Universidade Aberta

Modelos de Regressão Linear e Logística

utilizando o software R

João Pedro Bento Clemente da Silva

Mestrado em Estatística, Matemática e Computação

Área de especialização em Estatística Computacional

2016

Universidade Aberta

Modelos de Regressão Linear e Logística

utilizando o software R

João Pedro Bento Clemente da Silva

Orientadora: Profª Doutora Teresa Paula Costa Azinheira Oliveira

Co-orientador: Prof. Doutor Amílcar Manuel do Rosário Oliveira

2016

Resumo

A Regressão é uma técnica estatística utilizada na investigação da relação entre variáveis que surgem em problemas das mais variadas áreas da ciência. De uma forma geral, o investigador procura aferir a influência de uma variável explicativa *X* sobre o valor esperado de uma variável de resposta denominada *Y*. Por exemplo, qual o efeito do aumento de preços na procura; qual o efeito da pressão sanguínea no risco de acidente cardiovascular ou qual o efeito das descargas na probabilidade de cheia. Os modelos de regressão são vários e são definidos consoante o número de variáveis explicativas, cujos efeitos na variável *Y*, que se pretendem estudar. Nesta dissertação, serão abordados os modelos de regressão linear simples, Regressão linear múltipla e logística através do software estatístico R em problemas de contexto real no âmbito da análise do risco.

***Palavras chave**: regressão Linear simples, Regressão linear Múltipla, Regressão Logística, software R, Análise do risco.*

Abstract

Regression is a wide used statistical technique in several science fields with the purpose of studying the relationship between variables. Reseachers seek to know how a explanatory variable X relates to the expected value of the response variable Y. For exemple, a researcher might need to know the effect of price increase in a good´s demand; the effect of blood pressure on a heart attack risk or the effect of waterdischarge on flood probability. Regression models are defined acording to the number of explanatory variables and its effects on Y. In this paper, simple, multiple and simple logistic regression models will be analized and aproached with R software with real data regarding risk analysis.

Keywords: Simple Linear regression, Multiple Regression, Logistic Regression, R software, Risk analysis

Agradeço à minha família todo o apoio e paciência

e aos meus orientadores

pela orientação prestada, pelo incentivo,

disponibilidade e apoio demonstrados.

Índice Geral

Índice de tabelas

Indíce de Figuras

Lista de abreviaturas, siglas e acrónimos

g.l.	graus de liberdade
I.C.	Intervalos de Confiança
SQR	Soma dos Quadrados dos Erros
SQT	Soma dos Quadrados Total
SQE	Soma dos Quadrados dos Erros
QME	Quadrado Médio dos Erros
SD_y	Desvio padrão da variável *y*
SD_x	Desvio padrão da variável *x*
ANOVA	Análise de Variância
MMQ	Método dos Minimos Quadrados
RLS	Regressão linear Simples
MRLS	Modelo de Regressão Linear Simples
MRLM	Modelo de Regressão Linear Múltipla
ROC	*Receiver Operating Characteristic*
OR	*Odds Ratio*
VAR	Variância
VP	Verdadeiros positivos
FP	Falsos Positivos
FN	Falsos Negativos
VN	Verdadeiros Negativos
H_0	Hipótese Nula

H_1	Hipótese Alternativa
VD	variável dependente
VI	Variável independente (ou explicativa)
σ	Desvio padrão
α	Nivel de significância – Erro Tipo I
R^2	Coeficiente de determinação
ε	erro aleatório

INTRODUÇÃO

Foi a navegação maritíma e o recurso à astronomia no século XVIII que originou os primeiros grandes problemas de regressão. O matemático Adrien Marie Legendre (1752-1833) desenvolveu no inicio do século dezanove o Método dos Minimos Quadrados. No entanto, Johann Carl Friedrich Gauss (1777-1855) afirmou ter desenvolvido esse mesmo método e demonstrou que este era a solução ótima para conjuntos de dados cujos erros eram normalmente distribuidos. Esta metodologia foi aplicada de forma quase exclusiva apenas às ciências denominadas fisicas e em particular à astronomia, até finais do século XIX. Francis Galton (1822-1911) criou o conceito de regressão em relação à média e define uma conceção de correlação na década de 1880 que será mais tarde definido como coeficiente de correlação cuja definição e simbologia será posteriomente aperfeiçoada por Karl Pearson (1857-1936) . Galton utilizou a noção de correlação para explicar que os filhos de pais altos tendem a ser altos mas não tanto como os pais e que filhos de pais baixos tendem a ser baixos mas não tão baixos quanto os pais. Assim, Galton foi o primeiro investigador a utilizar a regressão no estudo de fenómenos associados ao ser humano.

A Regressão é uma técnica utilizada na investigação da relação entre variáveis que surgem em problemas das mais variadas áreas da ciência. De uma forma geral, o investigador procura aferir a influência de uma variável explicativa *X* sobre o valor esperado de uma variável de resposta denominada habitualmente por *Y*. Por exemplo, qual o efeito do aumento de preços na procura; qual a relação do número de pedidos de indemnização e os montantes pagos por uma dada seguradora; qual o efeito da pressão sanguínea no risco de acidente cardiovascular ou qual o efeito das descargas na probabilidade de cheia. Os modelos de regressão são vários e são definidos consoante o número e distribuição das variáveis explicativas, cujos efeitos na variável *Y*, que se pretendem estudar. Se o investigador está interessado na relação de apenas uma variável explicativa com a variável resposta então aplica-se a Regressão Linear Simples**.** Mas se pretende relacionar a variável resposta com mais do que uma variável explicativa, então um modelo de Regressão Linear Múltipla será o mais apropriado. Caso a variável resposta seja uma variável categórica, ou seja, a variável apresenta como possíveis realizações uma qualidade (ou atributo) e não

uma mensuração, utiliza-se o Modelo de Regressão Logística. Neste trabalho serão aplicados os vários modelos referidos através do software estatístico R em problemas de contexto real.

O objetivo desta dissertação é apresentar a regressão linear simples, regressão linear múltipla e regressão logística como suporte para a análise e avaliação de situações de risco em várias áreas e apresentar casos práticos com dados tratados e modelados em R. Assim, neste trabalho pretende-se utilizar a regressão nas formas acima citadas para inferir sobre montantes pagos por uma dada seguradora em função do número de pedidos de indemnização; volume de ações transacionadas num dado período; qual a probabilidade de acidente cardíaco em função da idade e qual a probabilidade de penetração de um tumor na cápsula gástrica.

O presente trabalho está dividido em cinco capítulos que, em seguida, serão descritos de forma resumida. O primeiro e segundo capítulos consistem na definição dos modelos de regressão linear simples e regressão linear múltipla respetivamente, na sua forma analítica e matricial, bem como da análise de variância, cálculo de intervalos de confiança e da análise de resíduos e respetivos testes inerentes aos pressupostos e adequação dos modelos adotados para os problemas em análise. No final do primeiro capitulo será descrito um exemplo de realização de um modelo de regressão linear simples(MRLS) e sua análise, tal como as principais funções e sintaxes utilizadas em R para tal. No final do segundo capítulo será exemplificada a obtenção e análise de um modelo de regressão linear múltipla (MRLM) tal como as principais funções e sintaxes utilizadas em R para tal No terceiro e quarto capítulos serão abordados e descritos os modelos de regressão logística simples e regressão logística múltipla, repetivamente, e as suas características tal como os testes mais utilizados para verificação do modelo em análise. Os modelos logisticos também são designados por modelos lineares generalizados (MLG). Nestes dois capitulos, tal como o primeiro e no segundo, também serão incluídos exemplos de criação e análise de MLG, simples e múltiplo respetivamente. As principais funções e sintaxes utilizadas em R para a realização destas metodologias também serão descritas e apresentadas, assim como os gráficos e figuras considerados necessários para a sua análise.

No capítulo cinco serão apresentados os casos de contexto real, associados a temas no âmbito da análise do risco com dados reais e aplicação de modelos de regressão adequados

para a sua possível modelação. O tratamento de dados será realizado através do software R. Todos os gráficos apresentados ao logo desta dissertação foram integralmente realizados pelo autor da mesma. Na conclusão e considerações finais serão descritas as conclusões gerais desta dissertação quanto aos modelos analisados e avaliados após a sua aplicação. Nesta ultima parte, também serão apontadas algumas das principais fragilidades dos modelos de regressão linear e regressão logística, assim como outras das metodologias utilizadas atualmente no âmbito da análise do risco.

CAPÍTULO 1

1. REGRESSÃO LINEAR SIMPLES

A análise de regressão consiste na criação de um modelo matemático que relacione, essencialmente, a variável dependente denominada por *Y* e a variável independente *X*. Estas variáveis, no âmbito da análise de regressão, tomam as denominações de variável de resposta e variável explicativa. O objetivo é verificar e analisar como o valor esperado da variável de resposta, *E[Y]*, é afectado consoante a alteração das condições que interagem com a variável *Y*. A variável explicativa *X* deverá fornecer informação sobre o comportamento da variável *Y*. Além de *X*, do modelo também consta a inclusão de parâmetros, habitualmente designados pela letra β e que são estimados a partir dos dados recolhidos.

O modelo de regressão linear simples consiste na utilização de apenas uma variável explicativa para a determinação do comportamento esperado de *Y*. Assim, a relação linear simples entre *Y* e *X* é definida pela equação linear

$$E[Y_i] = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1.1)$$

Em que:

β_0 é denominado intercepto ou coeficiente linear e corresponde á ordenada na origem da representação gráfica do modelo linear;

β_1 é denominado coeficiente de regressão ou angular e representa o declive da reta associada ao modelo linear;

O índice *i* indica a referência de cada observação com *i=1,2,...,n*;

X_i são as *n* observações da variável explicativa.

ε_i representa o desvio entre cada observação real Y_i e o correspondente valor estimado pelo modelo $E[Y_i]$. Os erros ε_i têm valor médio zero e variância σ^2. A distribuição de ε_i é assumida como Normal e habitualmente é descrita na forma $\varepsilon_i \sim N(0,\sigma^2)$.

1.1. Estimação dos parâmetros β_0 e β_1 : Método dos mínimos quadrados

O primeiro passo na análise da regressão linear é obter as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ relativas aos parâmetros β_0 e β_1 do modelo linear. Os valores dessas estimativas serão obtidos a partir de uma amostra de *n* pares de valores (X_i,Y_i), *i=1,...,n*, observados. Karl Gauss (1777-1855) entre outras investigações, propôs estimar os parâmetros, β_0 e β_1, visando minimizar a soma dos quadrados dos desvios, ε_i, com *i=1,...,n*, denominando este processo de método dos mínimos quadrados.

Seja S a soma dos quadrados dos desvios ε_i tal que:

$$S = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (1.2)$$

Para determinar as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ é necessário minimizar a expressão (1.2) em relação aos parâmetros β_0 e β_1. Para tal é necessário obter as derivadas parciais:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \qquad \text{e} \qquad \frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i$$

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

$$-2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \qquad \text{e} \qquad -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Obtendo assim, após simplificação, o sistema

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Resolvendo o sistema obtém-se:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \Leftrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_i \qquad (1.3)$$

em que $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i$ e $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ são as médias das variáveis *X* e *Y*.

Após substituição e simplificação na segunda equação do sistema obtém-se:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (1.4)$$

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

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \ .$$

1.2. Variância residual

A variância residual também deve ser estimada de forma a garantir análises precisas sobre as inferências feitas a partir do modelo linear. A estimativa da variância residual σ^2 consiste no quociente entre a soma dos quadrados dos resíduos (ou erros), e os graus de liberdade, assumindo que os erros são independentes e com média zero. Assim, temos que:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2} \tag{1.5}$$

O número de graus de liberdade (*g.l.)* é determinado através da subtração do valor dos *n* casos e o número de parâmetros do modelo linear, ou seja *n-2*. A regressão linear simples requer apenas dois parâmetros. O estimador $\hat{\sigma}^2$ também é denominado Quadrado médio dos erros (*QME).* O *QME* também pode ser calculado da seguinte forma:

$$QME = S_{yy} - \frac{S^2{}_{xy}}{S_{xx}} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} \tag{1.6}$$

em que $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ e $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

1.3. Propriedades dos estimadores pelo método dos mínimos quadrados

i. Valor esperado de $\hat{\beta}_1$.

Seja $C_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$. Então,

$$E[\hat{\beta}_1] = E\left(\sum_{i=1}^{n} C_i Y_i\right) = \sum_{i=1}^{n} C_i E[Y_i] = \sum_{i=1}^{n} C_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} C_i + \beta_1 \sum_{i=1}^{n} C_i x_i .$$

Como

$$\sum_{i=1}^{n} C_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0$$

e

$$\sum_{i=1}^{n} C_i x_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 1 ,$$

Conclui-se que $E\left[\hat{\beta}_1\right] = \beta_1$.

ii. Variância de $\hat{\beta}_1$**.**

$$Var\left(\hat{\beta}_1\right) = Var\left(\sum_{i=1}^{n} C_i y_i\right) = \sum_{i=1}^{n} C_i{}^2 Var\left(y_i\right) = \sum_{i=1}^{n} \left(\frac{\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \right)^2 \sigma^2 =$$

$$= \frac{\sigma^2}{S_{xx}}$$

iii. Valor esperado de $\hat{\beta}_0$

$$E\left[\hat{\beta}_0\right] = E\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] = E\left[\bar{y}\right] - \bar{x}E\left[\hat{\beta}_1\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] - \bar{x}\beta_1 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[y_i\right] - \bar{x}\beta_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0 - \beta_1 x_i\right) - \bar{x}\beta_1 =$$

$$= \beta_0 - \beta_1 \sum_{i=1}^{n} \frac{x_i}{n} - \bar{x}\beta_1 = \beta_0.$$

iv. Variância de $\hat{\beta}_0$

$$Var\left(\hat{\beta}_0\right) = Var\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) = Var\left(\bar{y}\right) + Var\left(\hat{\beta}_1 \bar{x}\right) - 2Cov\left(\bar{y}, \hat{\beta}_1 \bar{x}\right) =$$

$$= \sigma^2 \left(\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}} \right).$$

v. Covariância entre $\hat{\beta}_0$ **e** $\hat{\beta}_1$**.**

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\sigma^2 \frac{\bar{x}}{S_{xx}}$$

vi. Distribuição amostral de $\hat{\beta}_0$ **e** $\hat{\beta}_1$.

$$\hat{\beta}_0 \sim N\left[\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right] \quad \text{e} \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

As variâncias dos parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$ estão localizadas na matriz de covariância

$$\hat{cov}(\beta) = \begin{bmatrix} \hat{\sigma}^2(\hat{\beta}_0) & \hat{cov}(\beta_0, \beta_1) \\ \hat{cov}(\beta_1, \beta_0) & \hat{\sigma}^2(\hat{\beta}_1) \end{bmatrix}$$

1.4. Intervalos de confiança e testes sobre β_0 e β_1.

A realização de testes de hipóteses sobre o parâmetro β_1 é uma das formas de avaliar a capacidade da variável explicativa *X* como preditor da variável *Y*. A hipótese $\beta_1 = 0$ significa que não existe uma relação linear entre *X* e *Y*. O teste de hipótese requer os seguintes pressupostos: Para cada x_i, os erros ε são independentes, normalmente distribuídos com valor médio nulo e variância σ^2. Com estes pressupostos, $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores não enviesados de β_0 e β_1 respetivamente. Com as variâncias e distribuições anteriormente definidas, o teste de hipóteses consiste em testar:

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

Assim, sob a assunção de normalidade, o teste apropriado será

$$t_1 = \frac{\hat{\beta}_1}{\sqrt{\frac{QME}{S_{xx}}}}. \qquad (1.7)$$

A estatística t_1 tem uma distribuição *t-student* com *n-2* graus de liberdade.

Assim, rejeitamos H_0 com um nível de confiança $(1-\alpha)100\%$ se $|t_1| > t_{\left(1-\frac{\alpha}{2}, n-2\right)}$.

O intervalo de confiança para $\hat{\beta}_1$ com $(1-\alpha)100\%$ é

$$\left[\hat{\beta}_1 \mp t_{\left(1-\frac{\alpha}{2}, n-2\right)} \sqrt{\frac{QME}{S_{xx}}}\right] \qquad (1.8)$$

Para o intercepto β_0, também são calculados intervalos de confiança, supondo que se quer verificar se β_0 toma um dado valor sendo β_{00} esse valor. Assim, considerando as hipóteses

$$H_0 : \beta_0 = \beta_{00} \text{ vs } H_1 : \beta_0 \neq \beta_{00}$$

A estatística de teste é

$$t_0 = \frac{\hat{\beta}_0}{\sqrt{QME\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{(n-2)} \tag{1.9}$$

Então, rejeitamos H_0 com um nível de confiança $(1-\alpha)100\%$ se $|t_0| > t_{\left(1-\frac{\alpha}{2}, n-2\right)}$. Quando H_0 não é rejeitado, podemos utilizar o modelo de regressão sem intercepto.

O intervalo de confiança é dado por

$$\left[\hat{\beta}_0 \mp t_{\left(1-\frac{\alpha}{2}, n-2\right)} \sqrt{QME\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right] \tag{1.10}$$

1.5. Intervalo de confiança para a resposta média

Em alguns problemas pode ser necessário estimar a resposta média em função de uma variável explicativa especifica x_0. O parâmetro de interesse é $\mu^* = \beta_0 + \beta_1 x_0$. Um estimador pontual de μ^* pode ser obtido a partir do modelo ajustado $\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$ que podemos escrever como $\hat{y}(x_0)$.

Assim, temos que:

$$E\left[\hat{\mu}^*\right] = \beta_0 + \beta_1 x_0 = \mu^*$$

$$Var\left[\mu^*\right] = \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]$$

O intervalo de confiança para a resposta média é dado por:

$$\left[\hat{y}\left(x_0\right) \mp t_{\left(1-\frac{\alpha}{2}, n-2\right)}\sqrt{QME\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)}\right] \tag{1.11}$$

Para vários valores considerados de x_0, dentro do intervalo de realização dos dados, determinam-se vários valores para $\hat{y}\left(x_0\right)$. Assim, é possível calcular intervalos de confiança para $\hat{y}\left(x_0\right)$ que formam um conjunto que representam as bandas de confiança para a reta de regressão.

1.6. Análise de Variância

A análise de variância é um método considerado adequado para avaliar a significância do modelo. Este método consiste na decomposição da soma dos quadrados e nos graus de liberdade associados à variável de resposta. O desvio de uma observação em relação à média pode ser decomposto como o desvio da observação em relação ao valor ajustado pela regressão somado ao desvio do valor ajustado em relação.

Assim, podemos escrever

$$\left(Y_i - \bar{Y}\right) = \left(Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i\right) = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right) \tag{1.12}$$

Soma de Quadrados

A partir de (1.12) temos que $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$, em que:

$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ é a Soma de Quadrados Total (SQT);

$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ é a soma de Quadrados de Regressão (SQR) e

$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ é a soma de Quadrados dos erros ou Resíduos (SQE).

Então, podemos escrever que

$$SQT = SQR + SQE \tag{1.13}$$

Graus de Liberdade

Cada soma de quadrados tem um número de graus de liberdade associado à regressão linear simples. A decomposição de graus de liberdade é:

SQT: $gl_{Total} = n - 1$;

SQR: $gl_{\text{Re}gressão} = 1$

SQE: $gl_{Erros} = n - 2$

Quadrado Médio

O quadrado médio é obtido pelo quociente entre Soma de Quadrados e o respetivo número de graus de liberdade.

$$QMR = \frac{SQR}{1} = SQR$$

$$QME = \frac{SQE}{n-2}$$

Tabela 1:*Tabela de Análise de Variância para a regressão Linear Simples (ANOVA)*

Fonte	G.L	Soma de Quadrados	Quadrado Médio	Teste F
Regressão	*1*	$SQR = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$	$QMR = SQR$	$F_0 = \frac{QMR}{QME}$
Resíduo	*n-2*	$SQE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$	$QME = \frac{SQE}{n-2}$	
Total	*n-1*	SQT		

A estatística F_0 obtida na tabela é utilizada para testar a significância da regressão através de um teste com as seguites hipóteses:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

Para tal, a regra de decisão consiste em rejeitar H_0 se $F_{Calc} \geq F_{(\alpha,1,n-2)}$

1.7. Coeficiente de determinação

O coeficiente de determinação, também conhecido como R^2 fornece informação auxiliar ao resultado da análise de variância da regressão, como uma forma de se verificar se o modelo proposto é adequado ou não para descrever o fenômeno.

O R^2 é obtido através do quociente:

$$R^2 = \frac{SQR}{SQT} \tag{1.14}$$

O valor de R^2 varia no intervalo de 0 a 1. Habitualmente, aceita-se que valores próximos de 1 indicam que o modelo proposto é adequado para descrever o fenômeno. O coeficiente R^2 indica a proporção da variação de *Y* que é "explicada" pela regressão, ou quanto da variação na variável dependente *Y* está sendo "explicada" pela variável independente *X*

1.8. Análise de Resíduos

A distribuição dos resíduos é utilizada para a verificação de adequabilidade do modelo linear. Uma das técnicas para tal consiste na representação, num referencial cartesiano, dos pontos cujas coordenadas representam os resíduos e os valores ajustados. Os resíduos são dados por:

$$e_i = \hat{e}_i = y_i - \hat{y}_i \tag{1.15}$$

Os pontos do gráfico devem distribuir-se de forma aleatória em torno da reta que corresponde ao resíduo zero, formando uma mancha de largura uniforme. Assim, será de esperar que os erros sejam independentes, de média nula e de variância constante.

Quando os resíduos não se comportam de forma aleatória, ou seja, seguem um padrão, a condição de independência poderá não ser satisfeita.

Isto pode traduzir o facto de não existir uma relação linear entre as variáveis ou então, não constam no modelo uma ou várias variáveis independentes que influenciam significativamente a variável dependente e, portanto também os erros.

A análise de resíduos, através da observação de gráficos residuais pode ser um método pouco preciso. No entanto, existem situações que representam claramente a falta de independência dos erros.

Na figura seguinte, os primeiros dois gráficos representam comportamentos padronizados dos resíduos, logo facilmente se concluí que não há independência. Se a dispersão dos resíduos aumentar ou diminuir com os valores das variáveis independentes x_i, ou com os valores estimados da variável dependente y_i, deve ser posta em causa a hipótese de variâncias constante dos e_i. Este comportamento é observável nos últimos dois gráficos da figura. Pelo contrário, no gráfico da Figura 2 os resíduos parecem estar distribuídos de forma aleatória, sustentando a independência dos erros.

Figura 1: *Exemplos de Gráficos residuais*

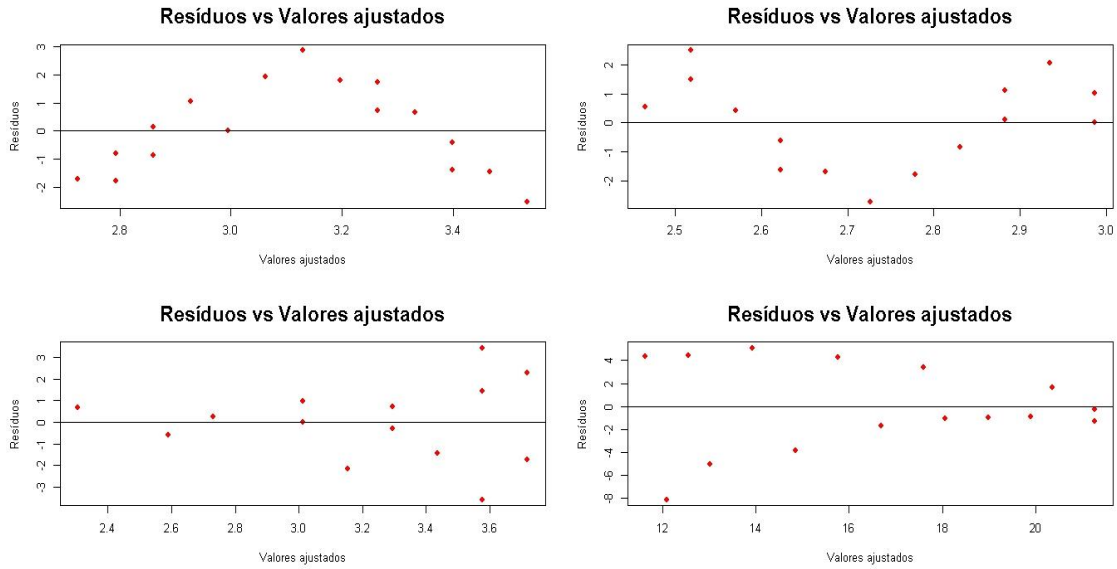
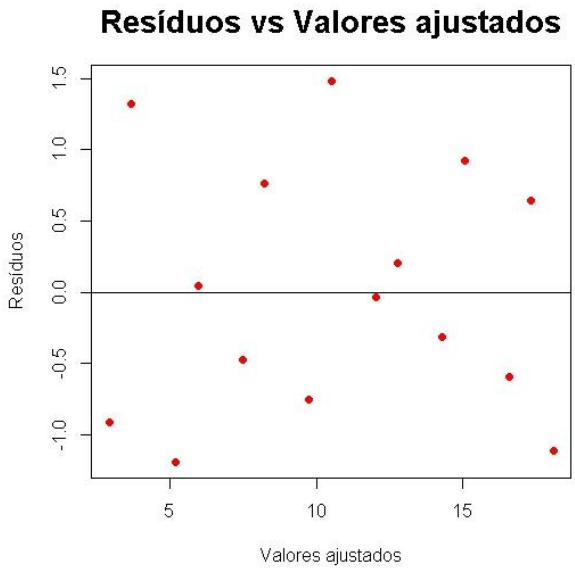


Figura 2: Exemplo de *Homocedasticidade*



A análise gráfica para a verificação da independência deve ser complementada com a realização de testes estatísticos de hipóteses.

Outros gráficos que permitem uma análise ao modelo determinado são:

Papel de probabilidade normal*:* Os pontos definidos por cada par de valores (valor teórico, valor "observado") são marcados num sistema de eixos dando origem a uma nuvem de pontos. Se esta nuvem de pontos evidenciar uma relação linear entre abcissas e

ordenadas, temos uma validação informal da normalidade da população de onde foi retirada a amostra, como pode ser observado na Figura 3.

