345	4,370	1	0,037	0,487	0,248	0,956
	Constant	-1,571	0,304	26,675	1	0,000	0,208		

a. Variable(s) entered on step 1: Escol_classes.

$$Ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \beta_0 + \beta_1 \mathrm{X}_{4\mathrm{i}} \,,$$

ou seja, neste caso a equação do modelo de regressão logística é dada por

$$Ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1{,}571 - 0{,}720\, X_{4i}$$

onde X_{4i} e o valor da Escolaridade

Para $\alpha = 0{,}05$ a estatística de Wald do quadrado mostra que só os indivíduos com mais escolaridade ($sig = 0{,}037$) são significativos para o modelo, melhorando significativamente a sua previsibilidade.

Nesta abordagem o $or = e^{\beta} = 0{,}487$ $IC =]0{,}248; 0{,}956[$ $l$ogo o modelo é significativo.

6.4. Análise dos resíduos

A análise dos resíduos permite identificar *outliers* e casos influentes na estimação do modelo e as medidas usadas são basicamente as mesmas da regressão linear, apenas é necessário entrar em consideração com as variâncias das observações, pois na regressão logística elas não são constantes.

Iniciámos o nosso estudo pela interpretação gráfica:

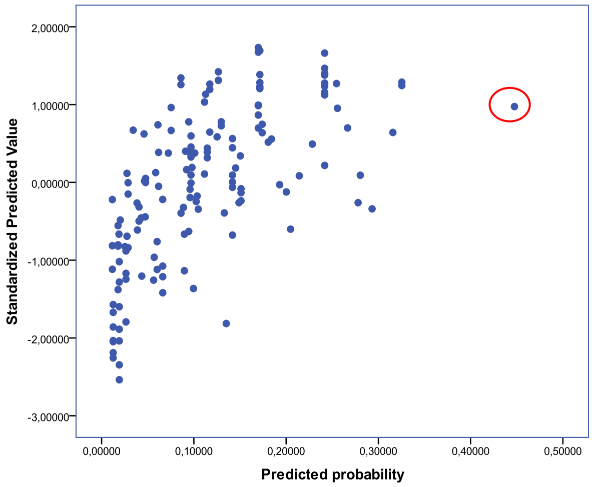


Figura 36: Análise de resíduos

Do gráfico representado na figura 36 podemos afirmar que este modelo apresenta *outliers*.

```
                Step number: 1

                Observed Groups and Predicted Probabilities

         80 +                                                                                +
            I                                                                                I
            I                  1                                                             I
 F          I                  1                                                             I
 R       60 +                  0                                                             +
 E          I                  0                                                             I
 Q          I    0             0                                                             I
 U          I    0             0                                                             I
 E       40 +    0             0                                                             +
 N          I    0    1        0                                                             I
 C          I    0    0        0                                                             I
 Y          I    0    0        0                                                             I
         20 +    0    0        0                                                             +
            I    0    0        0                                                             I
            I    0    0        0                                                             I
            I    0    0        0                                                             I
 Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
   Prob:   0        ,1        ,2        ,3        ,4        ,5        ,6        ,7        ,8        ,9         1
   Group:  0000000000000000000000000000000000000000000000000011111111111111111111111111111111111111111111111111
```

Predicted Probability is of Membership for Igual ou Superior a 50%. The Cut Value is .50. Symbols: 0 - Inferior a 50%; 1 - Igual ou Superior a 50%; Each Symbol Represents 5 Cases.

Figura 37 : Grupos observados e Probabilidades Previstas

Neste tipo de gráfico a distribuição desejável é em forma de U. Quando a distribuição é em forma de sino (como no nosso caso), o modelo classifica mal as observações cujas probabilidades se concentram em torno de 0,5.

Tabela 38 : Quadro de identificação dos *outliers*

Caso	Seleção Status[a]	Observada	Predita	Grupo Predito	Variável Temporária	
		Nível de aderência >4			Resid	Z Resid
48	S	1**	,047	0	,953	4,508
59	S	1**	,092	0	,908	3,144
63	S	1**	,092	0	,908	3,144
72	S	1**	,092	0	,908	3,144
109	S	1**	,092	0	,908	3,144
126	S	1**	,047	0	,953	4,508

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Pela leitura da tabela 38 concluímos que as observações 48, 59, 63, 72, 109 e 126 apresentam valores elevados de resíduos estandardizados (ZRESID), indicando que estes casos foram mal classificados no modelo (ou seja encontram-se mal classificados pelo modelo 3,8% – seis casos em 158 – das observações).

De forma análoga vamos analisar o modelo gerado pelo método ENTER, cujas tabelas do SPSS se encontram no anexo III.

Método Enter

Como 2LL é um indicador da qualidade do ajustamento do modelo aos dados, e o seu valor é muito elevado (102,408), podemos afirmar que o ajustamento não é bom.

Com o valor de $\chi^2(1) = 4{,}331$; e $p > 0{,}001$, nada podemos concluir sobre o valor preditivo da nossa variável dependente relativamente ao Nível de Aderência.

Qualidade do ajuste do modelo

O teste de ajustamento de Hosmer and Lemeshow permite concluir que os valores estimados pelo modelo estão próximos dos valores esperados, ou seja, o modelo ajusta-se aos dados uma vez que *o p-value* é 0,826.

Este modelo, relativamente à significância, não é significativo pois os IC dos *odds ratio* contêm sempre o valor 1.