Gráfico dos Resíduos versus valores ajustados*:* Diagrama de pontos em que os pares de valores são definidos por coordenadas que correspondem à raíz quadrada do resíduo e ao respetivo valor ajustado. Tal como no primeiro caso, os pontos não devem apresentar um padrão predominante. Este gráfico permite a deteção de valores atipicos (*outliers).*

Gráfico dos Resíduos versus valores alavanca *(leverage)*: No contexto de modelos lineares, a matriz *H* definida por $X(X^TX)^{-1}X^T$, corresponde à matriz de da solução dos minimos quadrados, sendo que os elementos da sua diagonal $h_{ii} = x_i^T \left(x_i^T x_i\right)x_i$ com $i = 1,2,...,n$ são denominadas medidas de alavancagem, que expressam quão extremas são as observações no espaço das covariáveis. Observações mais afastadas devem ser avaliadas com maior cuidado.

A notação matricial dos modelos lineares será aprofundada no próximo capítulo. Um exemplo sobre este tipo de gráfico pode ser observado na figura 3. Este gráfico produzido em R contém a distância de Cook. A distância de Cook mede a influência da observação *i* sobre todos *n* valores ajustados $\hat{Y}_i$, sendo definida por:

$$D_i = \frac{e_i^2}{QME} \frac{h_{ii}}{\left(1 - h_{ii}\right)^2}.$$

No entanto, também é possível obter e observar gráficos sobre a distância de Cook (Figura 3).

Teste de Durbin-Watson

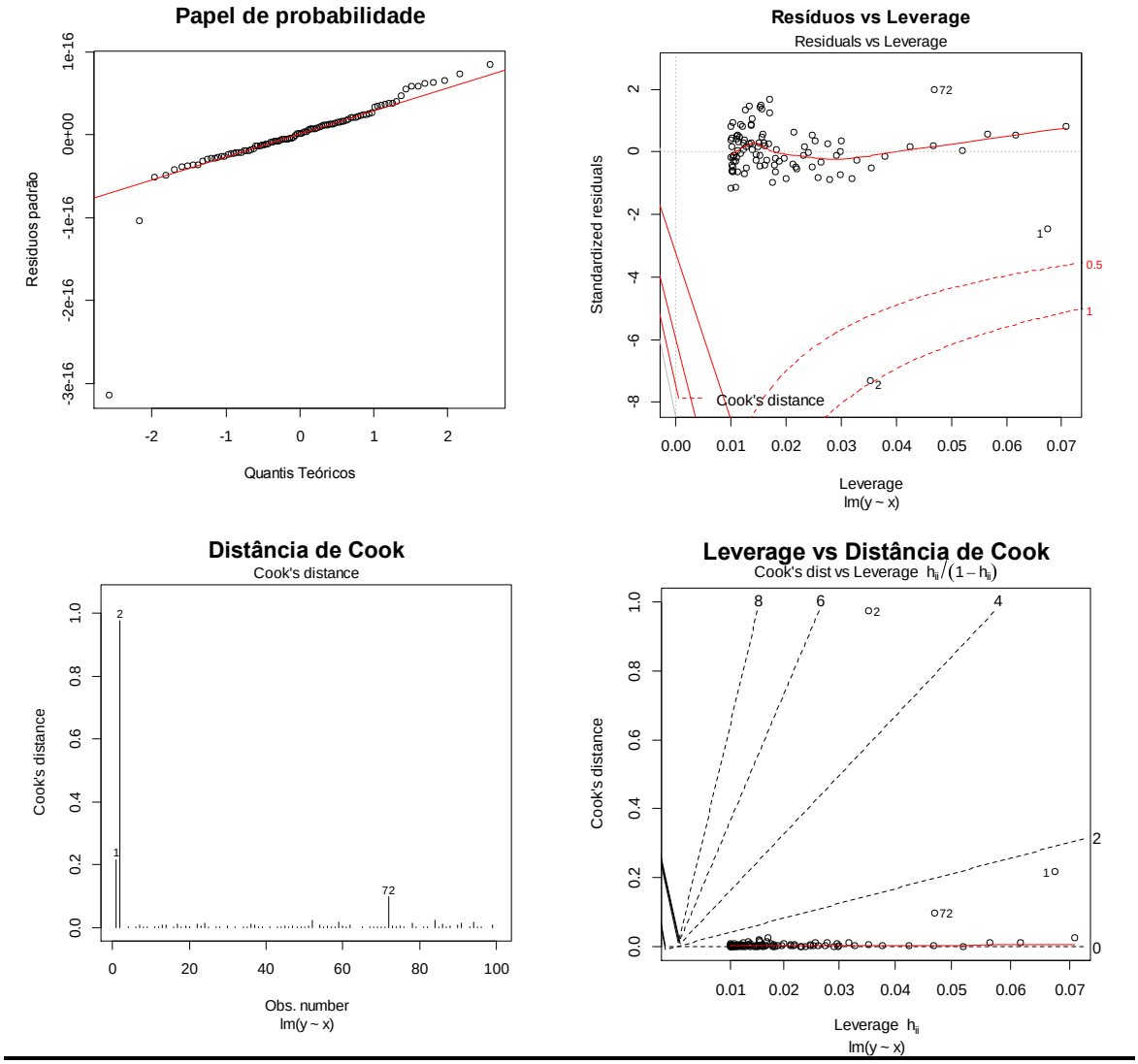
Se houver independência, a magnitude de um resíduo não influencia a magnitude do resíduo seguinte. Neste caso, a correlação entre resíduos sucessivos é nula $(\rho = 0)$. As hipóteses do teste, para aferir se a relação entre dois resíduos consecutivos é estatísticamente significativa, são então:

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

E estatística *d* de Durbin-Watson é dada por:

$$d = \frac{\sum_{i=1}^{n}(e_{i+1} - e_i)^2}{\sum_{i=1}^{n}(e_i)^2} \tag{1.15}$$

Figura 3:*Gráficos de Papel de Probabilidade, Alavancagem (Leverage), Distância de Cook e Alavancagem vs Distância de Cook*



Para a tomada de decisão, Compara-se o valor obtido para a estatística *d* com os valores críticos da tabela de Durbin-Watson, d_L e d_U , e toma-se a decisão recorrendo à tabela 2. A determinação dos valores d_L e d_U será abordada no exemplo de aplicação 1.9.

Tabela 2: *valores para o teste de Durbin-watson*

d	$[0, d_L[$	$[d_L , d_U[$	$[d_U ,4- d_U[$	$[4- d_U ,4- d_L [$	$[4- d_L ,4[$
Decisão	Rejeitar H_0	Teste Inconclusivo	Não rejeitar H_0	Teste Inconclusivo	Rejeitar H_0

Teste de Breusch-Pagan

O teste de Breusch-Pagan é bastante utilizado para testar a hipótese nula de que as variâncias dos erros são iguais (homoscedasticidade) em oposição à hipótese alternativa de que as variâncias dos erros são uma função multiplicativa de uma ou mais variáveis, sendo que esta(s) variável(eis) pode(m) pertencer ou não ao modelo em questão. É indicado para grandes amostras e quando a suposição de normalidade nos erros é assumida.

A estatística de teste neste caso é obtida da seguinte forma:

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

$$u_i = \frac{e_i^{\,2}}{SQE/n} \quad i=1,...,n \tag{1.16}$$

Finalmente, realiza-se a regressão entre $u = (u_1,...,u_1)$ (variável resposta) e o vetor $\hat{y}$ (variável explicativa) e obtemos a estatística do teste χ^2 calculando a Soma de Quadrados da Regressão de u sobre $\hat{y}$ e dividindo o valor encontrado por 2. Sob a hipótese nula, esta estatística tem distribuição Qui-quadrado com 1 grau de liberdade. Resumidamente, se não existe heteroscedasticidade, é de se esperar que os resíduos ao quadrado não aumentem ou diminuam com o aumento do valor predito $\hat{y}$, e assim, a estatística de teste deveria ser insignificante.

1.9. Comandos e sintaxes em R para obtenção e análise de um modelo de regressão linear simples

Os comandos (ou sintaxe) a utilizar para realizar a regressão linear simples com todos os elementos descritos anteriormente serão apresentados em seguida.

A sintaxe básica para obter o modelo de regressão é

lm(Y~modelo)

Onde *Y* é a variável de resposta e *modelo* é a fórmula correspondente ao modelo matemático determinado pelo investigador.

O comando *lm()* apenas fornece os coeficientes β_0 e β_1 sem qualquer informação estatística adicional.

Os dados utilizados neste exemplo foram retirados do ficheiro diamond.dat.txt. que está disponível no site da *American Statistical Associaton*, mais precisamente no link http://www.amstat.org/publications/jse/datasets/diamond.dat.txt. Os dados foram copiados e colados num ficheiro de texto com a denominação *diamonddata.txt*. Para inserir os dados em ambiente *R* pode utilizar-se o seguinte procedimento:

Localizar a diretoria em que se encontra o ficheiro e designar o objeto que será utilizado como tabela de dados, através da sintaxe:

```
> dados1<-read.table("C:/Users/Jonas/Desktop/diamonddata.txt")
```

Para este caso exemplificativo as variáveis estão descritas na tabela seguinte.

Tabela 3: *Variáveis sobre o exemplo de RLS em R*

Abreviatura	***Variável***	***Unidades de medida***
price	*variável dependente contínua relativa ao preço do diamante*	Dolar Singapura
carats	*variável independente que representa o número de quilates de um diamante em gramas (1 quilate =0.2 g)*	Valores contínuos

Pretende-se elaborar um modelo linear que permita obter e analisar o preço dos diamantes a partir do valor dos quilates. A função definida para a criação de modelos de regressão

linear em R é designada *lm()* e os argumentos principais consistem na relação entre a variável de resposta e a variável explicativa simbolizada pelo simbolo til(~) e também a designação do conjunto de dados que estão a ser utilizados. Então, utilizando o software R obtemos o seguinte *output* para o modelo linear *lm(price~carats,data=dados1*) :

Tabela 4: *output para modelo linear simples*

```
Call:
lm(formula = price ~ carats, data = dados1)
Coefficients:
(Intercept)      carats
    -259.6        3721.0
```

Para uma utilização mais simples e eficaz do modelo criado é necessário criar um objeto identificado com a função do modelo linear utilizado. Assim, podemos escrever:

modelo1<- lm(price~carats,data=dados1)

em que o objeto *modelo1* pode e deve ser utilizado como argumento para outros funções (ou comandos) importantes relacionados com a análise da regressão linear em ambiente R.

Para obter informação mais precisa e completa sobre o modelo linear criado, é necessário recorrer à função *summary()*.

Tabela 5:*output com informação mais detalhada sobre o modelo linear*

```
> summary(modelo1)

Call:

lm(formula = price ~ carats, data = dados1)

Residuals:

   Min     1Q     Median     3Q      Max
-85.159  -21.448  -0.869   18.972   79.370
Coefficients:

             Estimate    Std. Error    t value    Pr(>|t|)

(Intercept)   -259.63     17.32       -14.99     <2e-16 ***

carats         3721.02     81.79        45.50     <2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

Através do *output* produzido é possível verificar a informação acerca de medidas de dispersão e localização central sobre os resíduos. Quanto aos coeficientes do modelo

linear, é possível verificar os valores das estatísticas de teste permitem não aceitar as hipóteses nula para os testes $H_{0:}\beta_1 = 0$ vs $H_{1:}\beta_1 \neq 0$ e $H_{0:}\beta_0 = \beta_{00}$ vs $H_{1:}\beta_0 \neq \beta_{00}$. Neste caso, os coeficientes determinados são evidentemente não nulos pois a estatística de teste correspondente aos testes definidos em (1.7) e (1.9) apresenta valores inferiores a 0.05. No fundo da tabela estão assinalados os valores referentes ao desvio padrão do modelo, o coeficiente de regressão e a estatística *F* relativa ao modelo, e os respetivos graus de liberdade.

Neste *output* também é possível observar o valor do coeficiente de determinação ajustado R^2 denominado *adjusted r squared* . Neste caso, o valor obtido é 0,9778, ou seja, este modelo tem uma qualidade de ajuste de cerca de 98%.

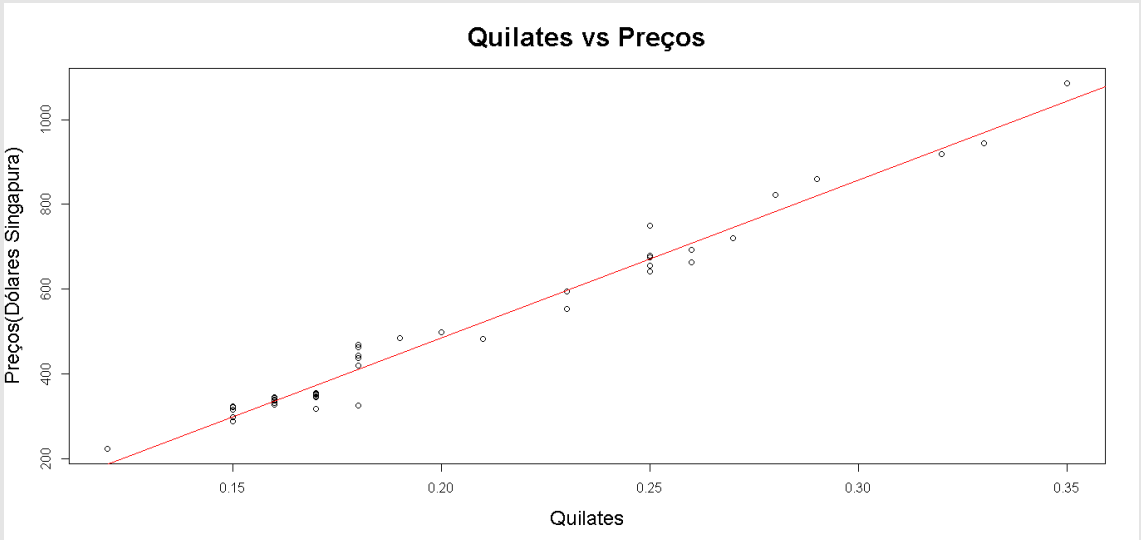
Então, o modelo linear obtido pode ser descrito através da expressão:

$$price = -259.63 + 3721.02 \times carats$$

Para visualizar o modelo linear obtido na forma de reta de regressão é necessário recorrer às funções *plot()* e *abline()*.

Figura 4: *Sintaxe e Gráfico do modelo de RLS*

```
> plot(price~carats,data=dados1,main="Quilates vs
Preços",cex.main=2,xlab="Quilates",ylab="Preços(Dólares Singapura)",cex.lab=1.5)

> abline(reg_lin,col="red")
```



Em seguida procede-se à obtenção e análise de valores associados aos parâmetros do modelo linear.

Para obter os I.C. dos parâmetros do modelo deve-se utilizar o comando *confint()*. Os valores referentes à variância dos parâmetros β_0 e β_1 podem ser determinados através da matriz de covariância dos mesmos parâmetros. Para obter a matriz é necessário utilizar a função *vcov()* através da sintaxe: *vcov(modelo).*

Tabela 6: *I.C. (95%) para os parâmetros do modelo linear obtido*

```
> confint(modelo1)
                 2.5 %       97.5 %
(Intercept)   -294.487     -224.7649
carats        3556.398    3885.6513
```

Tabela 7:*valores da matriz de covariâncias*

```
> vcov(modelo1)
                (Intercept)      carats
(Intercept)    299.9428      -1365.657
carats        -1365.6566      6688.930
```

Assim, a variância estimada para o parâmetro β_0 é 299.94 e para o parâmetro β_1 é 6688.93.

Para obter os intervalos de confiança para a resposta média é habitual utilizar a função *predict()*. O argumento base desta importante função consiste na designação do modelo que está a ser utilizado; na inclusão dos dados para previsão cujo nome deve ser identico à designação inicial da variável explicativa; no valor de confiança para os intervalos e no tipo de intervalo. Suponhamos que é necessário estimar o preço (*price*) de diamantes com 0.13, 0.14 e 0.24 quilates (*carats)*. O comando *predict()* deve ser utilizado com a seguinte sintaxe:

predict(modelo,data.frame(pred=newpred),level=0.95,interval=”confidence”)

em que *pred* é o objeto que contém as variáveis independentes originais e *newpred* é o objeto que contém os novos valores sobre os quais se pretende obter as estimativas, e *level* é o intervalo de confiança ao qual se atribui o valor pretendido.

Tabela 8:*Valores estimados e respetivos I.C.(95%)*

```
> newcarats<-c(0.13,0.14,0.24)

> predict(modelo1,data.frame(carats=newcarats),level=0.95,interval="confidence")

       fit          lwr        upr

1  224.1073   208.7888   239.4258

2  261.3176   247.2760   275.3592

3  633.4201   622.4484   644.3917
```

Na coluna *fit* podemos observar os valores correspondentes à estimação da variável resposta *price* para os novos valores de variável explicativa *carats* e nas colunas *lwr* e *upr* são apresentados os limites inferior e superior de cada valor estimado.

Para obter apenas os valores ajustados pelo modelo linear em função dos valores da variável "carats" dados inicialmente e os respetivos limites do I.C. pode utilizar-se o comando *predict(modelo linear,interval="conf").*

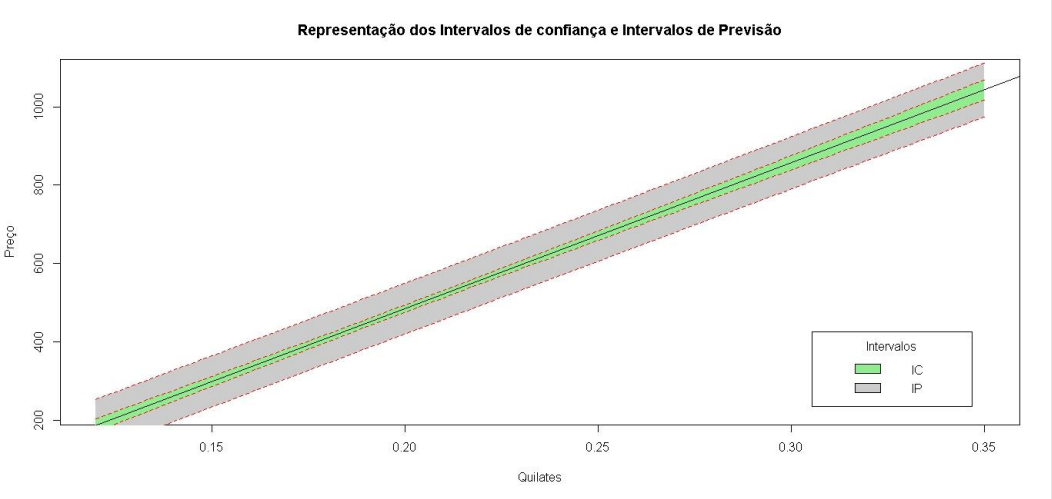
Tabela 9: *Valores ajustados e C.I. (95%)*

```
> ajust<-predict(modelo1,interval="conf")
> ajust
       fit             lwr          upr
1   372.9483    362.1217    383.7749
2   335.7381    323.9718    347.5043
3   372.9483    362.1217    383.7749
4   410.1586    400.0885    420.2286
…      ….            ….           ….
46  298.5278    285.6789    311.3768
47  707.8406    694.7997    720.8814
48  298.5278    285.6789    311.3768
```

Para uma apresentação mais completa é possível representar a reta de regressão e os limites, associados aos I.C. e aos Intervalos de previsão, para cada um dos valores ajustados, através da sintaxe descrita na Figura 5.

Figura 5: ***Sintaxe e*** *Gráfico sobre a reta de regressão e bandas de confiança e previsão*

```
x=dados1$carats

y=dados1$price

>plot(y~x,type="n",main="Representação dos Intervalos de confiança e Intervalos de Previsão",xlab="Quilates",ylab="Preço")

> newx <- seq(min(dados1$carats), max(dados1$carats), length.out=100)

> pred_s<-predict(modelo1,newdata=data.frame(carats=newx),interval="prediction")

> polygon(c(rev(newx), newx), c(rev(pred_s[ ,3]), pred_s[ ,2]), col = 'grey80', border = NA)

> lines(newx, pred_s[ ,2], lty = 'dashed', col = 'red')

> lines(newx, pred_s[ ,3], lty = 'dashed', col = 'red')

> preds<-predict(modelo1,newdata=data.frame(carats=newx),interval="confidence")

> polygon(c(rev(newx), newx), c(rev(preds[ ,3]), preds[ ,2]), col = 'lightgreen', border = NA)

> lines(newx, preds[ ,2], lty = 'dashed', col = 'red')

> lines(newx, preds[ ,3], lty = 'dashed', col = 'red')

> abline(modelo1)

    > legend("bottomright",inset=.05,title="Intervalos",c("IC","IP"),fill=c("lightgreen","grey80"))
```



A análise de variância do modelo linear é realizada através do comando *anova()* utilizando a sintaxe *anova(modelo).*

A estatística de teste F obtida é utilizada na avaliação das hipóteses sobre o parâmetro β_1 em que $H_0 : \beta_1 = 0$ e $H_1 : \beta_1 \neq 0$. Neste caso, a estatística F e o *pvalue* indicam que a hipótese nula deve ser rejeitada. De referir, que os valores das estatísticas de teste apresentadas neste *output* (Tabela 10) também são apresentados no *output* obtido a partir do comando *summary()*.

Tabela 10:*Output sobre tabela ANOVA do modelo linear*

```
> anova(modelo1)
Analysis of Variance Table
Response: price
            Df     Sum Sq    Mean Sq     F value       Pr(>F)
carats       1    2098596    2098596     2070      < 2.2e-16 ***
Residuals 46       46636         1014
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Antes de considerar e aceitar os resultados do modelo linear obtido como válidos para a inferência estatística, é importante avaliar a sua adequação aos dados através da análise de resíduos. Uma das várias formas de o fazer é analisar graficamente os resíduos e os valores ajustados. Se os resíduos forem aleatórios e tiverem uma distribuição normal então é válido considerar adequado o modelo linear obtido. Através do ambiente R é possível obter os principais gráficos para a análise de resíduos através da função *plot()*. Antes de escrever a sintaxe correspondente é necessário definir a janela gráfica para aceitar os 4 gráficos que, por predefinição, resultam da função *plot(modelo)*. A redefinição da janela gráfica pode ser feita de várias formas. Apresento em seguida, os comandos e sintaxes de duas dessas formas:

par(mfrow=c(2,2))

em que (2,2) significa duas linhas por duas colunas;

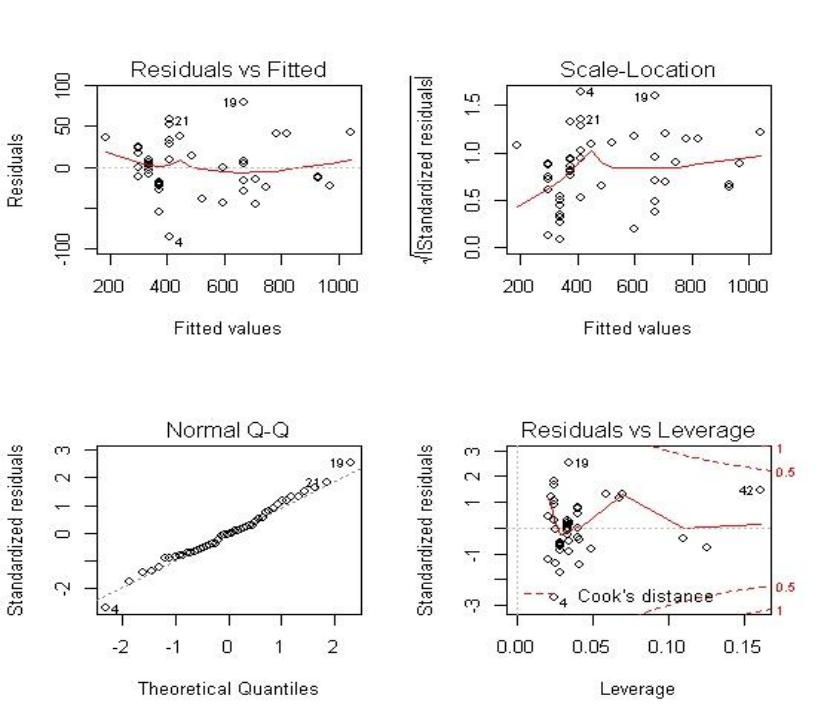
layout(matrix(1:4,2,2))

em que o *layout* da folha gráfica é definido por uma matriz 2×2, isto é, com quatro elementos.

No canto superior esquerdo é apresentado o gráfico de resíduos e valores ajustados. Os resíduos apresentam uma distribuição aleatória em torno da reta horizontal que representa o erro zero. Esta é a distribuição que os resíduos devem apresentar, ou seja, sem qualquer comportamento padronizado. O gráfico no canto inferior esquerdo é habitualmente denominado "papel de probabilidade" em português.

Figura 6: *Gráficos referentes à Análise de resíduos*

```
> layout(matrix(1:4,2,2))

> plot(reg_lin)
```



Neste exemplo, o gráfico sugere que os resíduos seguem uma distribuição normal devido ao seu alinhamento com a reta representada. Esta reta está relacionada com a frequência acumulada dos valores de uma distribuição normal para as abcissas. No canto superior direito, o gráfico "*scale location*" apresenta os valores ajustados contra a raiz dos valores absoluto dos resíduos. Tal como no primeiro caso, os pontos não devem apresentar nenhum padrão na sua distribuição. Finalmente, o gráfico no canto inferior direito, que

pode ser denominado como “Resíduos vs Alavancagem”, apresenta a distância de Cook, em que distâncias curtas sugerem que a remoção da observação correspondente terá pouco efeito no modelo linear e, distâncias superiores a um podem sugerir a presença de um valor aberrante.

Apesar da função *plot(modelo)* fornecer um conjunto de gráficos bastante completo para a análise de resíduos, existem outros comandos que podem ser úteis na análise e representação de resíduos e valores ajustados pelo modelo linear.

Sobre os resíduos é possível obter as principais medidas de localização central e também a sua distribuição através de um histograma. A sintaxe utilizada e os dados obtidos sobre os resíduos do modelo linear estão assinalados na tabela seguinte.

Tabela 11:*comandos e resíduos do modelo linear*

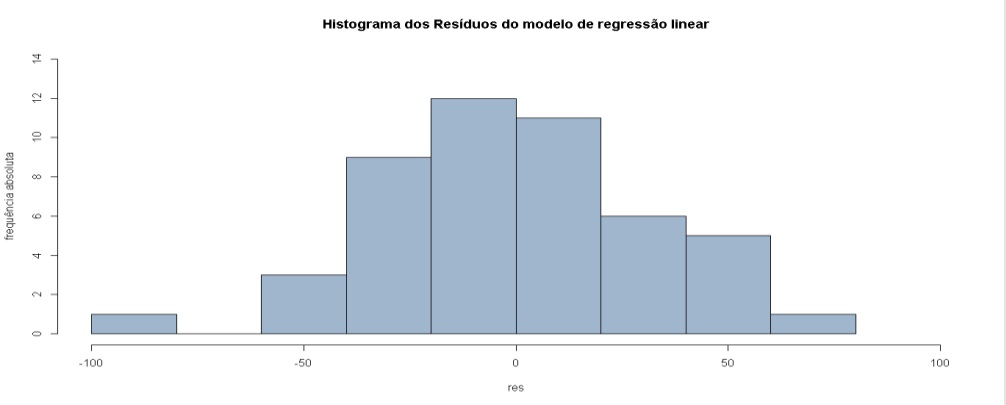
```
> res<-resid(modelo1)
> summary(res)
   Min.       1st Qu.   Median     Mean     3rd Qu.      Max.

-85.1600   -21.4500  -0.8688    0.0000    18.9700    79.3700

> res=data.frame(res)
> res
         res
1   -17.9483176

2     -7.7380691

3   -22.9483176

…  ……..……

46 -11.5278205

47 -14.8405542

48  17.4721795
```

Através da observação do histograma apresentado na Figura 7 pode verificar-se que a distribuição dos dados sobre os resíduos do modelo linear obtido segue uma distribuição aproximadamente normal.

Figura 7: *Sintaxe e Histograma sobre os resíduos do modelo de RLS*

```
> hist(res,main="Histograma dos Resíduos do modelo de regressão linear",col="slategray3",xlim=c(-
```

100,100))



Para além da análise gráfica aos resíduos, a utilização de testes estatísticos como o teste de Durbin-Watson e o teste de Breush-Pagan deve ser implementada de forma a solidificar as conclusões sobre a homocedasticidade dos resíduos.

Para realizar o teste de Durbin-Watson é necessário instalar um *package* que não se encontra na base de dados inicial do software R. Esse ficheiro tem a denominação *lmtest*. Após a sua instalação, insere-se o ficheiro no ambiente R através da sintaxe *library(lmtest)*. Para a obtenção dos valores referentes ao teste de Durbin-Watson é necessário utilizar o comando *dwtest(modelo)* tal como pode ser observado na tabela seguinte.

Tabela 12:*Output sobre o teste de Durbin-Watson*

```
> dwtest(modelo1)

        Durbin-Watson test

data:  reg_lin

DW = 1.9944, p-value = 0.4874

alternative hypothesis: true autocorrelation is greater than 0
```

Para localizar os valores críticos d_L e d_U na Tabela 12, devemos considerar, $\alpha = \frac{\alpha^*}{2}$ em que α^* é o nível de significância do teste. Estes valores permitem a elaboração de uma tabela semelhante à Tabela 2.

Tabela 13:*intervalos para conclusão do teste de Durbin-Watson*

DW= 1.99	[0, *1.42*[	[*1.42* , *1.5*[	[*1.5* ,2.5[	[2.5 ,2.58 [	[2.58 ,4[
Decisão	Rejeitar H_0	Teste Inconclusivo	Não rejeitar H_0	Teste Inconclusivo	Rejeitar H_0

Verifica-se que a estatística DW encontra-se no intervalo [d_U ,4- d_U[. Assim, de acordo com a estatística de teste obtida no *output* não rejeitamos a independência dos resíduos para um nível de significância de 0.05 .

Um dos testes mais utilizados para testar a homocedasticidade dos resíduos é o teste de Breusch-Pagan e a para a sua realização em R também é necessário recorrer ao *package lmtest.* Através da sintaxe *bptest(modelo)* obtém-se o seguinte *output*.

Tabela 14:*Output sobre o teste de Breush-Pagan*

```
> bptest(modelo1)

        studentized Breusch-Pagan test

data:  reg_lin

BP = 0.4202, df = 1, p-value = 0.5168
```

O *p value* obtido indica-nos que não devemos rejeitar a hipótese nula do teste de hipóteses. Assim, conclui-se que os resíduos são independentes.

Então, o modelo obtido pode ser considerado estatísticamente válido para a realização de inferência estatística no contexto do dados inicialmente analisados.

CAPÍTULO 2

2.1 REGRESSÃO LINEAR MÚLTIPLA

No capítulo 1, sobre regressão linear simples, definiram-se os principais conceitos e técnicas para se analisar e utilizar a relação linear entre duas variáveis. Esta análise conduz a uma equação que pode ser utilizada para se estimarem valores de uma variável resposta de interesse Y dados valores de uma variável independente X associada. Por vezes, é necessário mais do que uma variável explicativa (regressora) para modelar a variável resposta de interesse.

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

Os conceitos e técnicas para a realização e análise das relações lineares entre uma variável de resposta e várias variáveis regressoras são uma extensão natural do que foi apresentado na secção da regressão linear simples. Contudo, como é previsível, os cálculos tornam-se mais complexos. No entanto, esta complexidade muitas vezes é contornada e superada pela utilização de *softwares* tais como o R.

Assim, na regressão linear múltipla assume-se que existe uma relação linear entre uma variável de resposta e k variáveis regressoras, ou explicativas, x_j com $j = 1,....,k$.

As condições subjacentes à regressão linear múltipla são análogas à regressão linear simples. Em suma, temos que:

-As variáveis regressoras são fixas ou não aleatórias;

-Para cada conjunto de valores de x_j há um subconjunto de valores de Y. Para a construção dos intervalos de confiança e dos testes de hipóteses deve poder assumir-se que estes subconjuntos seguem uma distribuição normal;

-As variâncias dos subconjuntos de Y são iguais;

-Os valores de Y são estatísticamente independentes.

As variáveis podem ser apresentadas da seguinte forma:

Tabela 15: *Representação das variáveis explicativas e dependente num modelo de RLM*

Y	x_1	x_2	...	x_k
y_1	x_{11}	x_{21}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}	...	x_{nk}

A forma geral da equação de Regressão Linear Múltipla é :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_i \qquad (2.1)$$

Em que :

Y é a Variável Dependente;

β_0 corresponde a um coeficiente técnico fixo, a um valor de base a partir do qual começa *Y;*

β_j, com $j = 1,...., k$ corresponde aos coeficientes técnicos relacionados com as Variáveis Independentes, representando a variação esperada na resposta Y para cada unidade de variação em x_j;

x_j são as Variáveis Independentes (ou regressoras)

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

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

onde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ e } \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{bmatrix}$$

Sobre a constituição das diferentes matrizes pode referir-se o seguinte:

-A Matriz Y, é o vetor coluna $(n \times 1)$ constituído pelas observações da variável resposta;

-Na matriz $X_{n\times(k+1)}$, as linhas são constituídas pelos valores das variáveis independentes com $i = 1,...,n$ e $j = 1,...,k$;

-A matriz $\beta_{(k+1)\times 1}$ é o vetor coluna dos coeficientes de regressão;

-A matriz $\varepsilon_{n\times 1}$ é o vetor coluna dos erros aleatórios.

2.2. Método dos mínimos Quadrados

O método dos mínimos quadrados, tal como referido anteriormente, tem como objetivo encontrar o vetor $\hat{\beta}$ que minimiza

$$SQE = L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta) = Y^T Y - Y^T X\beta - \beta^T X^T X\beta =$$
$$= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

Sendo que $\beta^T X^T Y$ é um escalar e consequentemente, a sua transposta $Y^T X\beta$ tem o mesmo valor. O estimador dos mínimos quadrados $\hat{\beta}$ será a solução das seguintes equações:

$$\frac{\partial L}{\partial \beta} = 0 \Leftrightarrow X^T X\hat{\beta} = X^T Y$$

Assim, sabendo que, em geral, a matriz $X^T X$ é invertível, os estimadores para os parâmetros β_j, com $j = 1,...,k$, são dados pelo vetor

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{2.3}$$

Logo, o modelo de regressão linear ajustado e o vetor de resíduos são, respetivamente

$$\hat{Y} = X\hat{\beta} \tag{2.4}$$

e

$$e = Y - \hat{Y} \tag{2.5}$$

2.3. Propriedades dos Estimadores

Pelo Teorema de Gauss-Markov temos que o estimador de mínimos quadrados $\hat{\beta}$ é não enviesado e tem variância mínima entre todos os estimadores não enviesados que são combinações lineares dos Y_i (Weisberg, 2005).

i. Valor esperado de $\hat{\beta}$

$$\begin{aligned} E\left(\hat{\beta}\right) &= E\left[\left(X^T X\right)^{-1} X^T Y\right] = E\left[\left(X^T X\right)^{-1} X^T \left(X\beta + \varepsilon\right)\right] = \\ &= E\left[\left(X^T X\right)^{-1} X^T X\beta + \left(X^T X\right)^{-1} X^T \varepsilon\right] = E\left[I\beta\right] + E\left[\left(X^T X\right)^{-1} X^T \varepsilon\right] = \\ &= \beta + \left(X^T X\right)^{-1} X^T E\left[\varepsilon\right] = \beta \end{aligned}$$

ii. Matriz de covariâncias de $\hat{\beta}$

A definição matricial de variância consiste em considerar

$$\operatorname{cov}(W) = E\left[WW^T\right] - E[W]E(W)^T$$

em que W é um vetor de variáveis aleatórias. Então, tem-se que

$$cov(\hat{\beta}) = E[\hat{\beta}\hat{\beta}^T] - E[\hat{\beta}]E(\hat{\beta})^T =$$
$$= E\left\{\left[(X^TX)^{-1}X^TY\right]\left[(X^TX)^{-1}X^TY\right]^T\right\} - \beta\beta^T =$$
$$= (X^TX)^{-1}X^TE[YY^T]X(X^TX)^{-1} - \beta\beta^T =$$
$$= (X^TX)^{-1}X^T[Cov(Y) + E(Y)E(Y)^T]X(X^TX)^{-1} - \beta\beta^T =$$
$$= (X^TX)^{-1}X^TCov(Y)X(X^TX)^{-1} + (X^TX)^{-1}X^TE(Y)E(Y)^TX(X^TX)^{-1} - \beta\beta^T =$$
$$= \sigma^2(X^TX)^{-1}I + (X^TX)^{-1}X^TX\beta(X\beta)^TX(X^TX)^{-1} - \beta\beta^T =$$
$$= \sigma^2(X^TX)^{-1} + \beta\beta^T - \beta\beta^T =$$
$$= \sigma^2(X^TX)^{-1}$$

Em que $Cov(Y) = \sigma^2$ e $E(Y) = X\beta$.

iii. Estimador não viciado para σ^2

$$\hat{\sigma}^2 = QME = \frac{SQE}{n-k-1}$$

iv. Matriz de covariâncias estimada de $\hat{\beta}$

Seja $\hat{Cov}(\hat{\beta}) = \hat{\sigma}^2(X^TX)^{-1}$ e $C = \begin{bmatrix} C_{00} & & & \\ & C_{11} & & \\ & & \ddots & \\ & & & C_{kk} \end{bmatrix}$ a diagonal da matriz $(X^TX)^{-1}$.

Então, é possível escrever a variância estimada dos $\hat{\beta}_j$ como

$$\hat{\sigma}^2(\hat{\beta}_j) = \hat{\sigma}^2 C_{jj}$$

2.4. Análise de Variância

O objetivo inferencial consiste em avaliar se algumas das variáveis independentes podem ou não influenciar a variável de resposta. Isto é, se o modelo ajustado é ou não significativo. Esta hipótese teórica pode ser formalizada na forma:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \quad \text{vs} \quad H_1 : \exists\beta_j : \beta_j \neq 0$$

Considere-se que

$$SQT = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = Y^T Y - \frac{Y^T JY}{n} = Y^T \left(I - \frac{J}{n}\right) Y$$

em que $J = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$;

$$SQE = Y^T Y - \hat{\beta} X^T Y = Y^T \left(I - X\left(X^T X\right)^{-1} X^T\right) Y = Y^T \left(I - H\right) Y$$

em que:

$$H = X\left[X^T X\right]^{-1} X^T \tag{2.6}$$

Assim, tem-se que:

$$SQR = SQT - SQE = Y^T \left(H - \frac{J}{n}\right) Y \tag{2.7}$$

A variância explicada pela regressão é estimada dividindo *SQR* pelos respetivos graus de liberdade *k* (número de variáveis independentes no modelo) e a variância dos erros pode ser estimada dividindo a *SQE* pelos respetivos graus de liberdade *(n-k-1)*. Então, a estatística de teste é dada pelo quociente seguinte, que segue uma distribuição de Snedecor com *(k,n-k-1) g.l.:*

$$F = \frac{\frac{SQR}{k}}{\frac{SQE}{(n-k-1}} = \frac{QMR}{QME} \sim F_{(k,n-k-1)}$$

A tabela ANOVA da regressão com a estatística F é descrita na Tabela 16.

E quanto à estatística de teste F, H_0 é rejeitada se $F_0 > F_{(1-\alpha,k,n-k-1)}$ e se *pvalue* dado por $P\left[F_{(1-\alpha,k,n-k-1)} > F_0\right] < \alpha$, em que α é o nível de significância considerado.

Tabela 16: *Análise de variância sobre o modelo de RLM*

Fonte	G.L	Soma de Quadrados	Quadrado Médio	Teste F
Regressão	*k*	SQR	$QMR = \frac{SQR}{k}$	$F_0 = \frac{QMR}{QME}$
Resíduo	*n-k-1*	SQE	$QME = \frac{SQE}{n-k-1}$	
Total	*n-1*	SQT		

2.5. Coeficiente de determinação R^2

Tal como na regressão linear simples, a proporção de variabilidade em *Y* explicada pelos termos regressores é dada pelo quociente

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \tag{2.8}$$

Quanto mais próximo R^2 estiver de 1, maior é a explicação da variável de resposta pelo modelo ajustado.

Dado que o valor de R^2 aumenta à medida que são adicionados termos ao modelo, é prudente utilizar um valor ajustado de R^2 (C.Montegomery, 2009). Esse valor é definido da seguinte forma:

$$R_a{}^2 = 1 - \left(\frac{n-1}{n-k}\right)\left(1 - R^2\right) \tag{2.9}$$

2.6. Testes individuais para os coeficientes de Regressão

A necessidade de testar hipóteses sobre os coeficientes de regressão é justificada pela determinação da importância que cada um destes pode ter, ou não, no modelo a utilizar.

Adicionar uma variável ao modelo de regressão provoca um aumento na SQR e um decréscimo na SQE, e também aumenta a variância do valor ajustado $\hat{Y}$. As hipóteses para testar a significância de qualquer coeficiente de regressão individualmente são dadas por

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0 \ \text{com} \ \ j = 0,1,...,k$$

Se H_0 não é rejeitada, então x_j pode ser retirado do modelo dado que esta variável não influência a resposta de forma significativa. A estatística de teste para esta hipótese é

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \tag{2.10}$$

Com $\hat{\beta}_j \sim N_k\left(\beta; \sigma^2 C\right)$ em que $C = \left(X^T X\right)^{-1}$. Por vezes, também se encontra a denominação $se_{(\hat{\beta}_j)}$ para $\sqrt{\hat{\sigma}^2 C_{jj}}$.

A hipótese nula é rejeitada se $|t_0| > t_{\left(1-\frac{\alpha}{2}, n-k-1\right)}$.

2.7. Intervalos de confiança para os coeficientes de regressão e para valores de resposta previstos

Ao considerar a estatística *t0* apresentada anteriormente, um intervalo com $100(1-\alpha)\%$ de confiança para os coeficientes de regressão β_j , com $j = 0,1,...,k$, é dado por

$$\left[\hat{\beta}_j \mp t_{\left(\frac{\alpha}{2}, n-k-1\right)} \sqrt{\hat{\sigma}^2 C_{jj}}\right] \tag{2.11}$$

Um modelo de regressão pode ser utilizado para prever observações futuras da variável de resposta *y*. Seja $x_0^T = [1, x_{01}, x_{02}, ..., x_{0k}]$. A resposta esperada para este ponto é expressa na forma $\hat{y}(x_0) = x^T{}_0 \hat{\beta}$

O intervalo com confiança de $100(1-\alpha)\%$ para a resposta média $\hat{y}(x_0)$ é dado por

$$\left[\hat{y}(x_0) \mp t_{\left(\frac{\alpha}{2}, n-k-1\right)} \sqrt{\hat{\sigma}^2 x^T{}_0 \left(X^T X\right)^{-1} x_0}\right] \tag{2.12}$$

2.8. Análise de Resíduos

Para proceder á análise de resíduos do modelo de regressão linear múltipla utilizam-se vários testes, entre os quais, o de Durbin-Watson e Breusch-Pagan, já descritos anteriormente e também a análise dos diagramas "Resíduos vs Valores ajustados" e o "Papel de probabilidade".

2.9. Testes de ajustamento à Normalidade

2.9.1 Teste de Kolmogorov-Smirnov

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

As hipóteses a testar são:

H_0 : A amostra provém de uma distribuição normal (teórica);

H_1: A amostra não provém de uma distribuição teórica específica (distribuição normal), sendo, neste caso, uma distribuição *não normal* .

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

2.9.2 Teste de Shapiro-Wilk

O teste de Shapiro-Wilk determina uma variável estatística (W) calculada sobre os valores amostrais ordenados e elevados ao quadrado, com o objetivo de aferir se uma amostra aleatória x_i (com $i=1,\ldots,n$) provém de uma distribuição normal. Devido seu grande poder de resolução, este método tem sido profusamente adotado nos testes de normalidade utilizando-se preferencialmente em amostras de dimensão inferior a 30.

A variável W é calculada da seguinte forma:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.13)$$

em que x_i são os valores da amostra ordenados (x_1 é o menor). A constante b é determinada da seguinte forma:

$$b = \begin{cases} \sum_{i=1}^{\frac{n}{2}} a_{n-i+1} \times (x_{n-i+1} - x_i) \text{ se n é par} \\ \sum_{i=1}^{(n+1)/2} a_{n-i+1} \times (x_{n-i+1} - x_i) \text{ se n é ímpar} \end{cases}$$

em que a_{n-i+1} são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho *n* de uma distribuição Normal.

As hipóteses a testar são:

H_0: x_i tem distribuição normal

H_1: x_i não tem distribuição normal.

Rejeita-se a hipótese nula se o *pvalue* for inferior ao valor α previamente escolhido e assim há evidência de que os dados testados não provém de uma população com distribuição normal.

2.10. Comandos e sintaxes em R para a obtenção e análise de um modelo de regressão linear múltipla

A regressão linear múltipla e análise ao modelo obtido podem ser realizadas em R através de vário comandos (ou sintaxes) cuja identificação e descrição são propostas neste ponto através de um exemplo prático. Os dados utilizados neste exemplo foram retirados do site "The data and story library" com o link http://lib.stat.cmu.edu/DASL/Datafiles/enrolldat.html . Os dados foram copiados e colados num ficheiro de texto com a denominação *DadosRegrMultipla.txt*. Para inserir os dados em ambiente *R* pode utilizar-se o seguinte procedimento:

Localizar a diretoria em que se encontra o ficheiro e designar o objeto que será utilizado como tabela de dados, através da sintaxe:

```
> dados2<-read.table("C:/Users/Jonas/Desktop/DadosRegrMultipla.txt",header=T)
```

O ficheiro contém dados sobre cinco variáveis cuja identificação e definição podem ser observadas na tabela seguinte.

Tabela 17: *Descrição das variáveis do modelo MRLM*

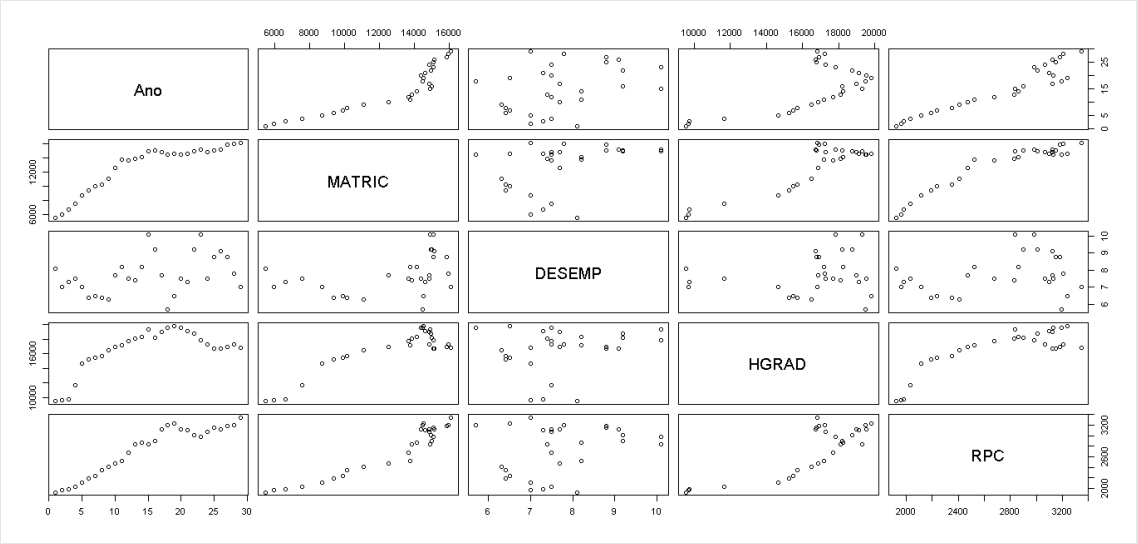
Abreviatura	***Variável***	***Unidades de medida***
MATRIC	*variável dependente contínua relativa ao número de matrículas em Bacharelatos*	
DESMP	*variável independente contínua que define a taxa de desemprego em Janeiro de cada ano registado.*	*Número real entre 0 e 1*
HGRAD	*variável independente contínua que representa o Número de conclusões de Highschool (nível secundário);*	
RPC	*variável independente que representa o rendimento per capita*	Milhares de dólares

Antes da obtenção do modelo de regressão linear múltipla é possível obter diagramas de dispersão para as várias variáveis em análise através da função *plot(dados)* ou *pairs(dados)* que podem ser observados na Figura 8.

A leitura desta matriz de diagramas de dispersão pode ser feita da seguinte forma: Na primeira coluna, a variável "Ano" está representada no eixo horizontal enquanto a variável MATRIC, na segunda linha, está representada no eixo vertical. As variáveis DESEMP, HGRAD e RPC estão representadas nas 3ª, 4ª e 5ª linhas respetivamente. A obtenção de um diagrama semelhante pode ser efetuada através do ficheiro *pschy* e utilizando o comando *Ipairs.panels(dados) ;* Ao utilizar este comando, obtém-se o diagrama e os respetivos valores dos coeficientes de correlação de *Pearson*.

Figura 8: *diagramas de dispersão que correlacionam as várias variáveis do exemplo*

```
>plot(dados2)
```



No capítulo 5, na parte 5.2, também será utilizado uma função denominada *pairs.panels()* que permite obter uma série de diagramas de dispersão sobre a distribuição das variáveis duas a duas. Esta função é inserida em ambiente R através do *package psych*. Além da visualição gráfica, também se obtém, em simultâneo, os valores de correlação entre as várias variáveis na mesma figura.

Após verificar a correlação entre as várias variáveis procede-se à determinação do modelo de regressão linear múltiplo. A sintaxe básica para obter o modelo de regressão é

lm(Y~modelo)

Onde *Y* é a variável de resposta e *modelo* é a fórmula correspondente ao modelo matemático determinado pelo investigador.

Neste caso a variável resposta é *MATRIC* e as variáveis explicativas são *DESEMP,HGRAD* e *RPC.* Pretende-se elaborar um modelo linear que permita obter e analisar o número de matrículas em cursos de nível de bacharelato a partir do valor da taxa de desemprego, número de alunos que concluíram o nível secundário e o rendimento *per capita*. Então, o modelo linear múltiplo é obtido através da sintaxe *lm(MATRIC~DESEMP+HGRAD+RPC,data=dados2*). A função *lm()* apenas fornece os coeficientes β_i sem qualquer informação estatística adicional. No exemplo sobre a

regressão linear simples verificou-se que é possível obter informação mais detalhada utilizando o comando *summary(modelo).*

Para uma utilização mais simples e eficaz do modelo criado é necessário criar um objeto identificado com o comando do modelo utilizado. Assim, podemos escrever:

modelo2<- lm(MATRIC~DESEMP+HGRAD+RPC,data=dados2)

em que o objeto *modelo2* pode e deve ser utilizado como argumento de outros comandos importantes relacionados com a regressão linear em ambiente R.

Através do output produzido e apresentado na Tabela 18 é possível verificar a informação acerca de medidas de dispersão e localização central sobre os resíduos. Quanto aos coeficientes do modelo linear, é possível verificar os valores dos coeficientes do modelo mas também as estatísticas de teste permitem não rejeitar ou rejeitar as hipóteses nula para os testes. Neste caso, os coeficientes determinados são evidentemente não nulos pois a estatística de teste apresenta valores inferiores a 0.05. No fundo da tabela estão assinalados os valores referentes ao desvio padrão do modelo, o coeficiente de regressão e a estatística *F* relativa ao modelo, e os respetivos graus de liberdade.

Neste *output* também é possível observar o valor do coeficiente de determinação ajustado R^2 denominado *adjusted R squared* . Neste caso, o valor obtido é 0,9576 ou seja, este modelo tem uma qualidade de ajuste de cerca de 96%. O coeficiente de determinação ajustado é mais adequado para modelos com várias variáveis pois o número de variáveis é utilizado no cálculo do valor final do coeficiente.

Então, o modelo linear obtido pode ser descrito através da expressão:

$$MATRIC= -9153+450\times DESEMP+0.4065\times HGRAD+4.275\times RPC$$

Em seguida procede-se à obtenção e análise de valores associados aos parâmetros do modelo linear.

Tabela 18:*Output para o modelo de regressão linear múltipla*

```
> summary(modelo2)

Call:

lm(formula = MATRIC ~ DESEMP + HGRAD + RPC, data = dados2)

Residuals:

     Min       1Q       Median     3Q       Max

-1148.84    -489.71     -1.88     387.40   1425.75

Coefficients:

                 Estimate       Std. Error    t value     Pr(>|t|)

(Intercept )    -9.153e+03     1.053e+03     -8.691    5.02e-09 ***

DESEMP           4.501e+02     1.182e+02      3.809    0.000807 ***

HGRAD            4.065e-01     7.602e-02      5.347    1.52e-05 ***

RPC              4.275e+00     4.947e-01      8.642    5.59e-09 ***

Multiple R-squared:  0.9621,    Adjusted R-squared:  0.9576

F-statistic: 211.5 on 3 and 25 DF,  p-value: < 2.2e-16
```

Para obter os I.C. dos parâmetros do modelo deve-se utilizar o comando *confint()*.