Graficamente, atendendo aos valores que se encontram no anexo II, vem:

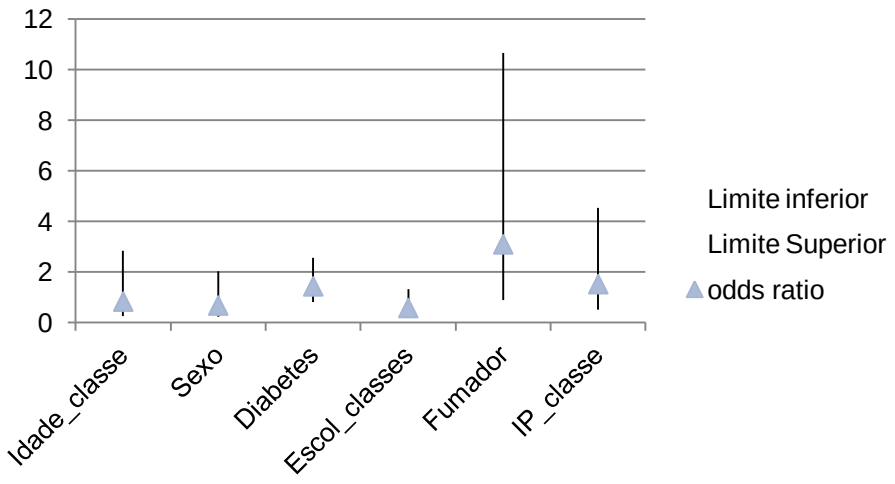


Figura 38: *Odds ratio* e respetivos Intervalos de Confiança

A equação do modelo de regressão logística é dado por:

$$Ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2{,}375 - 0{,}17\,X_{3i} - 0{,}376\,X_{2i} + 0{,}368\,X_{1i} - 0{,}531\,X_{4i} + 1{,}134\,X_{5i} + 0{,}428\,X_{6i}$$

Onde X_{3i} é a Idade, X_{2i} é o Sexo, X_{1i} é o valor do *Status Diabético*, X_{4i} é o valor da Escolaridade, X_{5i} é ser Fumador e X_{6i} é o Índice de Placa.

Para $\alpha = 0{,}05$ a estatística de Wald mostra que nenhuma variável é significativa para o modelo quando pretendemos prever o valor do Nível de Aderência.

Graficamente, como a distribuição é em forma de sino (Figura no anexo II), o modelo classifica mal as observações cujas probabilidades se concentram em torno de 0,5.

Analisando os *outliers,* concluímos que para além das observações 48, 59, 63, 72, 109 e 126 que se encontravam referenciadas no modelo anterior, ainda surgem mais duas (45 e 149) que apresentam valores elevados de resíduos estandardizados (ZRESID), indicando que estes casos foram mal classificados no modelo.

7. Conclusão geral da análise estatística e recomendações aos especialistas

Quando utilizamos a regressão logística devemos primeiramente determinar o modelo que melhor se ajuste aos dados em análise, com o intuito de se obter um modelo moderado e biologicamente razoável, que permita descrever a relação entre a variável resultado e um conjunto de variáveis independentes.

Algumas avaliações devem ser realizadas para se ter alguma ideia da eficácia e adequação do modelo.

- De entre as técnicas utilizadas para avaliar a eficácia do modelo, o coeficiente de correlação é uma primeira possibilidade.
- Outra medida de adequação é o coeficiente de determinação do modelo.
- A determinação do "melhor" modelo, ou do modelo que melhor se ajusta aos dados, está relacionada com a estimativa dos parâmetros que tornem os resíduos tão próximos de zero quanto possível.
- Por fim deve-se testar a significância estatística dos parâmetros do modelo.

No nosso estudo, comparando os diabéticos com os não diabéticos, verificou-se que só não existem diferenças quanto ao Índice de Placa. Quanto às restantes variáveis (Nível de Aderência, Hemorragia Pós Sondagem e Recessão Gengival), os dois grupos apresentam diferenças significativas, independentemente do método utilizado na avaliação.

Da nossa análise, podemos concluir que o teste de significância da equação de Regressão Linear Múltipla indicou que todos os modelos construídos podem ser considerados significativos para um nível de significância de 5%, o que indica homocedasticidade, e consequentemente os modelos de regressão são válidos. Em suma, **o modelo é altamente significativo**. Como o *p-value* encontrado foi inferior a 0,05, podemos assegurar que o modelo de regressão considerado é melhor que a média para predizer os valores do NA.

A interpretação dos parâmetros de um modelo de regressão logística é obtida comparando a probabilidade de sucesso com a probabilidade de fracasso, usando a função *odds ratio – or*, 38,4% da percentagem de variância é explicada pelo modelo.

A Probabilidade de um individuo com NA >4 ser portador de doença periodontal é 39 vezes superior à probabilidade de se obter um individuo com NA >4 que não tenha doença periodontal. Ou seja quando o fator NA >4 está presente é mais provável o individuo já ter doença do que o contrário.

Quando se aplicou a regressão logística, a Escolaridade e o *Status Diabético* revelaram ser as variáveis explanatórias mais importantes. Quando foi utilizado para a seleção das variáveis o **Método *Stepwise (Forward),*** de acordo com os resultados apresentados, podemos concluir que o modelo não se apresentou satisfatório, logo apresentou com um fraco poder preditivo. Após ajustamento, o modelo classifica corretamente 88,6% dos indivíduos que apresentam NA <4.