Tabela 19:*I.C.(95%) para os parâmetros* β_i

```
> confint(modelo2)

                    2.5 %              97.5 %

(Intercept)    -11322.321486      -6984.1874394

DESEMP           206.752768         693.4962390

HGRAD              0.249921           0.5630464

RPC                3.256059           5.2936565
```

Os valores referentes à variância dos parâmetros $\beta_0, \beta_1, \beta_2$ e β_3 podem ser determinados através da matriz de covariância dos mesmos parâmetros. Para obter a matriz é necessário utilizar a função *vcov()* através da sintaxe:

vcov(modelo)

Tabela 20:*Matriz de covariâncias*

> vcov(modelo2)

	(Intercept)	DESEMP	HGRAD	RPC
(Intercept)	***1109190.37295***	-83634.666002	-18.636985657	-51.37495594
DESEMP	-83634.66600	***13963.697300***	0.885986966	-14.20425583
HGRAD	-18.63699	0.885987	**0.005778788**	-0.03066991
RPC	-51.37496	-14.204256	-0.030669910	***0.24470172***

Assim, a variância estimada para os parâmetros $\beta_0, \beta_1, \beta_2$ e β_3 é, aproximadamente, 1109190, 13964, 0.0058 e 0.245 respetivamente.

Os intervalos de confiança para a resposta média podem ser obtidos utilizando o comando *predict()* da seguinte forma:

predict(modelo,level=0.95,interval="confidence")

Tabela 21:*Valores estimados e respetivos I.C.(95%)*

> predict(modelo2 level=0.95,interval="confidence")

	fit	lwr	upr
1	6596.038	5891.964	7300.112
2	6315.375	5637.174	6993.577
3	6548.091	5877.661	7218.521
…	….	…	…
27	15277.351	14780.417	15774.285
28	15071.306	14598.775	15543.838
29	15132.446	14427.442	15837.451

Na coluna *fit* podemos observar os valores correspondentes à estimação da variável resposta *MATRIC* e nas colunas *lwr* e *upr* são apresentados os limites inferior e superior do I.C. de cada valor estimado.

Os intervalos de previsão para o número de matrículas em função das variáveis explicativas dadas é calculado através da sintaxe

predict(modelo,level=0.95,interval="prediction")

Suponha-se que se pretendia estimar o intervalo de previsão do número de matrículas para um ano em que a taxa de desemprego seria de 10%, o número de alunos que concluíam o liceu seria de 20000 e o rendimento *per capita* seria de 3000 dólares. Então, o comando a utilizar e a respetiva sintaxe seria:

Tabela 22: *Intervalo de previsão*

```
> predict(modelo2,data.frame(DESEMP=10,HGRAD=20000,RPC=3000),interval="prediction")
       fit             lwr          upr
1 16302.24     14755.24     17849.24
```

Deve-se notar que a validade destes intervalos de previsão pressupõe a distribuição Gaussiana dos dados. A análise de variância do modelo linear é realizada através do comando *anova()* utilizando a sintaxe *anova(modelo).* A *ANOVA* para o modelo obtido pode ser observada na Tabela 23.

Tabela 23: *Output sobre ANOVA*

```
> anova(modelo2)
Analysis of Variance Table
Response: MATRIC
           Df     Sum Sq      Mean Sq      F value       Pr(>F)
DESEMP     1     45407767    45407767     101.02      2.894e-10 ***
HGRAD      1    206279143  206279143    458.92      < 2.2e-16 ***
RPC          1     33568255     33568255     74.68       5.594e-09 ***
Residuals 25     11237313      449493
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As principais conclusões que se podem obter a partir desta tabela já foram descritas a partir da Tabela 18 resultante do *output* realizado com o comando *summary(modelo).*

Antes de considerar e aceitar os resultados do modelo linear obtido é importante avaliar a sua adequação aos dados através da análise de resíduos. Uma das várias formas de o fazer é analisar graficamente os resíduos e os valores ajustados. Se os resíduos forem aleatórios e tiverem uma distribuição normal então é válido considerar adequado o modelo linear obtido. Através do ambiente R é possível obter os principais gráficos para a análise de resíduos através da função *plot()*. Antes de escrever a sintaxe correspondente é necessário definir a janela gráfica para aceitar os 4 gráficos que, por predefinição, resultam da função *plot(modelo).* A redefinição da janela gráfica pode ser feita de várias formas. Apresentam-se em seguida, os comandos e sintaxes de duas dessas formas:

par(mfrow=c(2,2))

em que (2,2) significa duas linhas por duas colunas;

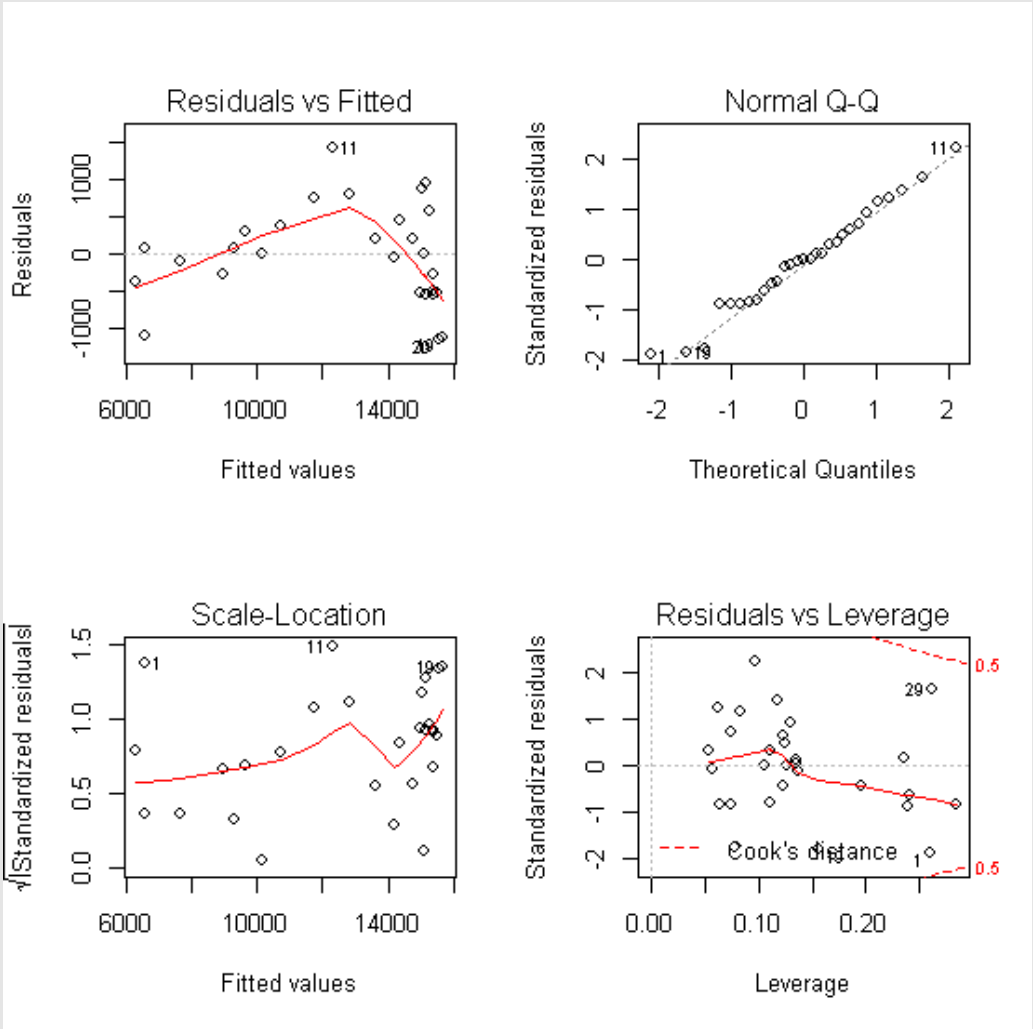
layout(matrix(1:4,2,2))

em que o *layout* da folha gráfica é definido por uma matriz 2×2, isto é, com quatro elementos.

Sobre os gráficos obtidos em R e apresentados na Figura 9 pode afirmar-se que os resíduos (vs valores ajustados) apresentam uma distribuição algo padronizada em torno da reta horizontal que representa o erro zero. O gráfico denominado "papel de probabilidade" sugere que os resíduos seguem uma distribuição normal devido ao seu alinhamento com a reta representada. Esta reta está relacionada com a frequência acumulada dos valores de uma distribuição normal para as abcissas. No canto superior direito, o gráfico "*scale location*" apresenta os valores ajustados contra a raiz dos valores absolutos dos resíduos. Tal como no primeiro caso, os pontos não devem apresentar nenhum padrão na sua distribuição. Finalmente, o gráfico no canto inferior direito, denominado como "Resíduos vs Alavancagem", apresenta a distância de Cook, em que distâncias curtas sugerem que a remoção da observação correspondente terá pouco efeito no modelo linear e, distâncias superiores a um podem sugerir a presença de um *outlier*. Neste caso, não é indicada nenhuma distância igual ou superior a 1. Apesar da função *plot(modelo)* fornecer um conjunto de gráficos bastante completo para a análise de resíduos, existem outros comandos que podem ser úteis na análise e representação de resíduos e valores ajustados pelo modelo linear.

Figura 9: *Gráficos sobre resíduos do modelo linear múltiplo*

```
> par(mfrow(2,2))

> plot(modelo2)
```



Sobre os resíduos é possível obter as principais medidas de localização central e também a sua distribuição através de um histograma. A sintaxe utilizada e os dados obtidos sobre os resíduos do modelo linear estão assinalados na tabela seguinte.

Tabela 24: *Output sobre a distribuição dos resíduos*

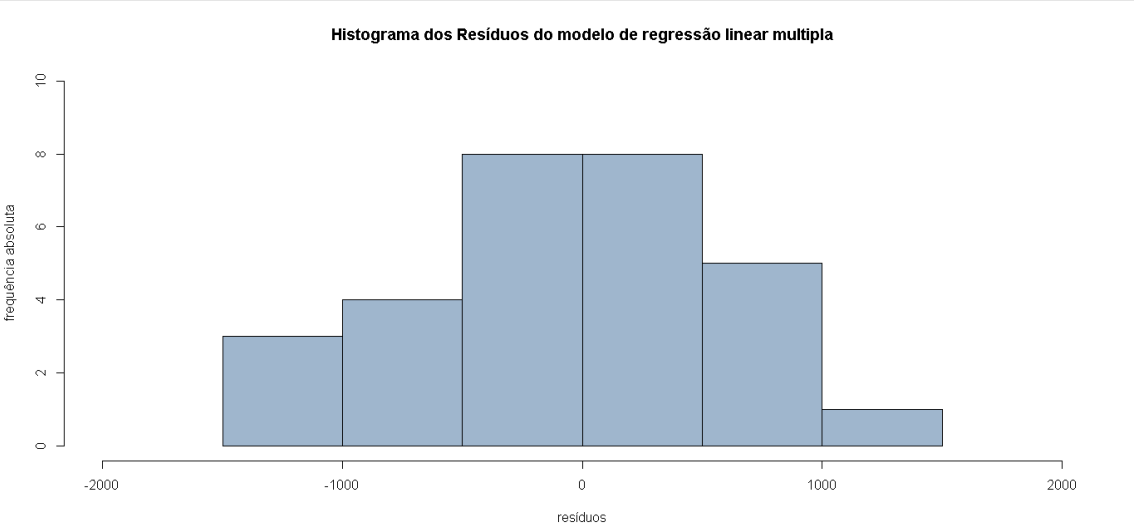
```
> res.modelo2<-resid(modelo2)
> summary(res.modelo2)

    Min.      1st Qu.   Median     Mean    3rd Qu.     Max.
-1149.000  -489.700    -1.876     0.000   387.400  1426.000
```

```
> res.modelo2=data.frame(res_m)
> res.modelo2

    Res.modelo2
 1 -1095.037937
 2  -370.375492
 3    80.909048
...   ..........
27   578.649118
28   866.693585
29   948.553573
```

Através da observação do histograma apresentado na Figura 10 pode verificar-se que a distribuição dos dados sobre os resíduos do modelo linear obtido segue uma distribuição aproximadamente normal.

Figura 10: *Histograma dos resíduos do MRLM*

```
> hist(res,main="Histograma dos Resíduos do modelo de regressão linear",col="slategray3",xlim=c(-2000,2000),ylab="frequência absoluta",xlab="resíduos")
```



Para além da análise gráfica aos resíduos, a utilização de testes estatísticos como o teste de Durbin-Watson e o teste de Breush-Pagan deve ser implementada de forma a solidificar as conclusões sobre a homocedasticidade dos resíduos.

Um dos testes mais utilizados para testar a homocedasticidade dos resíduos é o teste de Breusch-Pagan e a para a sua realização em R também é necessário o ficheiro *lmtest.* Através da sintaxe *bptest(modelo)* obtém-se o seguinte *output*. O *p value* obtido (0,8224)

para um nível de confiança de 0,05 indica-nos que não devemos rejeitar a hipótese nula do teste de hipóteses. Assim, conclui-se que os resíduos são independentes.

Tabela 25: *Output sobre os valores do teste de Breusch Pagan*

```
> bptest(reg_lin_m)

      studentized Breusch-Pagan test

data:  reg_lin_m
BP = 0.91268, df = 3, p-value = 0.8224
```

Logo, o modelo de regressão linear múltipla denominado "modelo2" cumpre os pressupostos necessários para a validação estatística do modelo e para a realização de inferência estatísticamente válida.

CAPÍTULO 3

3. REGRESSÃO LOGÍSTICA

A regressão logística é um modelo linear generalizado. Muitas vezes é necessário analisar casos em que a variável dependente é discreta, e não contínua. No modelo logístico a variável resposta Y_i é binária. Uma variável binária assume dois valores, habitualmente, $Y_i = 0$ e $Y_i = 1$ que podem se denominados "fracasso" e "sucesso", respetivamente. Neste caso, "sucesso" é o evento de interesse.

3.1. REGRESSÃO LOGÍSTICA SIMPLES

Seja x_i a variável explicativa e y_i o número de ocorrências de um dado evento, em que $i = 1,2,...,n$ representa o número de observações. Assume-se ainda, que a variável resposta tem distribuição binomial $(Y_i \sim B(\pi_i))$ com $\pi_i = E(Y_i)$. Assim,

$$P[Y_i = y_i] = \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \qquad (3.1)$$

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

$$\pi(x_i) = \frac{e^{\beta_0 + \beta_i x_i}}{1 + e^{\beta_0 + \beta_i x_i}}, i = 1,...,n \qquad (3.2)$$

Quando $\beta_1 < 0$, π é crescente e quando $\beta_1 > 0$, π é decrescente. Quando x tende para valores infinitamente grandes , $\pi(x)$ tende a zero (quando $\beta_1 < 0$) e tende para um (quando $\beta_1 > 0$). Caso $\beta_1 = 0$, a variável de resposta *Y* é independente da variável *X*.

Assim, desta forma, define-se a função de ligação necessária ao modelo logístico. Esta transformação é chamada de transformação *logit* de probabilidade $\pi(x)$, que pode ser escrita como

$$g(x) = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_i x_i \qquad (3.3)$$

Esta transformação é importante pois permite à função de ligação *g(x)* manter propriedades associadas ao modelo de regressão linear (Hosmer & Lemeshow, 2000).

De acordo com (1.1.) , ε significa o desvio de uma observação em relação à média condicional. A hipótese mais comum é que este ε segue uma distribuição Normal com média zero e variância constante, ao longo dos vários níveis da variável independente. Daqui, resulta que esta distribuição condicional da variável independente dado o valor da variável *x,* segue uma distribuição normal, com média $E[Y|x]$e variância constante. No entanto, tal não se verifica quando se está perante uma variável resposta dicotómica. Assim, nesta situação, segundo os autores Hosmer & Lemeshow, deve-se expressar o valor da variável como:

$$y = \pi(x) + \varepsilon$$

3.2 Estimação dos parâmetros β_0 e β_1.

Para estimar os parâmetros β_0 e β_1 da expressão 2.2. é utilizado o método da máxima verossimilhança que, de uma forma geral, fornece valores para os parâmetros desconhecidos que maximizam a probabilidade de se obter determinado conjunto de valores. Assumindo que $(x_0, y_0),...,(x_n, y_n)$ são independentes, a função de verossimilhança tem a seguinte forma:

$$\begin{aligned} & P[Y_i = y_1,...y_n \mid \beta_0, \beta_1] = \\ & = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} = \\ & = \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i} \right)^{y_i} (1-\pi_i)^{1-y_i} \end{aligned} \tag{3.4}$$

Após logaritmizar os dois membros da expressão obtém-se:

$$L(\beta_0, \beta_1 \mid (x_i, y_i)) = \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln(1 + e^{\beta_0 + \beta_1 x_i})$$

Os estimadores de máxima verossimilhança para os parâmetros β_0 e β_1 são os valores $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximizam o logaritmo da função de verossimilhança.

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

$$\frac{\partial}{\partial \beta_0} L(\beta_0, \beta_1 \mid (x_i, y_i)) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \qquad (3.5)$$

$$\frac{\partial}{\partial \beta_1} L(\beta_0, \beta_1 \mid (x_i, y_i)) = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \qquad (3.6)$$

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

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

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

Para resolver estas equações é necessário recorrer a métodos iterativos numéricos que não são abordados neste trabalho. Os resultados decorrentes da aplicação destes métodos são incluídos na matriz denominada Informação de Fisher com a seguinte forma:

$$I(\hat{\beta}) = \begin{bmatrix} \sum_{i=1}^{n} \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{\left(1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}\right)^2} & \sum_{i=1}^{n} x_i \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{\left(1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}\right)^2} \\ \sum_{i=1}^{n} x_i \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{\left(1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}\right)^2} & \sum_{i=1}^{n} x^2{}_i \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{\left(1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}\right)^2} \end{bmatrix} \qquad (3.7)$$

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

$$\pi_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_i x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_i x_i}} \qquad (3.8)$$

Os parâmetros β_0 e β_1 tem significados semelhantes aos análogos na regressão linear. Neste caso, β_1 é o incremento no valor da expressão (3.2) devido ao aumento de uma unidade em *x*. E β_0 corresponde a "log odds" de "sucesso" contra fracasso no caso em que *x=0.*

Seja $g(x) = \frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x}$. Ao tomar dois valores distintos da variável x com a diferença de uma unidade, x_j e x_{j+1} temos:

$$OR = \frac{g(x_{j+1})}{g(x_j)} = \frac{e^{\beta_0+\beta_1 x_{j+1}}}{e^{\beta_0+\beta_1 x_j}} \tag{3.9}$$

Temos ainda que:

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

Sabendo que a diferença entre as variáveis explicativas é de uma unidade, então:

$$ln(OR) = ln(e^{\beta_1}) = \beta_1 \tag{3.10}$$

Assim, temos o quão provável o resultado ocorrerá entre os indivíduos x_{j+1} em relação aos indivíduos x_j, fazendo, portanto, algumas análises:

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

3.2.1. Estimativa do desvio padrão

As variâncias e covariâncias dos estimadores $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1\right)$ são obtidos, invertendo a matriz de informação de Fisher, isto é, calculando $I^{-1}\left(\hat{\beta}\right)$.

O j-ésimo elemento da diagonal principal da matriz $I^{-1}\left(\hat{\beta}\right)$ é a variância do estimador $\hat{\beta}_j$ denominada $\hat{\sigma}^2\left(\hat{\beta}_j\right)$Os demais elementos da matriz são as covariâncias entre $\left(\hat{\beta}_j, \hat{\beta}_u\right)$ com $j \neq u$. Assim, o desvio padrão é definido como:

$$D\hat{P}\left(\hat{\beta}_j\right) = \sqrt{\hat{\sigma}^2\left(\hat{\beta}_j\right)} \tag{3.11}$$

3.2.2. Teste de Wald

O teste de Wald é utilizado para avaliar se um parâmetro é estatísticamente significativo. A estatística teste utilizada é obtida através da razão do coeficiente pelo seu respetivo erro padrão. Esta estatística de teste tem distribuição Normal, em que o seu valor é comparado com valores tabulados de acordo com o nível de significância definido. A estatística teste, para avaliar se o parâmetro b é igual a zero, é assim especificada:

$$W = \frac{\hat{\beta}}{\sqrt{var(\hat{\beta})}} \sim N(0,1) \tag{3.12}$$

3.2.3.Teste da razão de verossimilhança

Na regressão logística é necessário comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável em questão. A comparação dos observados com os valores preditos é baseado no logaritmo da verossimilhança. Para entender melhor esta comparação, é necessário pensar num valor observado da variável resposta também como sendo um valor predito resultante de um modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quanto observações.

A comparação dos observados com os valores preditos usando a função de verossimilhança é baseada na seguinte expressão:

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

Com o objetivo de assegurar a significância de uma variável independente, comparamos o valor de *D* com e sem a variável na equação. A mudança em *D* devido a inclusão da variável no modelo é obtida da seguinte maneira:

$$G = D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável)$$

Podemos então escrever a estatística G como:

$$G = -2ln(L_S) + 2ln(L_C) \qquad (3.13)$$

em que L_S é a verossimilhança do modelo sem a covariável e L_C é a verossimilhança do modelo com a covariável. As hipóteses a testar são

$$H_0: \beta_1 = 0 \ \ vs\ H_1: \ \beta_1 \neq 0$$

Sob a hipótese nula, a estatística G tem distribuição chi-quadrado com 1 grau de liberdade.

3.2.4 Teste Score

A estatística do teste Score é

$$TS = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\left(\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}}} \qquad (3.14)$$

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

No teste Score, segundo a distribuição assintótica Qui-Quadrado, as hipóteses a testar são:

$$H_0 : \beta_1 = 0 \, vs \, H_1 : \beta_1 \neq 0$$

3.3. Intervalos de Confiança

3.3.1 Intervalos de confiança para os parâmetros

A elaboração das estimativas do intervalo de confiança para os parâmetros tem por base a mesma teoria estatística que é utilizada para os testes de significância do modelo. Em particular, os intervalo de confiança para a inclinação e intercepto são baseados nos respetivos testes de Wald.

O intervalo de confiança ao nível de confiança $100(1-\alpha)\%$ para o parâmetro β_1 é:

$$IC_{\beta_1,1-\alpha} = \left[\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} DP\left(\hat{\beta}_1\right) \right] \tag{3.15}$$

E, para o intercepto β_0, é:

$$IC_{\beta_{01},1-\alpha} = \left[\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} DP\left(\hat{\beta}_0\right) \right] \tag{3.16}$$

Em que $z_{1-\frac{\alpha}{2}}$ é o valor critico da distribuição normal padrão correspondente a 100(1-α/2)%

3.3.2 Intervalo de confiança para o *logit*

O intervalo de confiança para *logit* é:

$$IC_{\hat{g}(x),1-\alpha} = \left[\hat{g}(x) \pm z_{1-\frac{\alpha}{2}} DP\left(\hat{g}(x)\right) \right] \tag{3.17}$$

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

3.3.3 Intervalo de Confiança para os valores ajustados

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

$$IC_{\pi,1-\alpha} = \left[\frac{e^{\hat{g}(x) \mp z_{(1-\alpha/2)} DP(\hat{g}(x))}}{1 + e^{\hat{g}(x) \mp z_{(1-\alpha/2)} DP(\hat{g}(x))}} \right] \tag{3.18}$$

3.3.4 Intervalo de Confiança para *Odds Ratio*

Sejam β_I e β_S os limites inferior e superior respetivamente, do $IC_{\beta_1,1-\alpha}$.

Assim, o intervalo de confiança para a *Odds Ratio* é:

$$IC_{OddsRatio,1-\alpha} = \left[e^{\beta_I}, e^{\beta_S} \right] \tag{3.19}$$

3.4 Comandos e sintaxes em R para a obtenção e análise de um modelo de regressão Logística Simples

Os comandos (ou sintaxe) a utilizar para realizar a regressão Logística Simples com todos os elementos indicados no capítulo 3 serão apresentados e descritos em seguida.

Os dados utilizados neste exemplo foram retirados do site "Machine Learning Repository" através do link https://archive.ics.uci.edu/ml/machine-learning-databases/space-shuttle/o-ring-erosion-only.data. Os dados foram copiados e inseridos num ficheiro de texto com a denominação *Nasa.txt*. Para inserir os dados em ambiente *R* pode utilizar-se o seguinte procedimento:

Localizar a diretoria em que se encontra o ficheiro e designar o objeto que será utilizado como tabela de dados, através da sintaxe:

> dados3 <-read.table("C:/Users/Jonas/Desktop/Nasa.txt",header=T)

O ficheiro contém dados sobre duas variáveis cuja definição pode ser observada na tabela ..

Tabela 26: *Descrição das variáveis para o MLG simples*

Abreviatura	***Variável***	***Unidades de medida***
Temp	*Variável independente contínua definida pelos vários registos de temperaturas;*	Graus centigrados
Fail	*variável independente categórica que designa a ocorrência ou não de falha.*	0-Não houve falha; 1-Houve falha

Antes da obtenção do modelo de regressão Logística Simples é possível obter um gráfico que permita a visualização da distribuição dos dados para as duas variáveis em análise através do comando *plot(dados)* que podem ser observados na Figura 11.

No software R não existe uma função específica para ajustar um modelo de regressão logística, e o motivo é simples: a regressão logística é apenas um caso de *modelo linear generalizado* (MLG), ou GLM em inglês. Os MLGs são ajustados no software R através da função *glm*, onde devemos especificar a formula (a definição do modelo) e *family* (a

distribuição assumida pela variável resposta com a função de ligação a ser usada). Isto é, Neste tipo de modelo especifica-se apenas a distribuição do erro e a função de ligação . A sintaxe básica para obter o modelo de regressão é

$$glm(Y \sim modelo, family = binomial(link = 'logit')))$$

Onde *Y* é a variável de resposta e *modelo* é a fórmula correspondente ao modelo matemático determinado pelo investigador. Assim, a sintaxe a utilizar e o respetivo *output* pode ser observados na Tabela 27.

Figura 11: *Registo Gráfico sobre Temperaturas e ocorrências de falhas*

```
plot(main="Temperaturas registadas e ocorrências",xlab="Temperatura",ylab="Falha no anel",fail~temp,data=dados3,col="navy",pch=19)
```

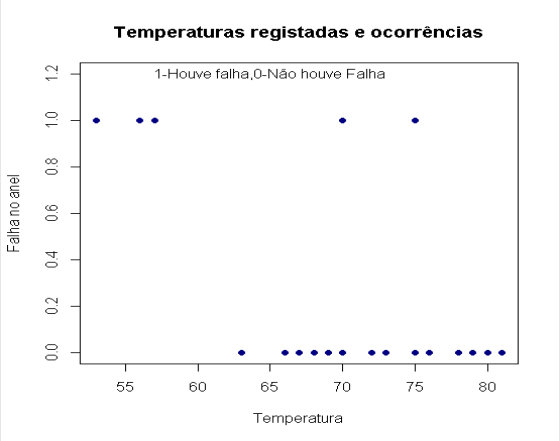


Tabela 27:*Sintaxe e Resumo do modelo logístico simples*

```
> modelo3 = glm(fail ~ temp, data = dados3, family = binomial(link = 'logit'))
> summary(modelo)
Call:
glm(formula = fail ~ temp, family = binomial(link = "logit"),
   data = dados_rls)

Deviance Residuals:
   Min      1Q      Median      3Q      Max
-1.2125  -0.8253  -0.4706     0.5907   2.0512

Coefficients:
              Estimate     Std. Error      z value      Pr(>|z|)
(Intercept)   10.87535       5.70291        1.907      0.0565 .
temp          -0.17132       0.08344       -2.053      0.0400 *
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance:  28.975  on 23  degrees of freedom
Residual deviance:  23.030  on 22  degrees of freedom
AIC: 27.03
Number of Fisher Scoring iterations: 4
```

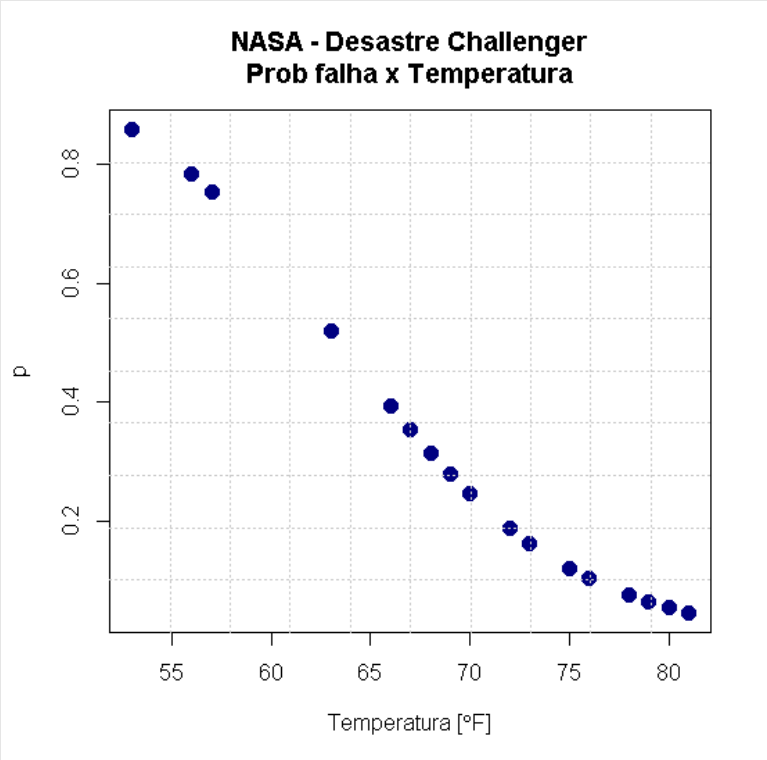
Dado que a variável de interesse é binária especificamos uma distribuição binomial e usamos a função logística como link (que, por *default*, surge associada ao modelo *glm*). Pelo *output* observa-se que a temperatura possui efeito significativo sobre a possibilidade de falhas (*p value ~ 0.04*). Observando os coeficientes do modelo, podemos escrever

$$ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 x = 10.87535 - 0.17132 * T$$

Onde π é a probabilidade de falha e T a temperatura. A análise visual do modelo obtido pode ser feita através do gráfico apresentado na Figura 12.

Figura 12:*Sintaxe e Gráfico do MLG simples*

```
> temp = dados_rls$temp
> p = modelo3$fitted.values
> xlb = expression(paste('Temperatura [', degree,'F]'))
> plot(temp, p,  col = 'navy', pch = 19, cex = 1.5, ylab = 'p', xlab = xlb, main = titulo)
> grid(10, 10, col = '#CCCCCC')
```



Ao observar o gráfico, verifica-se que é notória a influência da temperatura sobre a possibilidade de falha. Logo, ao variarmos a temperatura em *1 grau*, as chances de falha variam em $e^{-0.17132} = 0{,}8425515$. Ou seja, quando a temperatura aumenta *1 grau*, as chances de falha são reduzidas cerca de 16%. Esse valor é denominado de *odds ratio*, e é uma constante característica do modelo, que também pode ser calculado através de uma alteração dos coeficientes do modelo através da sintaxe descrita na Tabela 28.

Sobre a significância dos coeficientes do modelo é possível efetuar os testes (*3.12*) ,(*3.13)* e (*3.14)* em ambiente R.

Tabela 28:*exponenciação dos coeficientes do modelo*

```
> exp(modelo3$coefficients)
   (Intercept)          temp
5.285720e+04   8.425515e-01
```

Uma das formas de realizar o teste de Wald é através do *package aod* que será necessário instalar. Após a sua instalação e carregamento em ambiente R efetua-se o teste através do comando descrito na tabela seguinte.

Tabela 29:*Sintaxe e output do Teste de Wald*

```
> wald.test(b=coef(object=modelo3), Sigma=vcov(object=modelo3), Terms=2)
Wald test:
----------
Chi-squared test:
X2 = 4.2, df = 1, P(> X2) = 0.04
```

De acordo com os valores obtidos (*p-value=0.04)*, a hipótese nula é rejeitada indicando que a variável Temperatura tem influência no modelo obtido. Este teste surge no *output* obtido para o modelo logístico representado na Tabela 27.

Os modelos MGL são ajustados aos dados pelo método de máxima verossimilhança, proporcionando não apenas estimativas dos coeficientes de regressão, mas também estimando erros padrões dos coeficientes. O teste da razão de verossimilhança tem por objetivo de assegurar a significância de uma variável independente e a sua estatística G

obtida através da diferença $D(\text{modelo sem a variável}) - D(\text{modelo com a variável})$. Em R, a estatística G pode ser obtida da seguinte forma:

Tabela 30: *Cálculos para teste de verosimilhança*

```
Modelo3$deviance
[1] 23.03045
> modelo3$null.deviance
[1] 28.97459
> G2=modelo3$null.deviance-modelo3$deviance
> G2
[1] 5.944137
> 1-pchisq(G2,df=1)
[1] 0.01476632
```

Este teste também pode ser efetuado através de uma análise de variância utilizando o comando *anova(modelo, test="chisq")*. A sintaxe utilizada neste exemplo e o respetivo output estão descritos na tabela seguinte.

Tabela 31: *Análise de variância para o modelo logístico simples*

```
> anova(modelo3,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: fail
Terms added sequentially (first to last)
          Df   Deviance Resid.  Df        Resid.Dev      Pr(>Chi)
NULL                            23          28.975
temp      1          5.9441     22           23.030       0.01477 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Também é possível realizar o teste de razão de verossimilhança através da função *drop1()*, utilizando a sintaxe *drop1(modelo,test="chi")*.

Assim, sabendo que a estatística G tem distribuição chi-quadrado com 1 grau de liberdade, rejeitamos a hipótese nula e conclui-se que a variável Temperatura tem influência sobre a variável dependente.

Ainda sobre a significância do modelo determinado, é necessário fazer uma referência ao parâmetro AIC obtido no *output* . AIC, ou critério de *Akaike*, é uma ferramenta para seleção de modelos, pois oferece uma medida relativa quanto à qualidade do ajuste de um modelo estatístico. Este não se apresenta na forma de um teste de um modelo no sentido usual de testar uma hipótese nula, ou seja, o AIC não pode indicar nada sobre o quão bem o modelo ajusta os dados num sentido absoluto.

Na sua forma geral, AIC é dado por

$$AIC = 2K - 2\ln(L)$$

onde *k* é o número de parâmetros no modelo estatístico, e *L* é o valor maximizado da função de verossimilhança para o modelo estimado. Dado um conjunto de modelos candidatos, o modelo preferível é aquele com o valor mínimo de AIC.

Este indicador será desenvolvido na aplicação de R em modelos de regressão logística múltipla.

Para a determinação de Intervalos de confiança para os parâmetros β_1 e β_0 definidos por 3.14 e 3.15 é necessário utilizar a função *confint()* . Para este exemplo a sintaxe utilizada é:

confint(modelo)

E os resultados obtidos estão na tabela seguinte.

Tabela 32: *I.C. (95%) para os parâmetros do modelo logístico simples*

```
> confint(modelo3)
Waiting for profiling to be done...
                2.5 %       97.5 %
(Intercept)  1.1941770   24.9888102
temp        -0.3779407   -0.0305637
```

A leitura e interpretação dos valores dos parâmetros torna-se mais clara quando se procede à sua transformação através da exponenciação de base *e.*

Tabela 33*: I.C. (95%) para as Odds*

```
> exp(confint(modelo3))
Waiting for profiling to be done...
                 2.5 %          97.5 %
(Intercept)  3.3008401   7.120367e+10
temp         0.6852711   9.698986e-01
```

De notar que a função *confint.default()* utilizada através do *package MASS* gera I.C.´s através da estatística de Wald enquanto que a função *confint()* gera os valores para os limites dos I.C.´s através da verossimelhança.

A determinação dos I.C. para os valores ajustados neste modelo podem ser obtidos através do procedimento descrito a seguir. De notar que, tal como em muitos outros cálculos realizados em R, este é apenas uma das várias formas de o fazer.

O primeiro passo consiste em criar uma série de valores de temperaturas para introduzir na função resultante do modelo logístico através da sintaxe

predict.dados=53:81

O resultado obtido é uma série de 29 números naturais compreendidos entre 53 e 81. Estes valores são, respetivamente, o mínimo e máximo dos dados iniciais. Entre estes valores constam várias temperaturas para as quais não existe um registo inicial. A utilização da função associada ao modelo logístico deste exemplo pode ser realizada com recurso a um comando denominado *plogis()*. Assim, não é necessário criar um atributo à função logística obtida. Para a obtenção dos valores estimados é necessário associar o conjunto de dados criado à variável preditora *temp* do modelo logístico. Então, os valores estimados para o desvio padrão da função, associada ao modelo a partir dos dados criados, são obtidos através da função *predict()*. Estes valores estimados para o desvio padrão são necessários para o cálculo dos I.C. tal como pode ser observado na expressão 3.17. Na tabela seguinte pode observar-se os comandos e sintaxe utilizadas.

Tabela 34: *Valores ajustados, intervalos de predição e I.C. (95%)*

```
y = plogis(modelo3$coefficients[1] + modelo$coefficients[2] *predict.dados)

new.temp = data.frame(temp= predict.dados)

y.estim = predict(modelo3, new.temp , type = "link", se.fit = TRUE)

Sup_logit=y.est$fit+1.96*y.est$se.fit

Inf_logit=y.est$fit-1.96*y.est$se.fit

IC_S=plogis(Sup_logit)

IC_I=plogis( Inf_logit)
```

No entanto, para obter os valores ajustados sobre a probabilidade esperada para cada uma das temperaturas criadas é necessário utilizar a sintaxe:

$$y.estim2 = predict(modelo3, xy, type = "response", se.fit = TRUE)\ ,$$

em que se altera o parâmetro *type* de *link* para *response*.

Para apresentar os valores dos I.C. para os valores ajustados (3.18) e de previsão em tabela podemos realizar a sintaxe que se encontra na tabela seguinte.

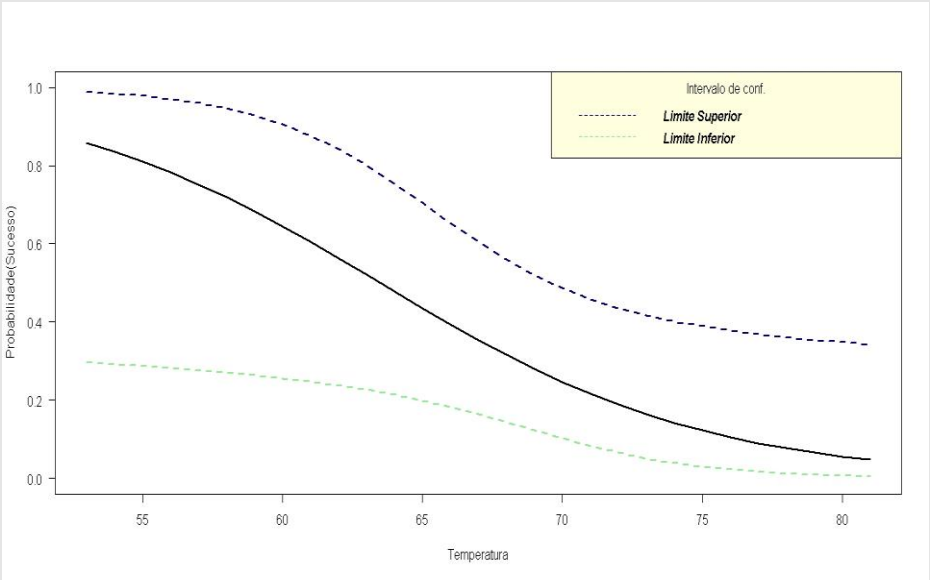
Tabela 35:*Criação de tabela de valores ajustados e I.C. (95%)*

```
> valores_ajust=y.estim2$fit
> Tabela<-data.frame(IC_I,valores_ajust,IC_S)
> Tabela
        IC_I          valores_ajust    IC_S
1  0.295421097   0.85758349     0.9885689
2  0.291097448   0.83535183     0.9842981
3  0.286456132   0.81041694     0.9785028
4  0.281437267   0.78268818     0.9706919
          ....          ....          ....
27 0.008883624   0.06543827     0.3535849
28 0.006498443   0.05570909     0.3473045
29 0.004738335   0.04735313     0.3416606
```

Também é possível fazer uma representação gráfica dos valores ajustados pelo modelo e dos I.C. calculados anteriormente (Figura 13).

Figura 13: *sintaxe para obtenção de gráfico de valores ajustados e I.C. e Gráfico sobre a representação* **do***s Valores ajustados pelo modelo e I.C. (95%)*

```
>plot(predict.dados,y,ylim=c(0,1),type="l",lwd=2,ylab="Probabilidade(Sucesso)",xlab="Temperatura",xaxt="
n",las=1)

> lines(predict.dados,IC_S,lty=2,lwd=2,col="navyblue")

> lines(predict.dados,IC_I,lty=2,lwd=2,col="lightgreen")

> axis(1)
>legend ("topright", legend = c("Limite Superior" , "Limite Inferior"), lty=c(2,2),
col=c("navyblue","lightgreen"), title="Intervalo de conf.", text.font=4, bg='lightyellow')
```



O cálculo dos I.C: para a *Odds Ratio* do modelo determinado segundo a definição 3.19 em R pode ser feito, sem recurso a qualquer comando ou função, da seguinte forma:

Tabela 36:*Cálculo dos I.C. (95%) para Odds Ratio*

```
> BI#limite inferior do IC do parâmetro B1

[1] -0.3779407

> BS#limite superior do IC do parâmetro B1

[1] -0.0305637

> OR_LI=exp(BI)#Limite inferior do IC para Odds Ratio

[1] 0.6852711

> OR_LS=exp(BS)#Limite superior do IC para Odds Ratio

[1] 0.9698986
```

Ou utilizando os comandos *exp, coef, confint* e *cbind* da seguinte forma

exp(cbind(Odds_Ratio_failVstemp=coef(modelo), confint(modelo)))

O output resultante pode ser observado na tabela seguinte.

Tabela 37:*Cálculo dos Odds Ratio e respetivos I.C.*

	Odds_Ratio_failVstemp	2.5 %	97.5 %
(Intercept)	5.285720e+04	3.3008401	7.120367e+10
temp	8.425515e-01	0.6852711	9.698986e-01

Todos os elementos definidos no capítulo 3 foram exemplificados e obtidos em R produzindo um MLG simples e respetivos parâmetros. Estes parâmetros foram testados e verificou-se a sua validade estatística em relação ao modelo obtido. Os I.C. definidos anteriormente foram calculados e os I.C. para os valores ajustados foram representados graficamente. A qualidade de ajuste do modelo não foi tratada porque não foi definida nesta parte.

CAPÍTULO 4

4.1 REGRESSÃO LOGISTÍCA MÚLTIPLA

Tal como no modelo de regressão linear, podemos ajustar um modelo logístico para a variável resposta com mais de uma variável explicativa, que é denominado modelo de Regressão Logística Múltipla. Seja *X* um conjunto sob a forma de vetor com *p* elementos tal que $X = (X_1, X_2, ..., X_p)$. A probabilidade condicional da ocorrência de um evento é representada pela definição $P(Y = 1 / X) = \pi(X)$ e a função de ligação ou *logit* é dado, respetivamente, por

$$E[Y] = \pi(X) = \frac{e^{g(X)}}{1 + e^{g(X)}} \tag{4.1}$$

e

$$g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{4.2}$$

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

$$D = \begin{cases} 0, se\ a\ caracteristica\ não\ estiver\ presente \\ 1, se\ a\ caracteristica\ estiver\ presente \end{cases}$$

4.2. Estimativas dos parâmetros do modelo e respetivos desvio-padrão

Para obter as estimativas dos componentes do vetor $\beta = (\beta_0, \beta_1, ..., \beta_p)$ correspondente aos parâmetros do modelo será utilizado o método da máxima verossimilhança. Então, temos que

$$L(\beta_0, \beta_1, ..., \beta_p \mid (x_i, m_i, y_i)) = \sum_{i=1}^{n} \left[y_i g(X) - ln(1 + e^{g(X)}) \right] \quad (4.3)$$

Em que *g(X)* é a expressão (4.2) e (x_i, y_i) são dados observados.

Derivando (4.3), igualando a zero e substituindo pelos estimadores dos parâmetros obtêm-se as seguintes equações:

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

As soluções destas equações, obtidas através de métodos numéricos, fornecem as estimativas para os parâmetros do modelo de regressão logística múltipla. Os processos iterativos de obtenção de soluções são análogos aos do modelo de regressão linear simples descritos no capítulo anterior.

As variâncias e covariâncias dos coeficientes são estimadas de acordo com teoria da estimação de máxima verossimilhança. Essa teoria assegura que os estimadores são obtidos a partir da matriz de segundas derivadas parciais da função log de verossimilhança. Estas derivadas parciais têm a seguinte forma geral:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \quad (4.4)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \quad (4.5)$$

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

Seja a matriz *(p+1)x(p+1)* que contém os termos negativos das derivadas parciais (4.2) e (4.3) e denotada por *I(β)* (ou Matriz de Informação de Fisher). As variâncias e covariâncias dos coeficientes estimados são obtidos a partir da inversa da matriz *I(β)*. Assim, $Var(\beta) = I^{-1}(\beta)$. A variância de β_j, é o *j-ésimo* elemento da diagonal da matriz e a

$Cov(\beta_j, \beta_l)$ é obtida através do elemento da matriz referente à linha de β_j e a coluna de β_l. Os estimadores das variâncias e covariâncias, $\hat{Var}(\hat{\beta})$ são obtidos de $Var(\beta)$ em $\hat{\beta}$.

A matriz de informação de Fisher estimada pode ser obtida por $\hat{I}(\hat{\beta}) = X^T V X$, em que

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} \quad \text{e} \quad V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}_{n\times n}$$

Então, o desvio padrão do coeficiente β_j é

$$\hat{\sigma}(\hat{\beta}_j) = \sqrt{\hat{Var}(\hat{\beta}_j)}$$

4.4. Avaliação da significância dos parâmetros

O método para avaliar a significância dos parâmetros é análogo ao modelo logístico simples.

4.4.1 Teste da Razão de Verossimilhança

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

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

$$D = -2ln(L_S) + 2ln(L_C) \qquad (4.6)$$

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

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

4.4.2 Teste de Wald

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

$H_0 : \beta_j = 0 \;\; vs \;\; H_1 : \beta_j \neq 0, j = 0,..., p$.

A estatística de teste é dada por

$$W_j = \frac{\hat{\beta}_j}{\hat{\sigma}\left(\hat{\beta}_j\right)} \tag{4.7}$$

De forma equivalente, o teste de Wald também pode ser obtido pela multiplicação dos seguintes vetores:

$$W = \hat{\beta}^T \left(X^T V X \right) \hat{\beta}$$

Com distribuição Qui-quadrado e *p+1* g.l. sob a hipótese que cada um dos *p+1* coeficientes é igual a zero.

4.5. Intervalos de Confiança

4.5.1 Intervalos de confiança para os parâmetros

O intervalo de confiança para um parâmetro β_j é baseado no respetivo teste de Wald. O intervalo de confiança de *100(1-α)%* para o parâmetro β_j é:

$$IC\left(\beta_j, 1-\alpha\right) = \left[\hat{\beta}_j \mp z_{1-\alpha} \sigma\left[\left(\hat{\beta}_j\right)\right]\right] \tag{4.8}$$

4.5.2 Intervalos de Confiança para *logit*

O estimador função (4.2) pode ser descrito matricialmente da seguinte forma:

$$\hat{g}(x) = x^T \hat{\beta} \tag{4.9}$$

em que $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ designa o estimador para os $p+1$ coeficientes do modelo e o vetor $x^T = (x_0, x_1, ..., x_p)$ representam a constante e um conjunto de valores das p covariáveis no modelo, em que $x_0 = 1$.

Assim, o estimador para a variância de (4.8) é:

$$\hat{Var}[\hat{g}(x)] = \sum_{j=0}^{p} x_j^2 \hat{Var}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \hat{Cov}(\hat{\beta}_j, \hat{\beta}_k) \tag{4.10}$$

ou

$$\hat{Var}[\hat{g}(x)] = x^T \left(X^T V X \right)^{-1} x \tag{4.11}$$

Então, o I.C. para o *logit* é

$$IC(\hat{g}(x), 1-\alpha) = [\hat{g}(x) \mp z_{1-\alpha/2} \sigma[(\hat{g}(x))]] \tag{4.12}$$

4.5.3 Intervalo de confiança para os valores ajustados

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

$$IC(\pi, 1-\alpha) = \left[\frac{e^{\hat{g}(x) \mp z_{1-\alpha/2} \sigma(\hat{g}(x))}}{1 + e^{\hat{g}(x) \mp z_{1-\alpha/2} \sigma(\hat{g}(x))}} \right] \tag{4.13}$$

4.5.4 Intervalo de confiança para *Odds Ratio*

Sejam β_I e β_S os limites inferior e superior respetivamente, do $IC_{\beta_1, 1-\alpha}$. Assim, o intervalo de confiança para a *Odds Ratio* é:

$$IC_{OddsRatio, 1-\alpha} = \left[e^{\beta_I}, e^{\beta_S} \right] \tag{4.14}$$

4.6 Análise ao modelo

Em qualquer modelo de regressão, é necessário proceder à análise dos resíduos para validação da qualidade do modelo estimado. Assim, pretende-se avaliar quais as ”distâncias” entre os valores observados e os valores estimados. Existem diversas medidas de modo a detetar diferenças significativas entre os valores observados e os valores estimados. Existem dois tipos de resíduos possíveis que poderão ser utilizados para avaliar a qualidade do ajustamento: os resíduos de Pearson e os resíduos da *Deviance*.

4.6.1. Resíduos de Pearson

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

$$r\left(y_i,\hat{\pi}_j\right)=r_j=\frac{y_i-\hat{\pi}_j}{\sqrt{\hat{\pi}_j\left(1-\hat{\pi}_j\right)}},\ j=1,2,...,n \tag{4.15}$$

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

$$\chi^2=\sum_{j=1}^{n} r\left(y_i,\hat{\pi}_j\right)^2 \tag{4.16}$$

com distribuição Qui-Quadrado para *n-(p+1)* g.l., em que *p* é o número de covariáveis do modelo, para testar a hipótese H_0:”O modelo ajustado é adequado”.

4.6.2. *Deviance*

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

$$d\left(y_j,\hat{\pi}_j\right)=\pm\left[2\left[y_j ln\left(\frac{y_j}{\hat{\pi}_j}\right)+\left(1-y_j\right)ln\left(\frac{1-y_j}{1-\hat{\pi}_j}\right)\right]\right]^{1/2} \tag{4.17}$$

A estatística de teste a utilizar é

$$D=\sum_{j=1}^{n} d\left(y_j,\hat{\pi}_j\right)^2 \tag{4.18}$$

A estatística D, sob a suposição inicial H_0 que o modelo ajustado é correto, tem distribuição assintótica χ^2 com $n-(p+1)$ g.l. , em que p é o número de covariáveis do modelo.

4.6.3. Teste de Hosmer-Lemeshow

O teste de Hosmer-Lemeshow é largamente utilizado na regressão logística com o objetivo de testar a qualidade do ajuste, isto é, o teste comprova se o modelo obtido pode explicar adequadamente os dados observados. Este teste baseia-se na divisão dos dados de acordo com as probabilidades previstas. As observações são separadas em g grupos de acordo com as probabilidades previstas. O número mais utilizado para g é 10. Assim, o primeiro grupo consiste nos elementos cuja probabilidade prevista é inferior a 10%. O segundo grupo é constituído pelos elementos cujas probabilidades previstas estão compreendidas entre 10% e 20%. A realização deste processo é realizado até formar o 10º grupo. Antes do cálculo da estatística teste, é necessário estimar a frequência esperada dentro de cada grupo, para tal é necessário dividir a variável resposta, que é dicotômica. Para $Y=1$, a frequência esperada estimada resulta da soma das probabilidades estimadas de todos os indivíduos dentro daquele grupo. Para $Y=0$, a frequência esperada estimada é obtida através da soma de 1-probabilidade estimada de todos os indivíduos dentro daquele grupo. A estatística de teste é dada por

$$C = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{e_k\left(1 - \frac{e_k}{n_k}\right)} \tag{4.19}$$

Em que :

$o_k = \sum_{j=1}^{n_k} y_{kj}$ são o número de casos registados no *k-ésimo* decil;

$e_k = \sum_{j=1}^{n_k} \hat{\pi}_{kj}$ são o número esperado de casos no *k-ésimo* decil;

y_{kj} e $\hat{\pi}_{kj}$ correspondem aos valores previstos e às probabilidades previstas para a observação *j* no grupo *k* de decil de risco.

O teste rejeita a capacidade de ajuste do modelo quando $C > \chi^2_{8,1-\alpha}$, em que $\chi^2_{t,1-\alpha}$ é o quantil *1-α* da distribuição Qui-Quadrado com *t* g.l.

4.7. Predição (Curva ROC)

Quando a variável resposta é binária é necessário determinar uma regra de predição, dado que a probabilidade estimada $\hat{\pi}$ está compreendida entre 0 e 1. Poderá ser intuitivo supor que se $\hat{\pi}_i$ apresentar um valor de próximo de 1, deveremos considerar que $\hat{Y}_i = 1$ e, se $\hat{\pi}_i$ corresponder um valor pequeno ou próximo de zero, deveremos considerar que $\hat{Y}_i = 0$. Mas como determinar o ponto que para os valores acima dele o indivíduo é classificado como "evento" ($\hat{Y}_i = 1$) e valores abaixo dele o indivíduo é classificado como "não evento" ($\hat{Y}_i = 0$)? Esse ponto é denominado como ponto de corte.