Quando recorremos à seleção de variáveis pelo método *Enter* (saídas do SPSS no anexo II), de acordo com os resultados apresentados, podemos concluir que o modelo ajustado não se apresentou satisfatório, logo com um fraco poder preditivo. Neste modelo foram incluídas todas as variáveis independentes, a estatística de Wald mostra que nenhuma variável é significativa para o modelo quando pretendemos prever o valor do NA. Um bom ajuste de modelo é indicado por um valor Qui-quadrado não-significante. Na tabela (Hosmer and Lemeshow Test – anexo II), o alto valor de significância (*p-value*) de 0,826, considerada a distribuição Qui-quadrado com 8 graus de liberdade, sugere que falha em se rejeitar a hipótese nula de que não há diferença entre os valores reais e os previstos. Em resumo, o modelo estima o ajuste dos dados num nível aceitável.

Em síntese:

<table>
<tr><td>Grau de relação entre as variáveis</td><td>Análise Multivariada</td><td rowspan="2">Com recurso à correlação foi comparada a situação periodontal entre os doentes diabéticos e não diabéticos, emparelhados para o sexo e idade, os diabéticos evidenciam pior saúde periodontal.
O NA dos não diabéticos e dos diabéticos estão diretamente correlacionados com Idade, HDL, Triglicerídeos e Índice de Placa.
Com base na Análise Multivariada podemos dizer que as variáveis que melhor permitem prever o valor do Nível de Aderência são a Idade, o HDL, Triglicerídeos e Índice de Placa.</td></tr>
<tr><td></td><td>Análise Multivariada</td></tr>
<tr><td rowspan="2">Significância das diferenças entre grupos</td><td>t-Teste</td><td rowspan="2">A comparação dos grupos diabéticos e não diabéticos foi feita pelo teste de M-W tendo as diferenças das suas medianas sido estatisticamente significativas para o NA.
O modelo de regressão múltipla é altamente significativo e a significância da equação de Regressão Linear Múltipla indicou que todos os modelos construídos podem ser considerados significativos para um nível de significância de 5%. Neste modelo 61,9% da variabilidade encontrada para o NA pode ser explicada pelas variáveis independentes.</td></tr>
<tr><td>ANOVA</td></tr>
<tr><td rowspan="3">Predição da pertença a um grupo</td><td>Análise Discriminante</td><td rowspan="3">O nosso modelo permite-nos afirmar que a probabilidade de um individuo apresentar doença periodontal é maior nos diabéticos e em indivíduos com menor grau de escolaridade.</td></tr>
<tr><td>Regressão Múltipla</td></tr>
<tr><td>Regressão logística</td></tr>
</table>

Neste trabalho propusemo-nos a avaliar a utilização da análise de regressão a um caso concreto, aplicando-a aos dados de um estudo sobre doença periodontal.

Convém referir que, de uma maneira geral, a análise de regressão pode ser utilizada com vários objetivos dentre os quais destacamos a Descrição, a Predição, o Controlo e a Estimação. Na prática, a análise de Regressão é utilizada para atingir simultaneamente mais do que um dos objetivos citados.

A primeira parte deste trabalho permitiu adquirir e ampliar conhecimentos relativos à problemática da regressão, clássica e logística, com particular destaques para os aspetos teóricos e representações gráficas, assim como um aprofundar de alguns conceitos básicos de doença periodontal. Na segunda parte, foi efetuada a análise estatística dos dados e sua discussão, de acordo com diferentes metodologias. Em função dos resultados obtidos e tendo em vista a sua aplicação na área da medicina, deixam-se algumas considerações e sugestões:

- É crucial o ajuste da escolha de variáveis e métodos em função do contexto biológico, ouvindo e estudando os fenómenos alvos do estudo estatístico de modo que todas as opções a fazer sejam devidamente fundamentadas tanto na vertente estatística como na vertente biológica.
- É importante sensibilizar os utilizadores da Estatística na área da Saúde para a importância da correta utilização dos métodos, não só validando pressupostos mas também na seleção dos modelos (PAPOILA, A. (2012)).
- Em trabalhos futuros ambicionamos explorar as componentes estatísticas da Análise Fatorial e Análise em Componentes Principais e proceder à comparação de resultados.
- A realização deste trabalho tornou-se uma experiência gratificante, apesar de ter exigido grande disponibilidade e esforço. Espera-se que este estudo possa, de alguma forma, contribuir, ainda que de forma modesta, para o desenvolvimento de alterações a nível das estratégias adotadas e da clareza da informação veiculada.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU M. N. S., SIQUEIRA A. L., CAIAFFAI W.T. (2009): *Regressão logística ordinal em estudos epidemiológicos***,** Rev Saúde Pública;43(1): 183-94.

ALLISON, PAUL D. (1999): Multiple Regression – A primer, Pine Forge Press.

BERTIN, J. (1973): *Sémiologie graphique*.(2.ª ed.) Gauthier-Villars. Paris.

BETZ, N.E. (1987): *Use of Discriminant Analyses in Counseling Psychology Research*. Journal of Counseling Psychology, 34 (4),393-403.

BRAGA, A. (2000): *Curva ROC: Aspectos fundamentais e Avaliação.* Braga: Tese de Doutoramento, Universidade do Minho.

CHALONER K., LARNTZ K. (1989): *Optimal Bayesian Design Applied to Logistic Regression Experiments,* Journal of Statistical Planning and Inference 21 191-208

CLEVELAND, WILLIAM S.; MCGILL, ROBERT (1987): *"Graphical perception: The visual decoding of quantitative information on graphical displays of data",* Journal of the Royal Statistical Society, 150,192-229.

COLIN R.B (2004): *Bioestatística usando R - apostila para biólogos*. Bragança.

COELHO-BARROS, E. A., SIMÕES P. A., ACHCAR J. A., MARTINEZ E. Z., SHIMANO A. C., (2008): *Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos*, Revista Colombiana de Estadística, volume 31, nº 1: 111- 129

DAGNELIE, P. (1973): *Estatística. Teoria e métodos* (2º Volume). Publicações Europa-América.

DUARTE, R. (2002): *Autovigilância e Métodos de Avaliação do controlo Metabólico do diabético*. Diabetologia Clínica In LIDEL – Ed Técnicas Lda,. Lisboa 57-77.

FARAWAY, J. (2002): *Practical Regression and Anova using R - http://csyue.nccu.edu.tw*

LAUREANO, M.; BOTELHO, M. (2010): *SPSS o meu manual de consulta rápida.* (1ª Edição) Edições Sílabo, Lda. Lisboa.

LAMSTER I., LALLA E., BORGNAKKE V., TAYLOR W. (2008): The Relationship Between Oral Health and Diabetes Mellitus, JADA October 2008 vol. 139 no. supl 519S-24S

HEDEKER, D. (2003): *A mixed-efects multinomial logistic regression model,* Statistics in Medicine, Statist. Med.; 22:1433–1446.

HOFFMAN, H. (2006): *Visualizing Simple Logistic Regression Models using Mosaic Plots.*

HOSMER, D. J., & LEMESHOW, S. (1989): *Applied Logistic Regression.* Copyright by John Wiley & Sons, Inc.

HOSMER D W, LEMESHOW S. (2000): *Applied Logistic Regression*, 2nd ed. New York; Wiley.

MADSEN L., FANG Y., *Joint Regression Analysis for Discrete Longitudinal Data*, Biometrics.

MARGOTTO, P. (s.d.): *Curva ROC: Como fazer e Interpretar no SPSS*. Curso de Medicina da Escola Superior de Ciências da Saúde (ESCS/CES/DF).

MARGOTTO, P: (2002): Entendendo Bioestatística Básica. Boletim Informativo Pediátrico (BIP)-Brasília, N0 65, p. 6.

MAROCO, J. (2007): *Análise Estatística com utilização do SPSS*. 3ª Ed., Edições Sílabo.

MARTINS, P. S. (2008). *Análise estatística de performance de um conjunto de testes auditivos.* Tese de Mestrado, Universidade de Aveiro.

MILTON J. S.; TSOKOS J.O. (1983): *Statistical Methods in the Biological and Health Sciences*. McGraw-Hill Book Company.

MONTGOMERY. D. C.: (2001): *Design and Analysis of Experiments*, 5th Ed, John Wiley & Sons.

MORRISON (1984): *Multivariate Statistical Methods*.2nd Edition, International Student Edition.

OLIVEIRA,T. A. (2004): *Estatística Aplicada*, Universidade Aberta.

PEREIRA, J. A. (2007). *Doença Periodontal e diabetes mellitus.* Tese de Doutoramento, Universidade de Porto.

PESTANA H. P., GAGEIRO J. N. (2000). *Análise de dados para ciências sociais. A complementaridade do SPSS*. 2ª ed. Edições Sílabo, Lda.

PRESHAW P. M., ALBA A. L., HERRERA D., JEPSEN S., KONSTANTINIDIS A., MAKRILAKIS K., TAYLOR R. (2012). *Periodontitis and diabetes: a two-way relationship.* Diabetologia. 2012 January; 55(1): 21–31.

RUSH SLOAN (2001): *Logistic Regression: The Standard Method of Analysis*, in Medical Research

SIEGEL, S. (1975) *Estatística Não-paramétrica Para as Ciências do Comportamento*, McGraw-Hill.

SCOT, M. (2002): *Applied Logistic Regression Analysis*, 2nd Edition, Sage Publications.

SGOLASTRA F, SEVERINO M., PIETROPAOLI D., GATTO R., MONACO A. (2012): *Effectiveness of Periodontal Treatment to Improve Metabolic Control in Patients With Chronic Periodontitis and Type 2 Diabetes: A Meta-Analysis of Randomized Clinical Trials*. Journal of Periodontology, October 29

STEEL, R. TORRIE, J. (1986): Bioestatistica Principios y procedimientos, 2thEd. Mc Graw Hill.

STEVEN C. B., HALBERT W., BEATRICE A. GOLOMBC (2001) *Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain*, Journal of Clinical Epidemiology 54 979–985.

TABACHNICK B., FIDELL L. (2007), *Using Multivariate Statistics* (5ª Edição).NY: Pearson Allyn & Baccon.

TURKMAN M. A., SILVA G. L. (2000), *Modelos Lineares Generalizados - da teoria à prática* – Lisboa.