Uma forma bastante utilizada para determinar o ponto de corte é através da Curva ROC (*Receiver Operating Characteristic Curve*). Geometricamente, a curva ROC é um gráfico de pares " $1-P\left(\hat{Y}_i = 0 \mid Y = 0\right)$" e " $P\left(\hat{Y}_i = 1 \mid Y = 1\right)$" representados num plano designado por plano ROC unitário. A designação de plano ROC unitário, deve-se ao facto das coordenadas deste gráfico representarem medidas de probabilidade, e por conseguinte variarem entre zero e um.

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

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

Precisão, Sensibilidade, Especificidade, Verdadeiro Preditivo Positivo e Verdadeiro Preditivo Negativo. A relação entre estes vários conceitos pode ser compreendida a partir da tabela de contingência que se apresenta a seguir.

Tabela 38: *Tabela de contingência 2X2 ou Matriz de Confusão*

	Valor Observado		
Valor Estimado		*Y=1*	*Y=0*
	$\hat{Y} = 1$	*VP*	*FP*
	$\hat{Y} = 0$	*FN*	*VN*

Assim, as definições dos conceitos acima citados são:

-***Precisão:*** proporção de predições corretas, sem considerar o que é positivo e o que é negativo mas sim o acerto total e é dada por:

$$ACC = \frac{VP + VN}{P + N}$$

em que P é o número total de eventos (Y=1, chamado aqui de positivo) e N é o número total de não eventos (Y=0, chamado aqui de negativo).

-***Sensibilidade***: proporção de verdadeiros positivos. Isto é, a avaliação da capacidade do modelo em classificar um indivíduo como evento $\hat{Y} = 1$ dado que realmente ele é evento (*Y=1*):

$$SENS = \frac{VP}{VP + FN}$$

-***Especificidade***: Proporção de verdadeiros negativos. Isto é, a avaliação da capacidade do modelo predizer um indivíduo como não evento $\hat{Y} = 0$ dado que ele realmente é não evento *(Y=0*).

$$ESPEC = \frac{VN}{VN + FP}$$

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

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

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

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

Uma forma bastante generalizada de aferir a capacidade discriminante de um modelo através da curva ROC é através do cálculo da área representada abaixo da curva que em inglês é designado pela abreviatura *auc*. Dado que esta área é uma parte da área quadrada unitária, o seu valor está compreendido entre 0 e 1. Dado que a determinação aleatória de eventos produz o segmento de reta entre (0,0) e (1,1) que divide o quadrado unitário em duas áreas de 0,5 , segundo (Fawcett, 2005) nenhum classificador realista deve ter uma *auc* inferior a 0,5.

4.8. Comandos e sintaxes em R para a obtenção e análise de um modelo de regressão logística múltipla

As funções e respetivas sintaxes a utilizar para realizar a regressão logística múltipla com todos os elementos descritos anteriormente serão apresentados em seguida. Os dados utilizados neste exemplo foram retirados do site da UCLA (Universidade da Califórnia em Los Angeles). Os dados foram introduzidos na consola de comandos do R através da seguinte sintaxe:

dados4 <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")

O ficheiro contém dados sobre 4 variáveis, codificadas comoRank, Gre, Gpa e admit, sendo descritas na Tabela 39.

Tabela 39:*Descrição das variáveis para o modelo de regressão logística múltipla*

Abreviatura	***Variável***	***Unidades de medida***
Rank	*variável independente relativa à graduação da escola de proveniência do candidato*	Escala de 1 a 4
Gre	*variável independente contínua que define a pontuação obtida pelo candidato em exames prévios*	
Gpa	*variável independente contínua que define a pontuação obtida pelo candidato em exames prévios*	
admit	*variável dependente*	0 (não admitido) e 1 (admitido)

Este ficheiro contem um total de 400 observações cuja origem não é identificada.

Antes de obter o modelo logístico é necessário analisar os dados e verificar se não existem valores nulos que, por defeito, estão codificados como *NA*. Uma das formas de o fazer é utilizar a sintaxe *Is.na(dados4).* Em seguida procede-se à observação da tabela resultante. Se todas as células tiverem preenchidas com o termo *FALSE* significa que não existe nenhum valor em falta.

Tabela 40:*localização de valores em branco NA*

```
> is.na(dados4)
      admit      gre   gpa    rank
 [1,] FALSE   FALSE   FALSE FALSE
 .....     .....     .....      .....       ....
```

A variável independente *rank* é categórica e está codificada entre 1 e 4 mas o programa não assume essa definição previamente. Assim, é necessário converter a variável de forma a que os valores registados sejam tratados como categorias. Para tal utiliza-se a seguinte sintaxe:

dados4$rank <- factor(dados4$rank)

Para obter o modelo de regressão logística múltipla recorre-se à função *glm*. Este modelo será denominado *modelo4*. A sintaxe a utilizar é:

modelo4 <- glm(admit~rank+gre+gpa,family=binomial(logit),data=dados4)

O output relativo ao modelo obtido através do comando *summary()* está disposto na tabela seguinte.

Tabela 41*:output relativo ao modelo de regressão logística múltipla*

```
> modelo4<-glm(admit~rank+gre+gpa,family=binomial(logit),data=dados4)
> summary(modelo4)
Call:
glm(formula = admit ~ rank + gre + gpa, family = binomial(logit),
   data = dados4)

Deviance Residuals:
   Min       1Q     Median      3Q       Max
-1.6268   -0.8662  -0.6388    1.1490   2.0790

Coefficients:
              Estimate       Std. Error    z value    Pr(>|z|)
(Intercept)  -3.989979     1.139951    -3.500    0.000465 ***
```

```
rank2       -0.675443   0.316490  -2.134  0.032829 *
rank3       -1.340204   0.345306  -3.881  0.000104 ***
rank4       -1.551464   0.417832  -3.713  0.000205 ***
gre          0.002264   0.001094   2.070  0.038465 *
gpa          0.804038   0.331819   2.423  0.015388 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

Sobre os valores estimados dos coeficientes presentes no *output* pode-se concluir que todos são estatísticamente significativos para o modelo pretendido. Note-se que todos os *p-value* são inferiores a um nível de confiança de 0,05, logo a hipótese nula que está subjacente à criação de um modelo de regressão, que é definida pela falta de preponderância das variáveis explicativas no modelo, em relação a todas as variáveis, não é aceite. Assim, não será necessário proceder à análise de vários modelos. Sobre os valores estimados dos coeficientes obtidos pode-se concluir que:

Para cada incremento de uma unidade na variável *gre* as chances *log* (ou *odds ratio*) de admissão aumentam para $e^{0,002}$ que é aproximadamente 1.

Para cada incremento de uma unidade na variável *gpa* as chances *log* de admissão aumentam para $e^{0,804}$;

A variável *rank*, dada a sua caracterização, pode ter uma interpretação ligeiramente diferente. Por exemplo, ao comparar um estudante que tenha frequentado uma escola de rank1 com um estudante que tenha frequentado uma escola de rank3, as chances *log* deste em ser admitido é $e^{-1,3402}$ em comparação com o primeiro.

Para o cálculo de probabilidade de admissão é necessário recorrer à função de ligação *logit*.

Essa função é definida por $\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}$, em que

$$g(x) = -3{,}9900 - 0{,}6754 rank2 - 1{,}3402 rank3 - 1{,}5515 rank4 + 0{,}002 gre + 0{,}8040 gpa.$$

Já foi referido que os coeficientes estimados são estatísticamente significativos para o modelo obtido. Os valores apresentados *output* provém da estatística de teste de Wald. No entanto, o teste de Wald pode ser realizado separadamente recorrendo ao pacote *aod*. Neste caso, pode ter relevância estudar e analisar os coeficientes para os diferentes valores da variável *rank*. Assim, o teste consistirá na verificação de igualdade entre os coeficientes das categorias 3 e 4 da variável *rank*. Para tal, é criado um vetor *l* definido pelos valores (0,0,1,-1,0,0). O vetor tem valores 1 e -1 na posição 3 e 4 respetivamente porque que as categorias 3 e 4 ocupam as terceira e quarta posições no modelo obtido. O programa procederá à multiplicação dos valores do vetor *l* pelos coeficientes do modelo e apenas comparará as variáveis não nulas.

Tabela 42:*Output sobre o Teste de Wald*

```
> l<-cbind(0,0,-1,1,0,0)
> wald.test(b = coef(modelo4), Sigma = vcov(modelo4), L=l)
Wald test:
----------

Chi-squared test:
X2 = 0.29, df = 1, P(> X2) = 0.59
```

Ao observar os resultados obtidos, verifica-se a obtenção de um valor de 0,29 na estatística Qui Quadrado, com um *g.l.* associado a um *p-value* de 0,59. Assim, conclui-se que não há evidência estatística sobre a diferença entre os coeficientes associados às categorias 3 e 4 da variável *rank*.

O teste de Wald para a significância dos parâmetros também pode ser realizado através do pacote *car* utilizando a função *anova()*.

Tabela 43: *Teste Wald (car)*

```
> library(car)

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

> Anova(modelo4, type="II", test="Wald")
Analysis of Deviance Table (Type II tests)
```

```
Response: admit
     Df    Chisq     Pr(>Chisq)
rank  3   20.8953    0.0001107 ***
gre   1    4.2843    0.0384651 *
gpa   1    5.8715    0.0153879 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A única diferença em relação ao teste anterior está na variável *rank*. Neste teste, o programa apresentou apenas um valor de teste para todas categorias da variável em análise.

Para a obtenção de intervalos de confiança para os coeficientes estimados dos parâmetros recorre-se ao comando *confint()*.

Tabela 44:*I.C. (95%) para parâmetros*

```
> cbind(Estimate=coef(modelo4), confint(modelo4))
Waiting for profiling to be done...
                Estimate          2.5 %          97.5 %
(Intercept) -3.989979073  -6.2716202334  -1.792547080
rank2       -0.675442928  -1.3008888002  -0.056745722
rank3       -1.340203916  -2.0276713127  -0.670372346
rank4       -1.551463677  -2.4000265384  -0.753542605
gre          0.002264426   0.0001375921   0.004435874
gpa          0.804037549   0.1602959439   1.464142727
```

Tal como foi referido no exemplo sobre regressão logística simples, muitas vezes é necessário exponenciar os valores obtidos de forma a calcular os valores de *odds ratio*. Na tabela seguinte podem observar-se os valores exponenciados das estimativas para os parâmetros e os respetivos I.C´s.

Tabela 45:*I.C. (95%) para parâmetros*

```
> exp(cbind(OR = coef(modelo4), confint(modelo4)))
Waiting for profiling to be done...
                 OR          2.5 %         97.5 %
(Intercept) 0.0185001   0.001889165   0.1665354
rank2       0.5089310   0.272289674   0.9448343
rank3       0.2617923   0.131641717   0.5115181
rank4       0.2119375   0.090715546   0.4706961
gre         1.0022670   1.000137602   1.0044457
gpa         2.2345448   1.173858216   4.3238349
```

O cálculo da probabilidade de admissão (ou não) de um individuo com características que possam (ou não) ter sido registadas nos dados inseridos é importante e a realização desse cálculo probabilístico em R pode definir-se pela criação de um novo conjunto de dados e

pela utilização do comando *predict()*. Na tabela seguinte estão os dados de 20 candidatos fictícios com valores de *gre* compreendidos entre os valores mínimo e máximo registados inicialmente; em que à variável *gpa* é atribuído o respetivo valor médio e são atribuídas as 4 categorias da variável *rank*. Os valores apresentados correspondem as 3 primeiras e 3 ultimas linhas de dados obtidos no output inicial.

Tabela 46:*Criação de dados para o cálculo de probabilidades a partir do modelo logístico*

```
> dados4.1<-with(dados4,data.frame(gre = rep(seq(from =min(dados4$gre), to =max(dados4$gre),
length.out = 5),4),gpa = mean(gpa), rank = factor(rep(1:4, each = 5))))
> dados4.1
     gre   gpa   rank
 1   220 3.3899   1
 2   365 3.3899   1
 3   510 3.3899   1
 …    …     ….      …
18   510 3.3899   4
19   655 3.3899   4
20   800 3.3899   4
```

É essencial que as variáveis definidas no novo conjunto de dados tenham a mesma designação que as variáveis do modelo logístico utilizado. Em seguida realiza-se o cálculo de probabilidade de admissão para os 20 indivíduos criados e inserem-se os valores calculados na tabela anterior. Simultaneamente, ao indicar *se=TRUE*, obtém-se o valor do desvio padrão para o cálculo dos I.C. (95%). Os limites do I.C. estão definidos com L.I. e L.S.

Tabela 47:*Probabilidades previstas e respetivos I.C.*

```
> dados4.2 <- within(dados4.2, {
+ PredictedProb <- plogis(fit)
+ LI <- plogis(fit - (1.96 * se.fit))
+ LS <- plogis(fit + (1.96 * se.fit))
+ })
> dados4.2
    gre   gpa   rank        fit      se.fit   residual.scale      LS          LI          PredictedProb
1   220 3.3899   1   -0.7661985  0.4961412          1      0.5513776  0.14948637      0.31730202
```

2	365	3.3899	1	-0.4378568	0.3718161	1	0.5722172	0.23746713	0.39225178
3	510	3.3899	1	-0.1095150	0.2840161	1	0.6099630	0.33935124	0.47264857
…									
17	365	3.3899	4	-1.9893204	0.4084120	1	0.2334677	0.05787743	0.12032877
18	510	3.3899	4	-1.6609787	0.3356348	1	0.2683256	0.08957601	0.15963066
19	655	3.3899	4	-1.3326370	0.3298906	1	0.3349120	0.12140174	0.20872351
20	800	3.3899	4	-1.0042952	0.3941213	1	0.4423038	0.14470206	0.26809777

Sobre a qualidade de ajuste do modelo obtido serão realizados análises e testes sobre os resíduos de Pearson, *Deviance* e aplicar o teste de Hosmer-Lemeshow.

Para a obtenção dos resíduos de Pearson é necessário especificar o tipo de resíduo no comando *residuals*.

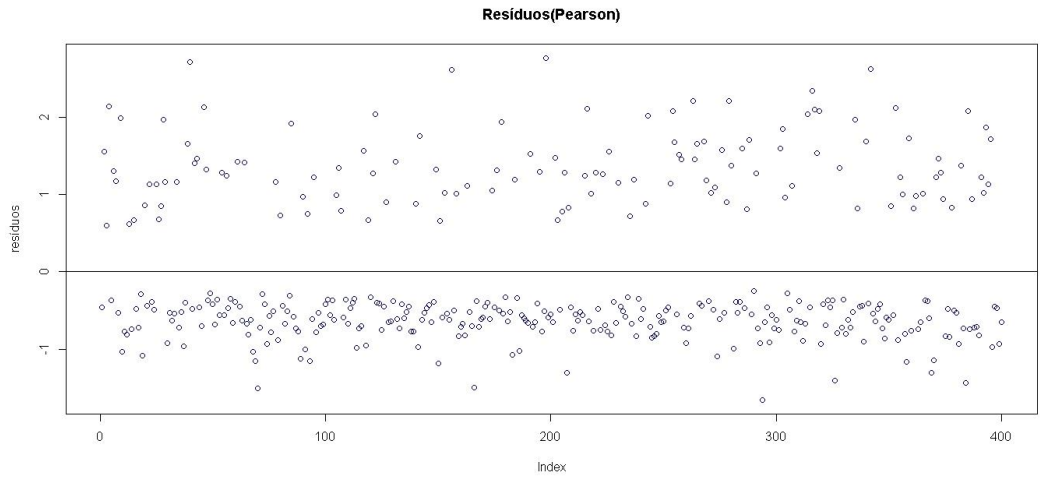
res4<-residuals(modelo4,type="pearson")

Para uma análise gráfica dos resíduos utiliza-se a seguinte sintaxe:

plot(res4,main="Resíduos(Pearson)",ylab="resíduos",col="darkblue")

abline(h=0)

Figura 14:*resíduos de Pearson*



A estatística de teste para a qualidade de ajuste a partir dos resíduos de Pearson dada em (4.16) é pode ser calculada da seguinte forma:

1-pchisq(sum((res4)^2),df=400-(3+1))

Neste caso, o valor do *p-value* determinado foi aproximadamente 0.47. Assim, não se rejeita a hipótese nula de adequação do modelo.

A obtenção dos resíduos *deviance* (4.17) é realizada de forma semelhante. Por defeito, o comando *residuals* fornece os resíduos *deviance.* Então, a sintaxe a utilizar é para obter os resíduos e o respetivo gráfico consiste nas seguintes linhas de comando:

```
res4.1<-residuals(modelo4)
plot(res4.1,main="Resíduos(Deviance)",ylab="resíduos",col="navyblue")
abline(h=0,col="red")
```

A distribuição dos resíduos *deviance* pode ser observada na Figura 15. Através do *package car* e da função *crPlots(modelo)* é possível representar os resíduos para cada uma das variáveis explicativas e verificar se o seu comportamento é relativamente linear. Esta função será exemplificada no capítulo 5, em 5.4.

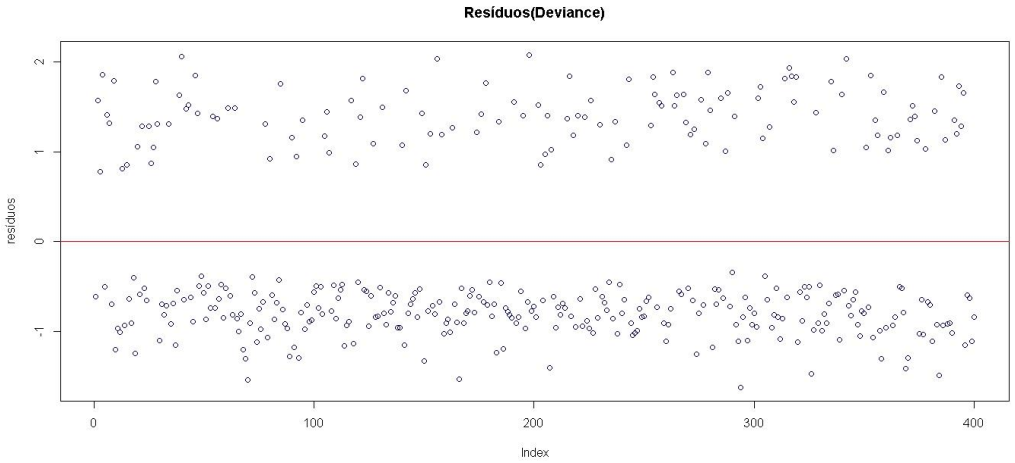
O teste para a qualidade de ajuste do modelo a partir destes resíduos (4.18) também é uma estatística Qui-quadrado realizada a partir da soma dos quadrados dos valores obtidos anteriormente.

```
1-pchisq(sum((res4.1)^2),df=400-(3+1))
```

A hipótese inicial H_0: "O modelo é ajustado" não deve ser aceite para um nível de significância inferior a 0,05 pois o *p-value* obtido é aproximadamente 0,02. A partir dos resíduos *Deviance* não há evidência estatística que o modelo em análise esteja ajustado.

A realização do teste de Hosmer-Lemeshow (4.19) consiste na comparação entre o número de uns e zeros observados e estimados nos dez grupos criados através de uma

Figura 15: *Diagrama sobre Distribuição dos resíduos (Deviance)*



estatística Qui quadrado. Para realizar o teste no ambiente R é necessário instalar o ficheiro *ResourceSelection*. Após instalar e carregar o ficheiro de dados na consola de comandos utiliza-se o comando *hosmer.test().* Os elementos que se devem colocar no argumento do comando são: os valores da variável resposta (admit) a partir do conjunto de dados inicial (dados4); os valores ajustados a partir do modelo logistico através do comando *fitted* e o número de grupos.

Tabela 48:*Output sobre Teste de Hosmer-Lemeshow*

```
> library(ResourceSelection)

>hl<-hoslem.test(dados4$admit,fitted(modelo4),g=10)

> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  dados4$admit, fitted(modelo4)

X-squared = 11.085, df = 8, p-value = 0.1969
```

A estatística de teste definida anteriormente como C, resultou em 11,085. Este valor não é superior ao valor tabelado de $\chi^2_{8,1-\alpha}$ que é 15,587. Assim, não se rejeita que o modelo logístico "modelo4" seja ajustado. Por outras palavras , não há evidência estatística que o modelo esteja mal ajustado. A representação dos grupos criados e dos valores da variável resposta observados e estimados também pode ser apresentada sob a forma de uma matriz através do comando *cbind()* tal como pode ser observado na tabela seguinte.

Tabela 49:*Grupos de dados para teste de Hosmer-Lemeshow*

```
> cbind(hl$expected,hl$observed)

                 yhat0     yhat1       y0   y1

[0.0588,0.136]  35.49158   4.508419    36    4
(0.136,0.181]   33.50546   6.494542    35    5
(0.181,0.217]   32.10793   7.892066    26   14
(0.217,0.261]   30.46332   9.536682    32    8
(0.261,0.298]   28.74853  11.251469    31    9
(0.298,0.34]    27.24296  12.757040    28   12
(0.34,0.376]    25.66337  14.336626    29   11
(0.376,0.437]   23.82702  16.172976    19   21
(0.437,0.531]   20.75104  19.248958    21   19
(0.531,0.738]   15.19878  24.801222    16   24
```

Na primeira coluna podem observar-se os intervalos de valores probabilísticos que limitam cada um dos 10 grupos. Nas 4ª e 5ª colunas estão registados os valores da variável de resposta *admit* (0 e 1) observados em cada um dos grupos. Por exemplo, no 1º grupo existem trinta e seis valores 0 e quatro valores 1 registados, enquanto no ultimo grupo existem dezasseis valores 0 e vinte e quatro 1 observados. Note-se que à medida que a probabilidade aumenta, o número de candidatos não admitidos diminui e o número de candidatos admitidos aumenta. Nas 2ª e 3ª colunas são apresentados os valores estimados para a variável resposta.

Por último procede-se à análise da predictabilidade do modelo. O método utilizado é a elaboração da curva ROC e conclusão sobre os valores obtidos. A matriz de confusão é obtida a partir dos dados iniciais da variável de resposta "admit" e dos dados estimados $\hat{Y}$. Através do comando *predict*() é possível recriar a tabela de contigência definida na Tabela 38.

Tabela 50: *Sintaxe para criar Tabela 2X2 (Matriz de Confusão)*

```
> pred <- predict(modelo4, type = 'response')
> table(dados4$admit, pred > 0.5)

    FALSE  TRUE
 0   254    19
 1    97    30
```

A representação da curva ROC em R pode ser realizada através dos comandos de diferentes *packages.* Neste caso optou-se pelo *package pROC.* Após instalar o ficheiro

pROC e carrega-lo na consola de comandos cria-se um objeto associado à função *plot.roc()*. No argumento deste comando são indicados os valores ajustados pelo modelo e os valores iniciais da variável resposta.

roc4<-plot.roc(dados4$admit,fitted(modelo4))

Para a elaboração do gráfico é utilizado o comando *plot()* da seguinte forma:

> plot(roc4, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2),

+ grid.col=c("green", "red"), max.auc.polygon=TRUE,

+ auc.polygon.col="lightgreen", print.thres=TRUE)

Call:

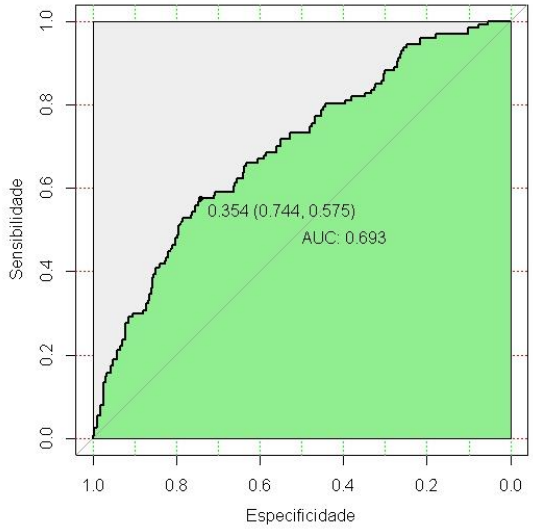
plot.roc.default(x = dados4$admit, predictor = fitted(modelo4))

Data: fitted(modelo4) in 273 controls (dados4$admit 0) < 127 cases (dados4$admit 1).

Area under the curve: 0.6928

No argumento do comando *plot* estão as instruções para inserir o valor *auc* no gráfico; para representar o polígono que corresponde à área *auc*; para inserir uma grelha; para colorir a área *auc* com a cor verde claro e para representar o ponto de corte e respetivas coordenadas. Após a finalização da sintaxe, é criado o gráfico e em simultâneo é obtido um *output* sobre o valor da área *auc* e sobre os valores totais das linhas da matriz de confusão. O gráfico é o seguinte:

Figura 16:***Curva ROC associada ao modelo logístico***



Através do pacote *pROC* também é possível obter valores para intervalos de confiança da sensibilidade e especificidade no entanto, as definições associadas a esses valores e a sua determinação em R não são tratados neste trabalho. O valor obtido para a área *auc* é de 0,693. Isto significa que a capacidade de predição do modelo não pode ser considerada excelente, pois é um valor abaixo de 0,70. Habitualmente, considera-se que um modelo com uma valor acima de 0,70 é fiável quanto à precisão na predictabilidade. No entanto, o valor obtido é muito próximo.

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

CAPÍTULO 5

5.1. Aplicabilidade dos modelos de regressão no âmbito da análise do risco

A análise do risco, segundo o autor (Simonovich, 1997) que cita vários autores, é sustentada por dois tipos de abordagem. Em primeiro temos a estruturação do processo de decisão e em segundo lugar a utilização de um conjunto de técnicas de avaliação do valor de diferentes decisões e é neste conjunto que se inscrevem técnicas matemáticas tais como a modelação.

O processo de modelação necessário para a análise do risco, nomeadamente a simulação, exige a estruturação do problema, obrigando o executivo a debruçar-se sobre o problema, esclarecendo-o, formulando-o em função de características reais e facilitando a comunicação entre os vários agentes decisores na organização.

Neste capítulo serão apresentados 4 casos onde é possível aplicar modelos de regressão para a obtenção de resultados válidos que permitam analisar, compreender e resolver problemas relacionados com a análise do risco. Todos estes dados foram objeto de outros estudos publicados cuja sua autoria será mencionada sempre que possível.