ANEXOS

Anexo I

Tabela das correlações

			PS	NA	Rm	HPSm
I	Não Diabéticos	r	**0,394**	**0,442**	**0,572**	**0,361**
		p	**<0,0001**	**<0,0001**	**<0,0001**	**0,001**
		cd	**0,15**	**0,20**	**0,33**	**0,13**
	Diabéticos	r	**0,278**	**0,514**	**0,607**	0,087
		p	**0,013**	**<0,0001†**	**<0,0001**	0,447
		cd	**0,08**	**0,26**	**0,37**	0,01
	z – dif. de r		*0,81*	*0,57*	*0,33*	***1,79***
GPJ	Não Diabéticos	r	0,083	0,176	0,114	0,126
		p	0,466	0,121	0,319	0,270
		cd	0,01	0,03	0,01	0,02
	Diabéticos	r	-0,057	-0,029	-0,029	0,049
		p	0,616	0,983†	0,797	0,666
		cd	0,00	0,00	0,00	0,00
	z – dif. de r		*0,87*	*1,28*	*0,88*	*0,47*
CT	Não Diabéticos	r	**0,388**	**0,325**	**0,258**	**0,401**
		p	**0,000**	**0,004**	**0,022**	**0,000**
		cd	**0,15**	**0,11**	**0,07**	**0,16**
	Diabéticos	r	0,033	0,187	0,121	0,142
		p	0,776	0,0992†	0,290	0,212
		cd	0,00	0,03	0,01	0,02
	z – dif. de r		***2,32***	*0,91*	*0,88*	***1,74***
HDL	Não Diabéticos	r	**-0,394**	**-0,451**	**-0,341**	**-0,371**
		p	**0,000**	**<0,0001**	**0,002**	**0,001**
		cd	**0,16**	**0,20**	**0,12**	**0,14**
	Diabéticos	r	-0,149	**-0,282**	**-0,269**	-0,034
		p	0,189	**0,0119†**	**0,017**	0,767
		cd	0,02	**0,08**	**0,07**	0,00
	z – dif. de r		*1,638*	*1,212*	*0,490*	***2,195***
LDL	Não Diabéticos	r	**0,453**	**0,408**	**0,265**	**0,490**
		p	**<0,0001**	**0,000**	**0,018**	**<0,0001**
		cd	**0,20**	**0,17**	**0,07**	**0,24**
	Diabéticos	r	0,017	0,102	0,113	0,127
		p	0,882	0,3712†	0,323	0,264
		cd	0,00	0,01	0,01	0,02
	z – dif. de r		***2,90***	***2,04***	*0,98*	***2,52***
T	Não Diabéticos	r	**0,262**	**0,255**	**0,339**	0,171
		p	**0,020**	**0,023**	**0,002**	0,133
		cd	**0,07**	**0,07**	**0,11**	0,03
	Diabéticos	r	**0,267**	**0,280**	0,207	0,218
		p	**0,0174†**	**0,0124†**	0,0674†	0,0533†
		cd	**0,07**	**0,08**	0,04	0,05
	z – dif. de r		*0,03*	*0,17*	*0,88*	*0,31*
IMC	Não Diabéticos	r	**0,353**	**0,342**	**0,400**	**0,419**
		p	**0,001**	**0,002**	**0,000**	**0,000**
		cd	**0,12**	**0,12**	**0,16**	**0,18**
	Diabéticos	r	0,098	0,215	0,214	0,181
		p	0,391	0,06†	0,059	0,110
		cd	0,01	0,05	0,05	0,00
	z – dif. de r		*1,67*	*0,84*	*1,28*	*1,62*
IP†	Não Diabéticos	r	**0,566**	**0,512**	**0,335**	**0,601**
		p	**<0,0001**	**<0,0001**	**0,003**	**<0,0001**
		cd	**0,32**	**0,26**	**0,11**	**0,36**
	Diabéticos	r	**0,447**	**0,298**	0,131	**0,546**
		p	**<0,0001**	**0,010**	0,248	**<0,0001**
		cd	**0,20**	**0,09**	0,02	**0,30**
	z – dif. de r		*0,99*	*1,59*	*1,33*	*0,50*