5.2. APLICAÇÃO DA REGRESSÃO LINEAR SIMPLES EM ANÁLISE DO RISCO

No âmbito da análise do risco, o número de pedidos de indemnização e os montantes pagos pelas seguradoras podem ser modelados através da regressão linear simples. Nestes caso, os dados são provenientes do "*Swedish Committee on Analysis of Risk Premium in Motor Insurance*" e as variáveis a designar estão descitas na Tabela 51. Os diferentes dados provém de várias zonas geográficas da Suécia e não é indicada qualquer data de registo dos mesmos.

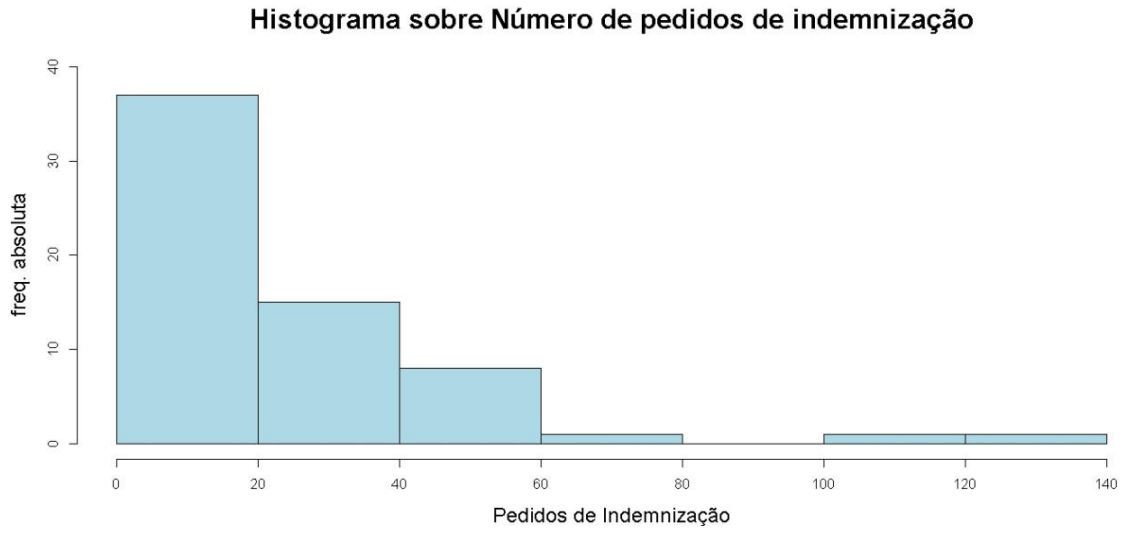
Antes de iniciar a definição do modelo linear é necessário proceder à caracterização dos dados. O total de dados obtidos é n=63, isto é ,os dados provém de 63 regiões geográficas diferentes.

Tabela 51: *Dados para o modelo de regressão linear simples*

Abreviatura	***Variável***	***Unidades de medida***
Y	*Total de pagamentos de indemnizações*	*(em milhares) em Coroas Suecas*
X	*variável independente Número de pedidos de indemnização*	Valores discretos

A distribuição dos dados sobre a distribuição dos pedidos de indemnização pode ser observada na figura seguinte com maior predominância do número de pedidos da classe [0,20[com uma frequência de 37 regiões.

Figura 17: *Histograma sobre a distribuição da variável x*



Graficamente, a distribuição dos montantes pagos por região pode ser observada na Figura 18.

Após a caracterização das duas variáveis, deve-se proceder à representação da distribuição bivariada através do diagrama de dispersão representado na Figura 19. Simultaneamente, verifica-se que o nível de correlação das duas variáveis através do cálculo do coeficiente de Pearson, é positivo forte.

Figura 18: *Histograma sobre a distribuição da variável Y*

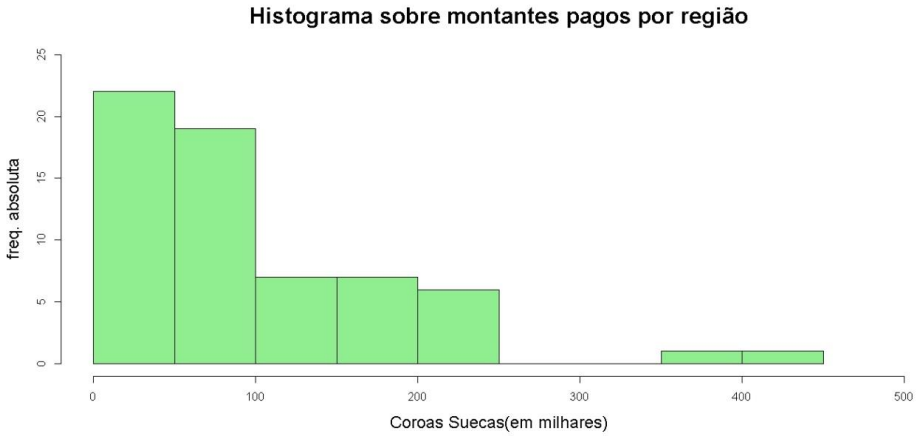
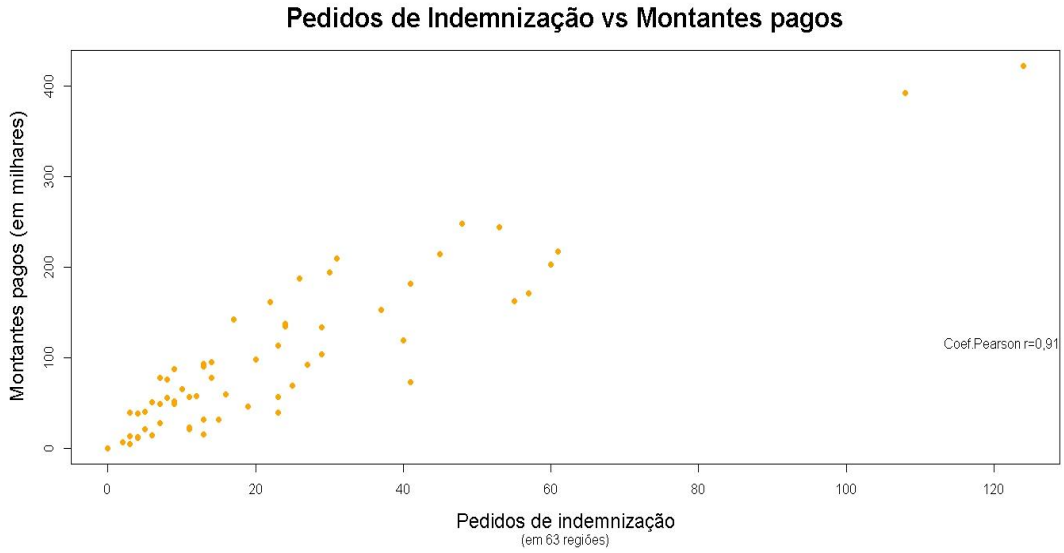


Figura 19: *Diagrama de dispersão para as variáveis X e Y*



Assim, a construção do modelo linear tem como base a relação entre os montantes pagos (variável de resposta) e o número de pedidos de indemnização (variável explicativa) para as 63 regiões registadas. O modelo de regressão linear simples determinado em R tem a seguinte expressão analitica:

$$Y=19{,}9945+3{,}4138X$$

Em que os I.C. (em 95%) para os parâmetros β_0 e β_1 são:

Tabela 52: *I.C.(95%) para os parâmetros do modelo linear simples*

parâmetros	*Limite inferior*	*Limite superior*
β_0	*7,2614*	*32,7276*
β_1	*3,0230*	*3,8047*

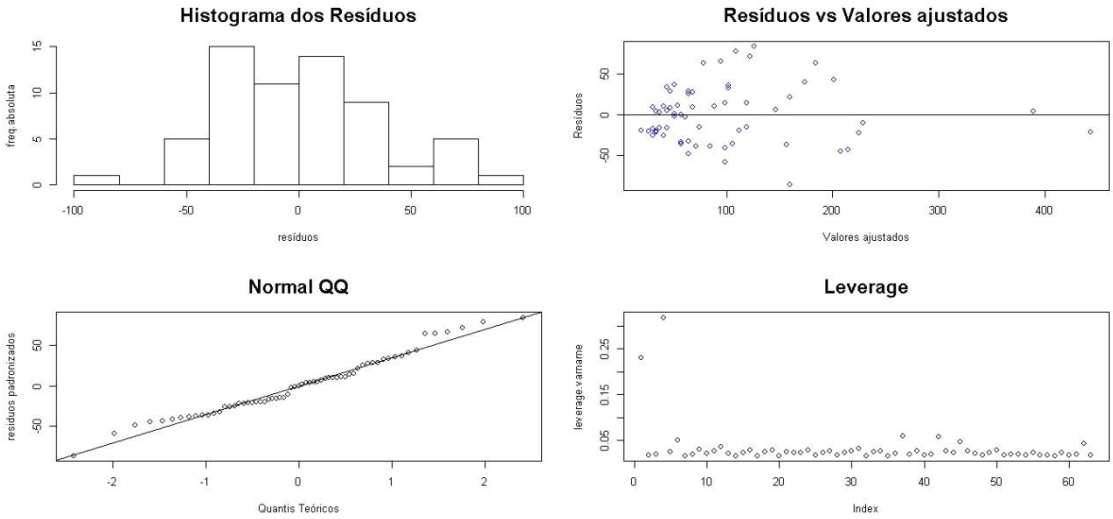
Quanto à análise de resíduos, aplicação do teste de Breusch-Pagan revela que deve ser aceite a hipótese nula de homocedasticidade e independência dos resíduos. Através da Tabela 53, observamos que para um nível de significância α=0,05, não rejeitamos H_0, ou seja, os resíduos são homoscedásticos segundo o teste de Breusch-Pagan.

Tabela 53:*Teste de Breusch-Pagan*

G.L.	*Estatística BP*	*p value*
1	1,71	0,1909

Graficamente, análise de resíduos pode ser realizada a partir dos gráficos seguintes.

Figura 20:*Conjunto de gráficos sobre análise de resíduos*



Através da observação dos gráficos é possível concluir que os resíduos tem uma distribuição normal, corroborado pelo *p-value* (0,22) determinado no teste Shapiro Wilk (ver anexo) ,e que estes são independentes dado que a mancha produzida não está definida segundo um padrão claro.

Assim, dado que os principais pressupostos para um modelo linear simples não foram violados, conclui-se que este modelo é estatísticamente válido para explicar a associação entre as variáveis relacionadas com pedidos de indemnização e montantes pagos e também para eventuais estimativas no âmbito da inferência estatística.

5.3. APLICAÇÃO DE REGRESSÃO LINEAR MÚLTIPLA

O investimento no mercado bolsista é uma atividade cujo risco é inerente. A decisão de investir no mercado bolsista é geralmente, baseada em alguns critérios. Em primeiro lugar, pretende-se um lucro elevado. Em segundo, deve-se ter em conta o risco do titulo em que se pretende investir. Este risco pode ser medido pela variabilidade do retorno associado. Por outro lado, os investidores preocupam-se com a possibilidade de vender os títulos no momento que considerem oportuno, isto é a capacidade de liquidar o conjunto de títulos em sua posse. Quanto maior for a capacidade de liquidação associada a uma ação, mais fácil será a sua venda. Os dados utilizados neste exemplo de aplicação de modelação linear múltipla foram obtidos para a realização de um estudo em que pretendeu estudar a relação entre o número de ações de várias empresas negociadas num dado período de tempo , também designado por volume, e outras características financeiras associadas a um título de bolsa. Para este trabalho, os dados referidos foram obtidos no site da *Wisconsin School of Business* cujo link é: http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html .

A indicação e descrição das variáveis está indicada na Tabela 54.

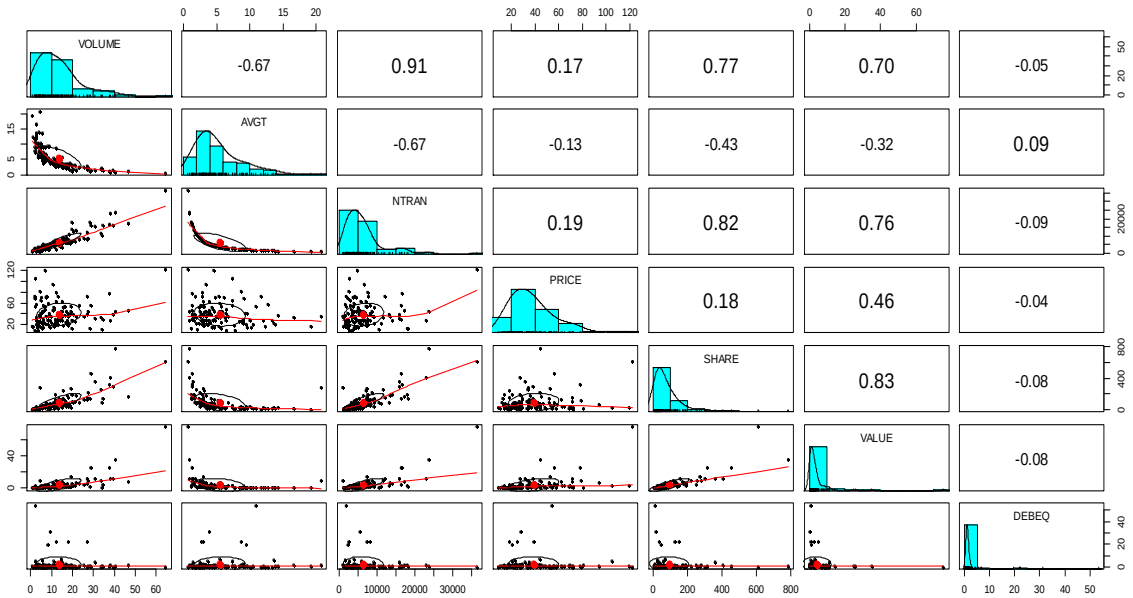
Tabela 54: *Variáveis para o modelo linear multiplo*

Abreviatura	***Variável***	***Unidades de medida***
VOLUME	*variável dependente contínua que designa o número total de transações de cada empresa em 3 meses.*	Milhões de ações
AVGT	*variável independente continua relativa ao tempo médio entre transação*	minutos
NTRAN	*variável independente continua que designa o número total de transações em 3 meses de cada empresa*	
PRICE	*Variável independente contínua que representa o preço de abertura para a ação de cada*	Dólar americano

	empresa em 2 janeiro de1985	
SHARE	*Variável independente contínua que representa o número de ações detidas por acionistas em 31 de Dezembro de 1984.*	milhões
VALUE	*Variável independente contínua que representa o valor bolsista da empresa*	Milhões de dólares americanos
DEBEQ	*Variável independente contínua que representa a razão entre fundos próprios e passivo.*	

Antes de obter o modelo linear que permita verificar a influência dos vários fatores no volume associado à liquidação de títulos analisa-se a correlação entre as várias variáveis através de uma série de diagramas e dos seus coeficientes de correlação.

Figura 21: *Diagramas de correlação entre as várias variáveis para o MRLM*



De notar que existe uma forte correlação positiva entre VOLUME e NTRAN e correlações positivas entre VOLUME e as variáveis SHARE e VALUE. As correlações mais fracas são entre VOLUME e as variáveis PRICE e DEBEQ.

Os coeficientes para o modelo linear múltiplo estão representados na tabela seguinte, tal como os valores de teste para a sua significância no modelo obtido.

Tabela 55: *Coeficientes e estatísticas de deste das covariáveis do modelo*

Variável	coeficiente	*t*	*P*>|*t*|
intercepto	6,1570	3,147	0,0021
AVGT	-0,4436	-2,960	0,0037
NTRAN	0,0014	7,873	$2{,}03^{-12}$
PRICE	-0,0138	-0,606	0,5456
SHARE	0,0090	1,302	0,2206
VALUE	0,0812	0,715	0,4759
DEBEQ	0,0609	1,027	0,3065

As estatísticas de teste correspondentes às variáveis AVGT e NTRAN conduzem à não rejeição da hipótese nula que pressupõe que estas são significativas no modelo obtido.

O *output* produzido em R também indica dois valores extremamente importantes para a análise de um modelo de regressão linear: os coeficientes de determinação R^2 e R^2 *ajustado*. Para este modelo, os valores obtidos são 0,85 e 0,84 aproximadamente. Optando pelo coeficiente R^2 *ajustado*, conclui-se que cerca de oitenta e quatro por cento dos valores da variável VOLUME pode ser ajustados pelo modelo linear obtido.

O número total de ações negociadas num determinado período parece ser influenciado por apenas duas variáveis. No entanto, após testar vários modelos através da função *stepAIC(modelo)* em que foram retiradas variáveis ao modelo inicial obteve-se um modelo que contém também a variável SHARE e cujos valores dos coeficientes de determinação são aproximadamente idênticos.

Tabela 56: *coeficientes e estatísticas de teste para um novo modelo*

Variável	coeficiente	*t*	*P*>|*t*|
intercepto	5,2765	3,888	0,000167
AVGT	-0,3968	-2,873	0,004814
NTRAN	0,0014	9,082	$2{,}79^{-15}$
SHARE	0,0119	1,994	0,048402

Note-se que o processo de exclusão de variáveis e adaptação do modelo inicial deve ter em atenção o contexto do problema e a importância dessas mesmas variáveis no contexto financeiro e, não apenas numa perspetiva matemática. Neste exemplo, optou-se por analisar os coeficientes e proceder de acordo com conclusões meramente estatísticas.

Então, o modelo linear escolhido pode ser descrito através da expressão:

$$VOLUME = 5{,}2765 - 0{,}3968 \times AVGT + 0{,}0014 \times NTRAN + 0{,}0119 \times SHARE$$

Ainda sobre os valores dos coeficientes das variáveis do modelo escolhido, apresenta-se também a tabela de análise de variância.

Tabela 57:*tabela ANOVA sobre o modelo*

Fonte	g.l	Soma de Quadrados	Quadrado médio	Teste F	*Pr(>F)*
AVGT	1	6273,9	6273,9	353,68	2^{-16}
NTRAN	1	5336,7	5336,7	300,85	2^{-16}
SHARE	1	70,6	70,6	3,98	0,0484
resíduos	119	2110,9	17,7		
Total	119	13792,1			

Após verificar os valores tabelados da estatística F verifica-se que as variáveis deste segundo modelo são estatísticamente significativas para a composição do mesmo.

Por exemplo, ao analisar os dados obtidos para a variável AVGT, para $\alpha=0,05$ obtemos que o valor tabelado é $F_{(0,95;1;119)} = 3{,}92$. Logo, $F_0 > F_{(0,95;1;119)}$, em que $F_0 = 353,68$.

Portanto, rejeita-se H_0 com um nível de confiança de 95% e conclui-se que a variável explicativa AVGT tem correlação com a variável resposta.

A análise e conclusão em relação às outras covariáveis é análoga.

Os intervalos de confiança calculados a partir das expressões (2.11) para cada um dos coeficientes deste modelo estão indicados na Tabela 58.

Antes de considerar e aceitar os resultados do modelo linear obtido é importante avaliar a sua adequação aos dados através da análise de resíduos. Para este propósito, foram utilizados os testes de Durbin-Watson e de Breusch_Pagan e também a análise gráfica dos resíduos determinados a partir do modelo.

Tabela 58: *I.C. (95%) para os coeficientes*

Variável	Limite Inferior (2,5%)	Limite Superior (97,5%)
intercepto	2,5895	7,9635
AVGT	$-6{,}7018^{-01}$	-0,1233
NTRAN	$1{,}1149^{-03}$	0,0017
SHARE	$8{,}4960^{-05}$	0,0237

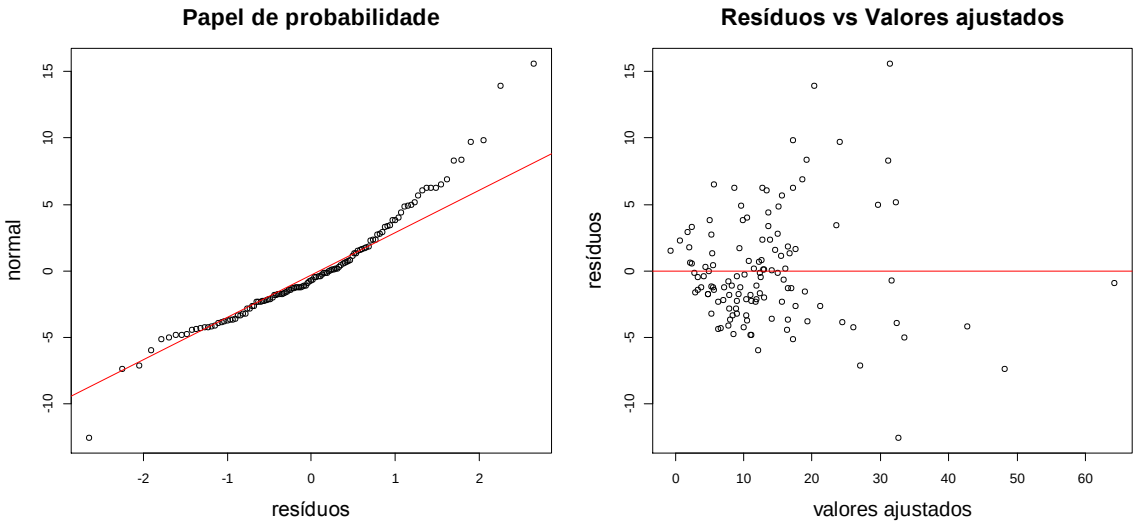
Tabela 59: *testes de análise de resíduos*

Teste	Estatística de teste	*p value*
Durbin-Watson	1,8916	0,2738
Breusch_Pagan	20,909	0,00011

O *p-value* obtido para o teste de *BP* de 0,00011 conduz à não aceitação da hipótese nula do teste de hipóteses. No entanto, o teste *DW* revela um *p-value* de 0,2738 que indica que não se deve rejeitar a hipótese nula de independência dos resíduos. O nível de significância em ambos os casos é de 0,05.

Assim, com os resultados obtidos, é de todo aconselhável a análise gráfica e o teste de normalidade aos resíduos.

Figura 22: *Gráficos de papel de probabilidade e Resíduos vs valores ajustados*



A partir dos gráficos obtidos verifica-se que os resíduos não apresentam o comportamento de dados distribuídos normalmente. Este fato é corroborado pelo teste de Shapiro-Wilk

realizado cujo *p-value* calculado é $7{,}45^{-5}$. Este valor não garante a não rejeição da hipótese nula: os dados provém de uma distribuição normal. Quanto ao diagrama de resíduos e valores ajustados verifica-se que não existe variância constante nos resíduos. Logo, não existe Homocedasticidade. De referir, que o modelo inicial foi submetido aos mesmos testes e os resultados foram semelhantes. Nesta fase, dados que os pressupostos para um modelo de regressão linear foram violados, não é prudente tomar este modelo como estatísticamente válido para explicar a relação entre as variáveis ou para proceder a inferências a partir do mesmo.

5.4. APLICAÇÃO DA REGRESSÃO LOGÍSTICA SIMPLES

A aplicabilidade da regressão logística é muito incidente no campo da medicina, na avaliação do risco de ocorrência de doenças graves, através do cálculo de probabilidade com o modelo logístico.

Neste exemplo, os dados provém do site da universidade de Messachussets amherst - *statistical software information*, cujo link é https://www.umass.edu/statdata/statdata/data/chdage.dat . As variáveis em análises são idade e a ocorrência ou não de episódios de doença coronária cardíaca (CHD em inglês) em 100 indivíduos. Pretende-se associar o risco de ocorrência de episódios de doença cardíaca à idade e simultaneamente criar um modelo matemático de regressão logística simples que permita ajustar os dados em análise e, elaborar conclusões estatísticamente válidas sobre o mesmo. A designação das variáveis em estudo está representada na tabela seguinte.

Tabela 60: *Designação das variáveis*

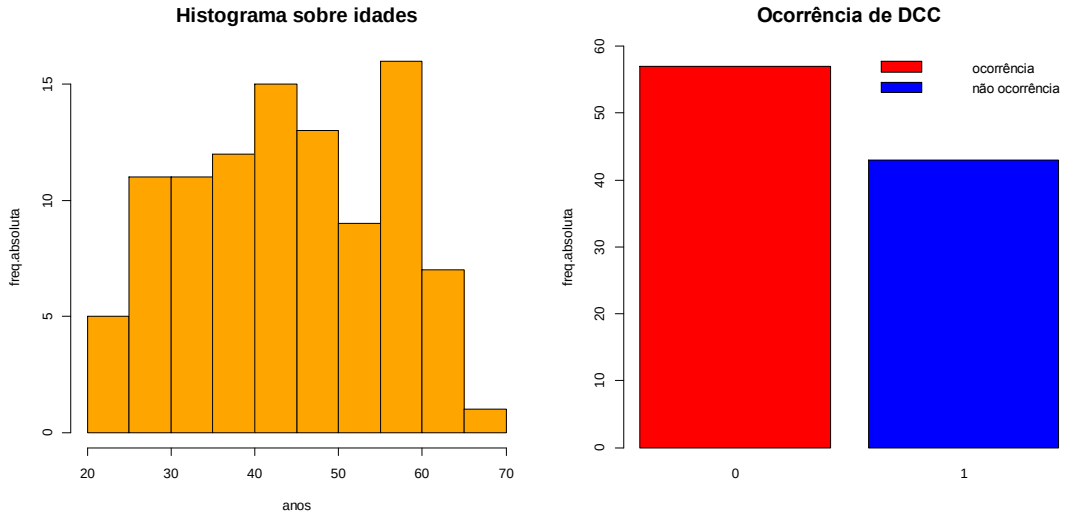
Abreviatura	***Variável***	***Unidades de medida***
idade	*variável explicativa continua relativa à idade de cada individuo*	anos
DCC	*variável independente categórica que designa a ocorrência ou não de episódios de doença cardíaca coronária*	0-Ocorrência; 1-Não Ocorrência

A distribuição dos dados das variáveis e algumas medidas estatísticas sobre as mesmas estão representadas na Tabela 61 e na Figura 23.

Tabela 61: *Medidas de localização sobre a variável idade*

Min.	Média	Mediana	Máx.
20	44,38	44	69

Figura 23:*Distribuição das variáveis*



O modelo de regressão Logística obtido a partir do ambiente R tem os coeficientes que estão assinalados na tabela seguinte e os valores de teste de Wald sobre a significância dos mesmos estão representados na quarta coluna.