Spearman -†

Anexo II

Saídas do SPSS nos modelos de regressão

(Stepwise Forward,Enter)

```
LOGISTIC REGRESSION VARIABLES Na_mais_4
 /METHOD=BSTEP(WALD) Idade_classe Sexo Diabetes Escol_classes Fumador IP_classe
 /CONTRAST (Sexo)=Indicator
 /CONTRAST (Fumador)=Indicator
 /CONTRAST (Diabetes)=Indicator
 /SAVE=COOK LEVER DFBETA
 /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Block 1: Method = Backward Stepwise (Wald)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	14,167	7	,048
	Block	14,167	7	,048
	Model	14,167	7	,048
Step 2[a]	Step	-,022	1	,881
	Block	14,145	6	,028
	Model	14,145	6	,028
Step 3[a]	Step	-,414	1	,520
	Block	13,731	5	,017
	Model	13,731	5	,017
Step 4[a]	Step	-,522	1	,470
	Block	13,209	4	,010
	Model	13,209	4	,010
Step 5[a]	Step	-5,950	2	,051
	Block	7,259	2	,027
	Model	7,259	2	,027
Step 6[a]	Step	-2,109	1	,146
	Block	5,150	1	,023
	Model	5,150	1	,023

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	97,899[a]	,086	,169
2	97,922[a]	,086	,169
3	98,336[a]	,083	,164
4	98,857[a]	,080	,158
5	104,808[b]	,045	,088
6	106,917[c]	,032	,063

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

b. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

c. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Classification Table[a]

	Observed		Predicted		
			Nível de aderência >4		
			Inferior a 50%	Igual ou Superior a 50%	Percentage Correct
Step 1	Nível de aderência >4	Inferior a 50%	140	0	100,0
		Igual ou Superior a 50%	18	0	,0
	Overall Percentage				88,6
Step 2	Nível de aderência >4	Inferior a 50%	140	0	100,0
		Igual ou Superior a 50%	18	0	,0
	Overall Percentage				88,6
Step 3	Nível de aderência >4	Inferior a 50%	140	0	100,0
		Igual ou Superior a 50%	18	0	,0
	Overall Percentage				88,6
Step 4	Nível de aderência >4	Inferior a 50%	140	0	100,0
		Igual ou Superior a 50%	18	0	,0
	Overall Percentage				88,6
Step 5	Nível de aderência >4	Inferior a 50%	140	0	100,0
		Igual ou Superior a 50%	18	0	,0
	Overall Percentage				88,6
Step 6	Nível de aderência >4	Inferior a 50%	140	0	100,0
		Igual ou Superior a 50%	18	0	,0
	Overall Percentage				88,6

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Idade_classe	,092	,614	,023	1	,881	1,096
	Sexo(1)	,345	,560	,380	1	,538	1,412
	Diabetes			1,685	2	,431	
	Diabetes(1)	-,732	,564	1,685	1	,194	,481
	Diabetes(2)	-19,565	8753,441	,000	1	,998	,000
	Escol_classes	-,485	,406	1,424	1	,233	,616
	Fumador(1)	-1,144	,632	3,271	1	,071	,319
	IP_classe	,434	,552	,619	1	,431	1,544
	Constant	-,831	,725	1,316	1	,251	,435
Step 2[a]	Sexo(1)	,355	,556	,409	1	,523	1,427
	Diabetes			1,694	2	,429	
	Diabetes(1)	-,735	,565	1,694	1	,193	,479
	Diabetes(2)	-19,533	8752,731	,000	1	,998	,000
	Escol_classes	-,465	,382	1,475	1	,224	,628
	Fumador(1)	-1,166	,616	3,586	1	,058	,312
	IP_classe	,435	,552	,620	1	,431	1,545
	Constant	-,801	,695	1,329	1	,249	,449
Step 3[a]	Diabetes			1,608	2	,448	
	Diabetes(1)	-,709	,559	1,608	1	,205	,492
	Diabetes(2)	-19,520	8756,092	,000	1	,998	,000
	Escol_classes	-,488	,385	1,603	1	,205	,614
	Fumador(1)	-1,053	,584	3,248	1	,072	,349
	IP_classe	,393	,548	,514	1	,473	1,481
	Constant	-,665	,659	1,019	1	,313	,514
Step 4[a]	Diabetes			1,495	2	,474	
	Diabetes(1)	-,687	,562	1,495	1	,221	,503
	Diabetes(2)	-19,511	8753,746	,000	1	,998	,000
	Escol_classes	-,547	,377	2,103	1	,147	,578
	Fumador(1)	-1,013	,580	3,055	1	,080	,363
	Constant	-,451	,578	,609	1	,435	,637
Step 5[a]	Escol_classes	-,767	,356	4,634	1	,031	,465
	Fumador(1)	-,840	,560	2,250	1	,134	,432
	Constant	-,921	,512	3,239	1	,072	,398
Step 6[a]	Escol_classes	-,720	,345	4,370	1	,037	,487
	Constant	-1,571	,304	26,675	1	,000	,208

a. Variable(s) entered on step 1: Idade_classe, Sexo, Diabetes, Escol_classes, Fumador, IP_classe.

Variables not in the Equation

			Score	df	Sig.
Step 2[a]	Variables	Idade_classe	,023	1	,881
	Overall Statistics		,023	1	,881
Step 3[b]	Variables	Idade_classe	,052	1	,820
		Sexo(1)	,411	1	,521
	Overall Statistics		,434	2	,805
Step 4[c]	Variables	Idade_classe	,045	1	,831
		Sexo(1)	,303	1	,582
		IP_classe	,517	1	,472
	Overall Statistics		,955	3	,812
Step 5[d]	Variables	Idade_classe	,039	1	,843
		Sexo(1)	,255	1	,613
		Diabetes	4,160	2	,125
		Diabetes(1)	,351	1	,554
		Diabetes(2)	2,546	1	,111
		IP_classe	,363	1	,547
	Overall Statistics		5,217	5	,390
Step 6[e]	Variables	Idade_classe	,006	1	,938
		Sexo(1)	,042	1	,838
		Diabetes	3,510	2	,173
		Diabetes(1)	,242	1	,623
		Diabetes(2)	2,319	1	,128
		Fumador(1)	2,346	1	,126
		IP_classe	,178	1	,673
	Overall Statistics		7,577	6	,271

a. Variable(s) removed on step 2: Idade_classe.

b. Variable(s) removed on step 3: Sexo.

c. Variable(s) removed on step 4: IP_classe.

d. Variable(s) removed on step 5: Diabetes.

e. Variable(s) removed on step 6: Fumador.

```
LOGISTIC REGRESSION VARIABLES Na_mais_4
 /METHOD=ENTER Idade_classe Sexo Diabetes Escol_classes Fumador IP_classe
 /SAVE=PRED LRESID ZRESID DEV
 /CLASSPLOT
 /CASEWISE OUTLIER(2)
 /PRINT=GOODFIT CORR CI(95)
 /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9,659	6	,140
	Block	9,659	6	,140
	Model	9,659	6	,140

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	102,408[a]	,059	,117

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4,331	8	,826

Contingency Table for Hosmer and Lemeshow Test

		Nível de aderência >4 = Inferior a 50%		Nível de aderência >4 = Igual ou Superior a 50%		Total
		Observed	Expected	Observed	Expected	
Step 1	1	18	17,601	0	,399	18
	2	16	16,415	1	,585	17
	3	16	16,123	1	,877	17
	4	15	13,888	0	1,112	15
	5	14	14,588	2	1,412	16
	6	12	13,312	3	1,688	15
	7	12	12,122	2	1,878	14
	8	14	13,330	2	2,670	16
	9	16	14,779	3	4,221	19
	10	7	7,843	4	3,157	11

Classification Table[a]

Observed			Predicted		
			Nível de aderência >4		
			Inferior a 50%	Igual ou Superior a 50%	Percentage Correct
Step 1	Nível de aderência >4	Inferior a 50%	140	0	100,0
		Igual ou Superior a 50%	18	0	,0
	Overall Percentage				88,6

a. The cut value is ,500

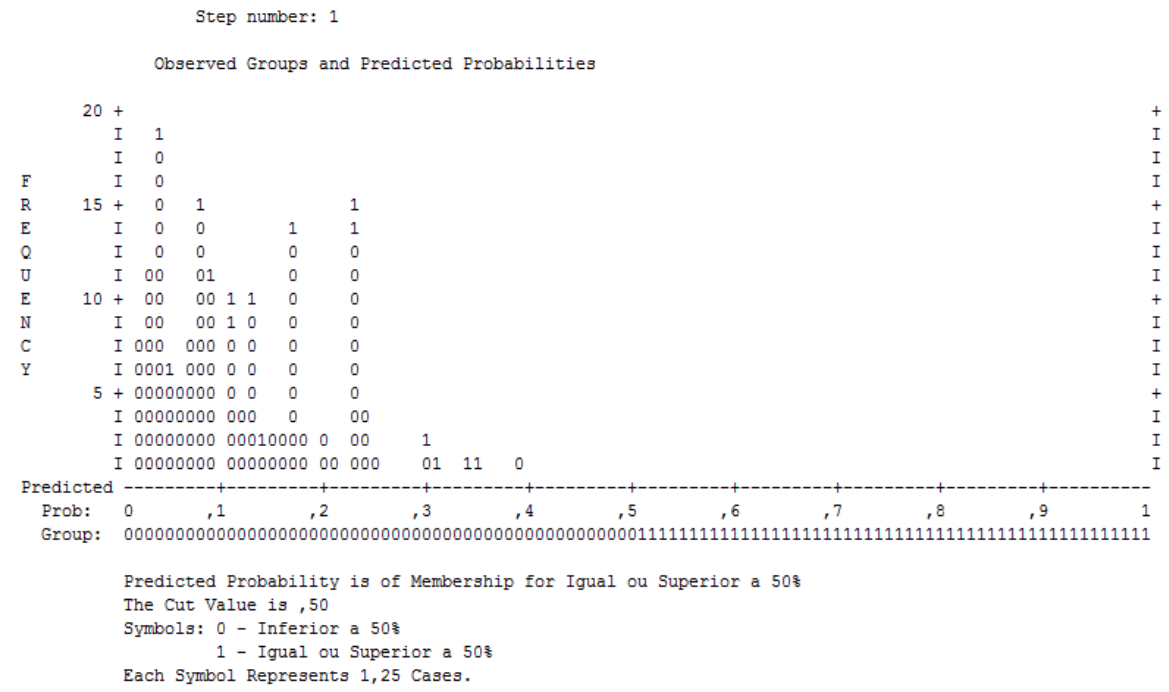
Correlation Matrix

		Constant	Idade_classe	Sexo	Diabetes	Escol_classes	Fumador	IP_classe
Step 1	Constant	1,000	-,139	-,173	-,611	-,369	-,288	-,574
	Idade_classe	-,139	1,000	,119	-,046	-,380	-,265	,000
	Sexo	-,173	,119	1,000	-,074	-,111	-,314	-,122
	Diabetes	-,611	-,046	-,074	1,000	,248	,189	,068
	Escol_classes	-,369	-,380	-,111	,248	1,000	,066	,205
	Fumador	-,288	-,265	-,314	,189	,066	1,000	,181
	IP_classe	-,574	,000	-,122	,068	,205	,181	1,000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Idade_classe	-,170	,620	,075	1	,785	,844	,250	2,847
	Sexo	-,376	,551	,465	1	,496	,687	,233	2,023
	Diabetes	,368	,294	1,566	1	,211	1,445	,812	2,573
	Escol_classes	-,531	,409	1,686	1	,194	,588	,264	1,311
	Fumador	1,134	,630	3,238	1	,072	3,107	,904	10,677
	IP_classe	,428	,554	,596	1	,440	1,533	,518	4,538
	Constant	-2,375	,700	11,516	1	,001	,093		

a. Variable(s) entered on step 1: Idade_classe, Sexo, Diabetes, Escol_classes, Fumador, IP_classe.



Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable	
		Nível de aderência >4			Resid	ZResid
45	S	1**	,107	0	,893	2,882
48	S	1**	,080	0	,920	3,392
59	S	1**	,031	0	,969	5,616
63	S	1**	,131	0	,869	2,573
72	S	1**	,103	0	,897	2,959
93	S	1**	,125	0	,875	2,647
109	S	1**	,104	0	,896	2,928
126	S	1**	,044	0	,956	4,657
149	S	1**	,089	0	,911	3,195

dimension0

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Anexo III

Estatística descritiva com recurso ao R

Análise Descritiva

Objetivos:

1. Construir gráficos.

    A. Construir caixa de bigodes

    B. Construir um diagrama de caule e folhas

    C. Construir um histograma

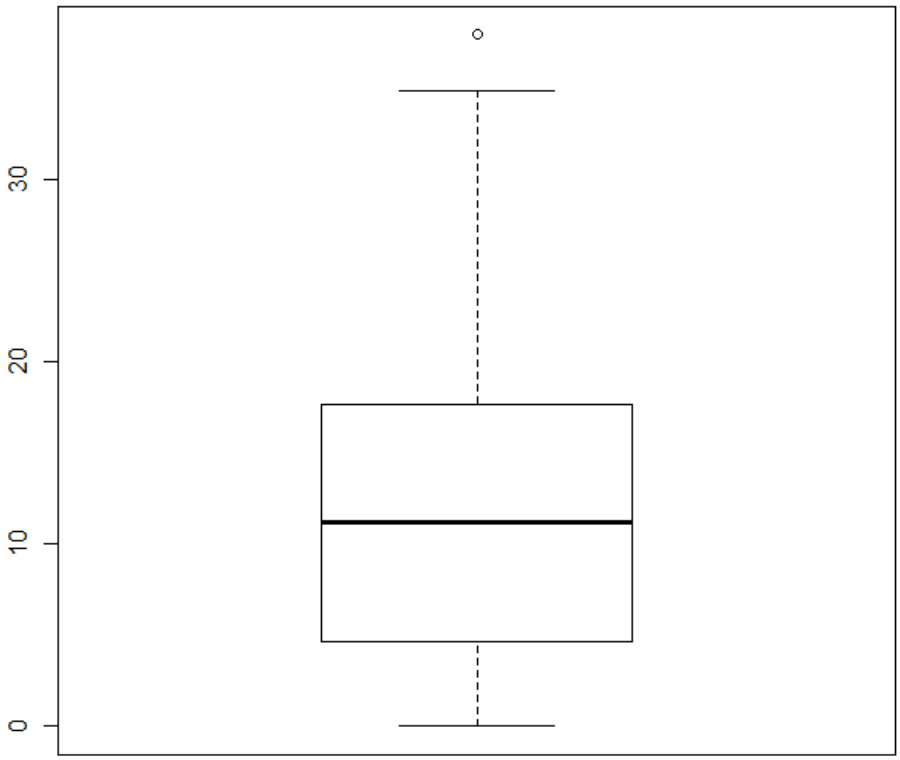
2. Determinar a média, variância, moda e desvio padrão da variável.

1. Construir gráficos

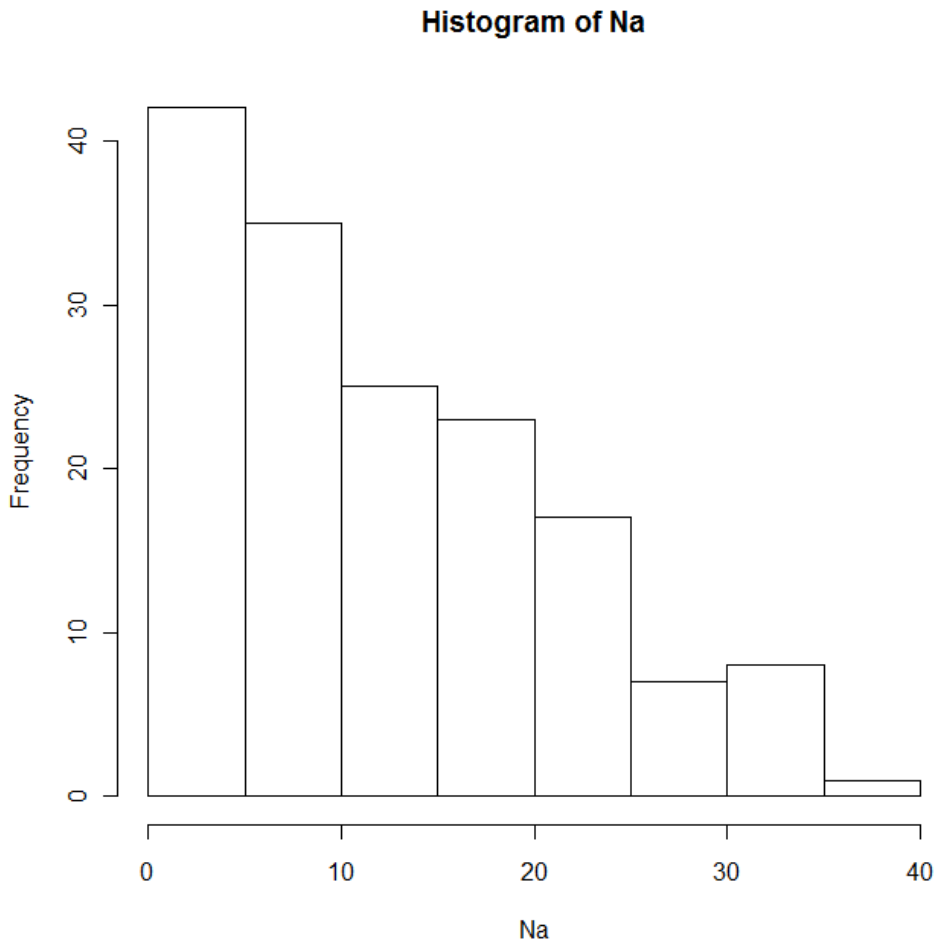
Rotinas:

```
> # Caixa de bigodes
> boxplot (Na)
>

> #Diagrama de caule e folhas
> stem (Na)

> #Histograma
> hist (Na)
>
```

Resultados:




```
The decimal point is at the |

 0 | 00000000000167711124667778
 2 | 35567228
 4 | 023556789112367889
 6 | 0034473335578
 8 | 0037902333789
10 | 41357888
12 | 22388999266789
14 | 2472589
16 | 0177777772457
18 | 24288
20 | 02580357
22 | 027783567
24 | 0479
26 | 68
28 | 10
30 | 898
32 | 553
34 | 19
36 |
38 | 0
```


Histogram of Na
Frequency
0
10
20
30
40
0
10
20
30
40
Na