Tabela 62: *coeficientes e valores de teste*

Variável	coeficiente	z	$P>\|z\|$
intercepto	-5,30945	-4,683	2.82^{-6}
idade	0,11092	4,610	4.02^{-6}

O modelo que traduz a relação entre a variável explicativa e a variável dependente é definido pela seguinte função de ligação estimada:

$$\hat{g}(x) = -5{,}30945 + 0{,}11092 idade$$

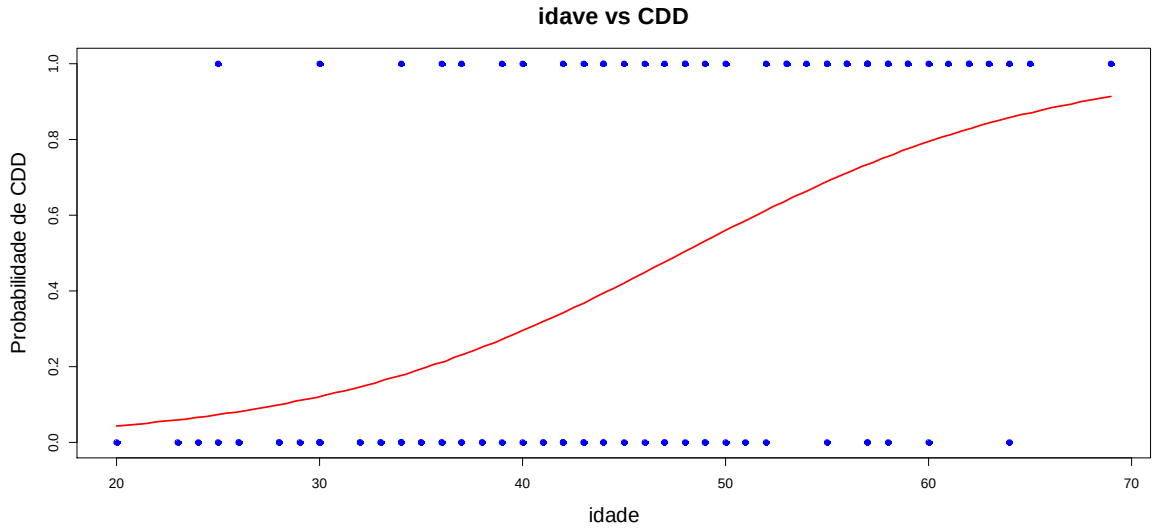
O modelo determinado também pode ser representado pela seguinte expressão:

$$\hat{\pi}(idade) = \frac{e^{-5,30945+0,11092 \times Idade}}{1+e^{-5,30945+0,11092 \times Idade}}$$

A partir dos dados dos valores de teste, $2,82^{-6}$ e $4,02^{-6}$ rejeita-se a hipótese nula para um *p-value* e conclui-se que os coeficientes determinados são diferentes de zero, isto é, são estatísticamente significativos para a modelação pretendida.

Na Figura 24 é possível observar a disposição dos dados das duas variáveis e o gráfico da função logística correspondente ao modelo determinado.

Figura 24: *representação gráfica do modelo logístico*



A interpretação dos coeficientes torna-se mais simples após o cálculo da exponenciação dos mesmos. A análise do risco sobre a incidência de episódios de CDD é realizada a partir dos valores obtidos para os parâmetros do MLG simples. Isto é, a partir do valor obtido conclui-se que as chances (*odds*) de se registar um episódio de CDD são de 1,12 aproximadamente quando a idade aumenta uma unidade.

A qualidade dos coeficientes do modelo é avaliada pelo teste de Wald e pelo teste de verossimilhança. Os valores do teste de wald estão indicados na quarta coluna da Tabela 62 e indicam que os coeficientes são estatísticamente significativos tal como foi referido na página anterior.

O valor teste de verossimilhança dado pela estatística *G* para o modelo é de 29,31 aproximadamente e com uma distribuição Qui-Quadrado com um g.l. também conduz à não aceitação da hipótese nula.

O teste Score, para averiguação da significância dos coeficientes, tem como resultado 26,399 e sob uma distribuição Qui-Quadrado não se aceita a hipótese nula com um *p-value* inferior de $2{,}777^{-7}$ para um nível de confiança de 0,05.

Logo, conclui-se que os coeficientes são estatísticamente significativos para o modelo logístico determinado, ou seja, a idade tem influência na ocorrência de episódios de CDD. Os intervalos de confiança para os coeficientes do modelo estão assinalados na tabela seguinte.

Tabela 63*:I.C. (95%) para os parâmetros do modelo*

Variável	Limite Inferior (2,5%)	Limite Superior (97,5%)
intercepto	-7,7259	-3,2462
idade	0,0669	0,1620

A leitura e interpretação dos valores dos parâmetros torna-se mais clara quando se procede à sua transformação através da exponenciação de base *e.* Os valores das *odds* associadas ao intercepto e à variável *idade* assim como os respetivos intervalos de confiança são apresentados na tabela seguinte.

Tabela 64: *I.C. (95%) dos parâmetros exponenciados*

Variável	*Odds*	Limite Inferior (2,5%)	Limite Superior (97,5%)
intercepto	0,0049	0,0004	0,0389
idade	1,1173	1,0692	1,1176

Em relação à "bondade" (ou qualidade) de ajuste do modelo a partir dos resíduos podem aplicar-se os testes de resíduos de Pearson e de Deviance definidos para o modelo de regressão múltipla.

Tabela 65: *Testes de resíduos*

Teste	*p-value(Qui Quadrado)*
Pearson	0,4272
Deviance	0,2896

De acordo com os resultados obtidos, aceita-se a hipótese nula de que o modelo obtido é ajustado aos dados. O nível de significância considerado é de 0,05.

Assim, pode concluir-se que este modelo de regressão logística "explica" estatísticamente a relação entre a idade e a ocorrência de episódios cardíacos para os individuos analisados e o modelo de regressão logística simples obtido é estatísticamente válido no âmbito da inferência estatística. No entanto, o estudo, análise e prevenção de episódios cardíacos não deverá estar sujeita a um modelo simples de uma variável independente.

5.5. MODELO DE REGRESSÃO LOGÍSTICA MÚLTIPLA NA ANÁLISE DO RISCO

Neste exemplo será utilizado uma base de dados retirada do site da universidade de Messachussets amherst - *statistical software information*, cujo link é https://www.umass.edu/statdata/statdata/data/pros.dat. Os dados que serão utilizados fazem parte de um estudo sobre a incidência de cancro na próstata em 380 indivíduos. O objetivo deste estudo era determinar se a penetração da cápsula prostática, ou não, por um tumor poderia ser prevista a partir dos dados, recolhidos em exame prévio, de várias variáveis (Hosmer & Lemeshow, 2000). A descrição das variáveis pode ser observada na Tabela 66.

Tabela 66: *Descrição das variáveis*

Abreviatura	***Variável***	***Unidades de medida***
AGE	*variável independente continua relativa à idade de cada individuo*	anos
RACE	*variável independente categórica que define a raça de cada individuo*	1-branca;2-negra
DPROS	*variável independente categórica relativa ao exame rectal digital*	1-Não há nódulo; 2-Apresenta Nódulo (lado esq); 3-Apresenta Nódulo (lado dir.); 4-Apresenta dois nódulos

DCAPS	*variável independente categórica sobre a deteção de envolvimento capsular no exame rectal digital*	1-Não;2-Sim
PSA	*Variável independente contínua relativa ao Nível de antígeno prostático especifico*	mg/ml
VOL	*Variável independente contínua relativa ao volume do tumor obtido em exame de ultrassom*	cm^3
GLEASON	*Variável independente relativa à pontuação do teste de Gleason.*	0-10
CAPSULE	*Variável dependente dicotómica relativa à penetração do tumor na cápsula prostática*	0-Não há penetração; 1-Há penetração

Após a codificação das variáveis RACE, DPROS e DCAPS em variáveis fatoriais, os coeficientes do modelo logístico múltiplo (ou modelo linear generalizado) obtidos inicialmente em R estão descritos na tabela seguinte.

Tabela 67: *Variáveis e respetivos coeficientes e valores de teste*

Variável	coeficiente	*z*	$P>\|z\|$
intercepto	-6,9665	-4,301	$1,7^{-5}$
AGE	-0,0118	-0,599	0,54949
RACE2	-0,6513	-1,379	0,16775
DPROS2	0,7302	2,034	0,04196
DPROS3	1,5095	4,002	$6,29^{-5}$
DPROS4	1,3872	3,002	0,00268
DCAPS2	0,4924	1,062	0,28819
PSA	0,0299	2,959	0,00308
VOL	-0,0115	-1,467	0,14244
GLEASON	0,9625	5,770	$7,94^{-9}$

O modelo que traduz a relação entre as covariáveis e a variável dependente é aquele que é definido pela seguinte função de ligação estimada:

$$\hat{g}(x) = -6{,}9665 - 0{,}0118AGE - 0{,}6513RACE2 + 0{,}7302DPROS2 + 1{,}5095DPROS3 + 11{,}3872DPROS4 + 0{,}4924DCAPS2 + 0{,}0299PSA - 0{,}0115VOL + 0{,}9625GLEASON$$

Para uma melhor interpretação do modelo exponenciam-se os coeficientes obtidos. Os valores podem ser observados na tabela seguinte.

Tabela 68: *coeficientes após exponenciação*

AGE	RACE2	DPROS2	DPROS3	DPROS4	DCAPS2	PSA	VOL	GLEASON
0,988	0,5213	2,0755	4,5244	4,0035	1,6362	1,0303	0,9886	2,6183

Estes valores indicam que, por exemplo, ao aumentar em uma unidade a idade (AGE) as chances de haver penetração da cápsula prostática é de 0.988 ou, se aumentar em uma unidade a escala de Gleason (GLEASON) as chances de haver penetração da cápsula prostática é de 2,6183.

A significância do coeficientes determinados pode ser analisada através do teste de razão de verossemelhança e através do teste de Wald. A hipótese nula a testar é de que os coeficientes são nulos, isto é não tem qualquer influência sobre a variável dependente. O estatística de teste G, associada ao teste da razão de verossimilhança (2.13), obtida através do software R é 132,0522. O *p-value* para a distribuição Qui-quadrado é $P\left[\chi^2(8) > 132{,}0522\right] = 0$ e é significativo para α=0,05. Assim, não se aceita a hipótese nula e concluí-se que pelo menos uma covariável (ou mais), é diferente de zero. Este modelo contém variáveis cuja estatística de teste indica claramente que estas não são significativas para o modelo em causa. Os valores do teste de Wald estão representados na quarta coluna da Tabela 67. Para um nível de significância de 0,05 pode-se concluir que as variáveis AGE, RACE2, DCAPS2 e VOL não são estatísticamente influentes no modelo obtido.

Para a obtenção de um modelo que melhor se ajuste aos dados poderá ser necessário reduzir o número de variáveis e comparar este novo modelo com o modelo completo. Sempre que uma variável categórica é incluída (ou excluída) de um modelo, todos os níveis desta variável devem ser incluídas (ou excluídas). Caso contrário, seria necessário a realização de um complexo processo de recodificação da variável (Hosmer & Lemeshow, 2000).

O teste de verossemelhança para a comparação do modelo reduzido obtido (ver anexo II) com o modelo completo resulta no valor de $G = -2[(-190{,}5612)-(-187{,}2676)] = 6{,}5872$. O *p-value* para a distribuição Qui-quadrado é $P[\chi^2(6) > 6{,}5872] = 0{,}3607$ e excede o nível de confiança 0,05. Assim, concluí-se que o modelo reduzido é tão bom quanto o modelo completo, o que pode conduzir à exclusão das variáveis relacionadas com idade e raça. No entanto, num estudo de carácter clinico estas são variáveis que certamente são importantes e que não devem ser excluídas apenas por critérios meramente estatisticos. Consequentemente, manter-se-á o estudo do modelo completo.

Os intervalos de confiança (95%) para os coeficientes do modelo apresentados na tabela 70 estão representados na Tabela 69. A sua interpretação pode ser a seguinte: há uma confiança de 95% que as *Odds* de penetração na cápsula prostática pelo tumor situam-se entre 0,951 e 1,027, aproximadamente, quando a idade (AGE) aumenta uma unidade.

Em seguida, procede-se à análise da qualidade do modelo logístico a partir dos seus resíduos cuja distribuição pode ser observada na

Figura 25. A análise aos gráficos de resíduos das covariáveis continuas permite aceitar que o comportamento linear da distribuição de dados não compromete o modelo obtido.

Tabela 69: *I.C. (95%) para os coeficientes*

Variável	Limite Inferior (2,5%)	Limite Superior (97,5%)
intercepto	3.6105^{-05}	0.0210
AGE	9.5061^{-01}	1.0272
RACE2	2.0041^{-01}	1.2903
DPROS2	1.0397	4.2730
DPROS3	2.1956	9.6842
DPROS4	1.6330	10.0639
DCAPS2	6.6968^{-01}	4.1878
PSA	1.0113	1.0522
VOL	9.7320^{-01}	1.0036
GLEASON	1.9105	3.6807

Em seguida, procede-se à análise da qualidade do modelo logístico a partir dos seus resíduos cuja distribuição pode ser observada na

Figura 25. A análise aos gráficos de resíduos das covariáveis continuas permite aceitar que o comportamento linear da distribuição de dados não compromete o modelo obtido.

A estatística de teste para a qualidade de ajuste a partir dos resíduos de Pearson é calculada através da expressão dada em (4.16) e o resultado obtido em R é 0,68 aproximadamente. Logo, aceita-se a hipótese nula de adequação do modelo.

O teste para a qualidade de ajuste do modelo a partir dos resíduos *Deviance* (4.18) realizado permite aceitar a hipótese inicial H_0: "O modelo é ajustado" pois o *p-value* obtido é aproximadamente 0,42, para um nível de confiança de 0,05. A partir dos resíduos *Deviance* há evidência estatística que o modelo em análise esteja bem ajustado.

A estatística de teste de Hosmer e Lemeshow ,definida anteriormente como C (4.19), resultou em 5,165. Este valor não é superior ao valor tabelado de $\chi^2_{8,1-\alpha}$ que é 15,587. Logo, não há evidência estatística que o modelo esteja mal ajustado.

Por último, será efetuada a análise à capacidade de predição do modelo logístico. Este processo consiste em criar a tabela de contingência denominada Matriz de Confusão e elaboração e análise da curva ROC. O valor de corte determinado previamente será de 0,5. Isto é, acima de 0,5 todos os valores probabilísticos preditos serão codificados como individuo com penetração da cápsula prostática por tumor.

Figura 25: *Resíduos do modelo logístico múltiplo*

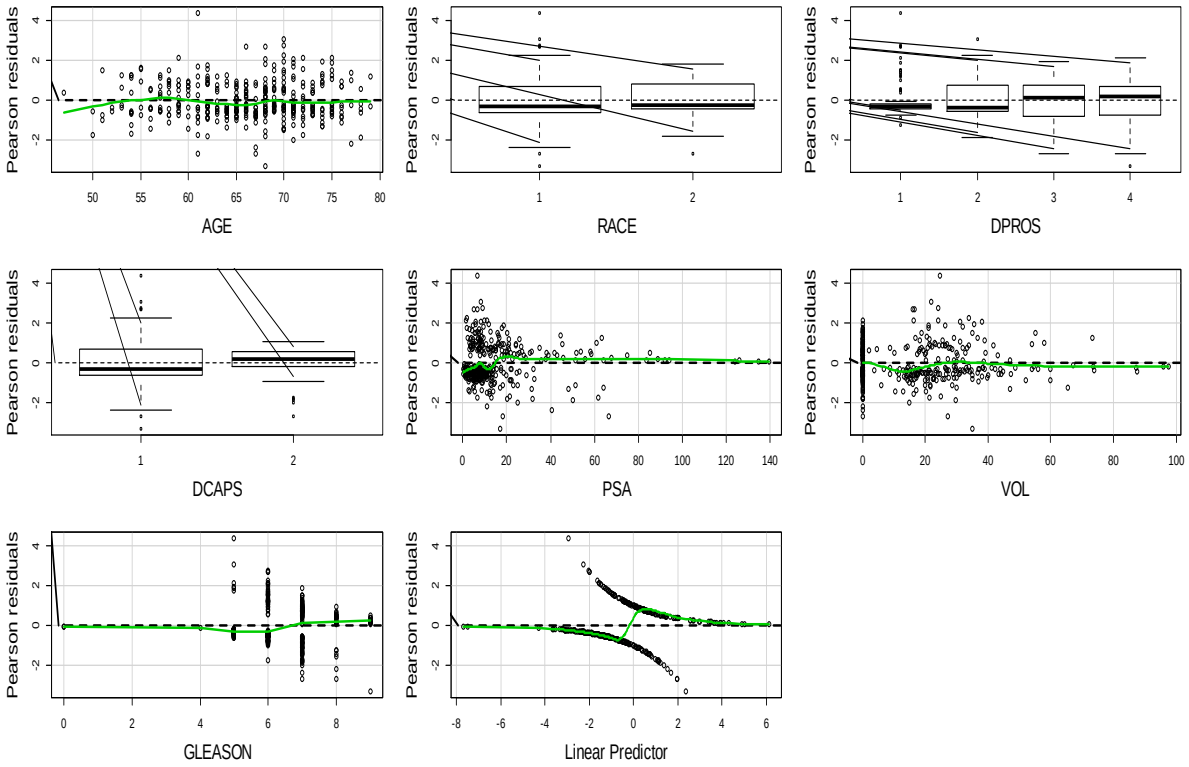


Tabela 70:*Tabela de contingência 2X2 ou Matriz de confusão*

n=376		Observação	
		Positivo	Negativo
Predição	Positivo	97	35
	Negativo	54	190

Os valores dos parâmetros associados à matriz de confusão que permitem uma melhor análise estão representados na *Tabela 71* e permitem, de alguma forma, concluir que o modelo completo é aceitável quanto à sua capacidade de predição. Note-se que todos os valores representados, com exceção da Especificidade, estão acima dos 70%. No entanto, é necessário notar que todos estes parâmetros podem estar sujeitos a vários desequilíbrios e interpretações erradas.

O conjunto de informações fornecidas pelo software R incluí também o teste de Mcnemar. Este teste é utilizado para analisar proporções de duas amostras relacionadas, isto é, tem como objetivo avaliar a eficiência de situações "antes" e "depois", em cada individuo é utilizado como o seu próprio controlo. Caso o *pvalue* determinado for inferior ao nível de confiança α , aceita-se H_0: "Não Existem diferenças entre os valores efetivos e os valores preditos".

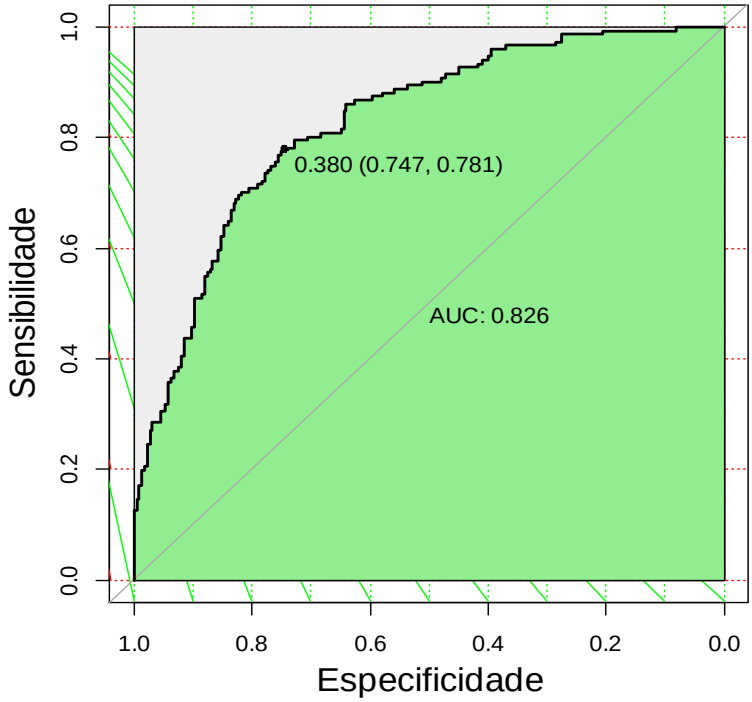
O teste de McNemar obtido apresenta um *pvalue* com o valor de 0,056 aproximadamente. Este valor conduz à não rejeição da hipótese nula, ou seja o número de casos positivos registados não se alterou significativamente em relação ao número de casos positivos preditos.

Tabela 71*:Parâmetros associados à capacidade de previsão*

Precisão	0.7633
Sensibilidade	0.6424
Especificidade	0.8444
Pos pred val	0.7348
Neg pred value	0.7787

A curva ROC obtida a partir dos valores de predição e dos valores observados está representada no gráfico 20 tal como o valor de 0,826 correspondente à AUC. O valor encontrado para a AUC (Figura 26) indica que este modelo tem uma boa capacidade de predição. Assim, é possível concluir que o modelo logístico múltiplo elaborado é estatísticamente adequado para a modelação da relação entre os dados das covariáveis correspondentes a questões clinicas apuradas previamente a um individuo e possibilidade de penetração de um tumor na cápsula prostática desse mesmo individuo permitindo uma análise ao risco de surgimento de cancro.

Figura 26: *Gráfico sobre a curva ROC*



CONCLUSÃO E CONSIDERAÇÕES FINAIS

Neste trabalho foram apresentadas e descritas de forma sucinta as principais definições e propriedades de modelos de regressão linear simples e múltipla e de modelos de regressão logística simples e múltipla. Também foram descritas definições de algumas estatísticas de teste associadas aos modelos de regressão. No final de cada parte, foram indicados e descritos comandos e sintaxes para a elaboração de modelos de regressão e realização de testes para análise dos mesmos, em ambiente R. No capítulo cinco procedeu-se à aplicação de modelos de regressão linear e logística a situações relacionadas com análise do risco, em que foram utilizados dados reais.

No primeiro exemplo relacionado com análise do risco utilizou-se um modelo de regressão linear simples para relacionar o número de pedidos de indeminização com montantes pagos no caso de acidentes com automóveis na Suécia. A partir do software R foi possível criar um MRLS a partir dos dados utilizados. O modelo obtido foi considerado válido para inferência estatística após verificar que os principais pressupostos foram satisfeitos. Em particular, os resíduos apresentaram variância constante e evidência de distribuição normal. No entanto, apesar de o MRLS ser o mais simples dos modelos apresentados neste trabalho, é de entre os quatro tipos de modelo de regressão, o modelo menos utilizado em situações de análise do risco. Atualmente, são indicadas várias fragilidades ao MRLS tais como a rígida assunção de normalidade e variância constante dos resíduos. Muitos problemas de análise do risco estão relacionados com dados temporais, como por exemplo valores mensais de precipitações ou valores diários de venda de ações. Na análise de dados temporais é muito provável que uma variável observada num dado período de tempo seja influenciada pelos valores registados num período anterior ou imediatamente anterior. Assim, ao ajustar estes dados através de um MRLS é frequente identificar autocorrelação nos resíduos. Isto é, a assunção de variância constante não é de todo satisfeita e, assim, a inferência a partir de deste modelo não será eficiente e credível.

Na parte 5.2. foi abordada a temática do risco de investimento em bolsa através da modelação linear múltipla de várias variáveis regressoras e o seu impacto na explicação dos valores da variável dependente “volume” contínua que designa o número total de transações em 3 meses. Em primeiro lugar procedeu-se à produção de diagramas e valores

de correlação entre as várias variáveis. E em seguida foi elaborado o MRLM cujas informações sobre coeficientes e estatísticas de teste sobre significância de coeficientes foram transcritas para a tabela 58. Numa primeira análise verificou-se que, apenas as variáveis relativas ao tempo médio entre transação e ao número total de transações em 3 meses eram significativas para a explicação do número total de acções vendidas em 3 meses. Assim tem-se:

$$VOLUME = 5{,}2765 - 0{,}3968 \times AVGT + 0{,}0014 \times NTRAN + 0{,}0119 \times SHARE$$

Procedi à retirada sucessiva de variáveis e verifiquei que o modelo apresentou os valores dos coeficientes de determinação R^2 e R^2 *ajustado*, aproximadamente iguais aos valores dos mesmos coeficientes referentes ao modelo inicial com todas as variáveis regressoras: 0,85 e 0,84 respetivamente. São valores elevados que indicam que mais de 84% dos dados das variáveis do modelo são explicadas pelo mesmo. Note-se que a retirada de variáveis pode ser realizada em função da significância estatística dos coeficientes mas deverá ter sempre em conta o contexto e especificidade do problema.

Para além do cálculo dos I.C. para os coeficientes, procedeu-se também à importante análise de resíduos através dos testes de Durbin-Watson e de Breusch-Pagan. Após a obtenção das respetivas estatísticas de teste observou-se que a estatística DW não conduz à rejeição da independência dos resíduos enquanto a estatística BP contradiz a conclusão anterior. Assim, a análise gráfica deverá ser conclusiva para a verificação de homocedasticidade. E neste caso verificou-se através do diagrama de Resíduos vs Valores ajustados (Figura 22) que que a variância não era constante. Logo, um dos principais pressupostos da regressão linear não se verifica. Existem procedimentos, tais como alteração de variável, para tentar contornar esta situação. Este procedimento foi realizado a titulo experimental em R e verificou-se que, neste caso não produziu qualquer efeito positivo do ponto de vista da homocedasticidade. Tal como tinha sido referido na análise ao MRLS, a modelação linear de dados temporais pode acarretar dificuldades quanto à independência dos resíduos. Para além na falta de indepêndencia nos resíduos, verificou-se também que a sua distribuição não provém de uma distribuição normal. Assim, a inferência estatística a partir deste modelo deve ser feita com bastante reserva ou não deve ser realizada em absoluto.

Um outro problema que pode afetar um MRLM é a correlação forte entre variáveis regressoras a incluir num dado modelo. Num MRLM, nenhuma ou poucas das covariáveis são estatísticamente significativas, no entanto a variância da variável dependente explicada pelas covariáveis é elevada (medida através de R^2). Tal acontece devido à sobreposição de informação de uma covariável em relação a outras covariáveis causada pela multicolinariedade. Assim, poderá ser difícil senão mesmo impossível distinguir a contribuição individual de cada covariável para a variância resultante (Y-K.Tu, 2005).

Os dois exemplos de aplicação de modelos logisticos para modelação em análise do risco têm a medicina como base de trabalho. Em 5.3. aplicou-se um modelo de regressão logística simples (MLG simples) para ajustar a variável explicativa idade e a variável dependente categórica DCC que designa a ocorrencia ou não de episódios de doença cardíaca coronária. Verificou-se que os coeficientes do modelo obtidos em R são estatísticamente significativos após a realização das estatísticas de Wald e Teste de Verossemelhança. E após a análise de resíduos de Pearson e de Deviance aceitou-se a hipótese de que o modelo é ajustado aos dados e assim este modelo revelou ser estatísticamente válido para inferência. A partir dos valores obtidos para *Odds*, verifica-se que, segundo este modelo, as chances de ocorrência de um episódio de doença cardíaca coronária são de 1,12 aproximadamente quando a idade aumenta uma unidade. É claro que a análise do risco de um episódio de DCC num dado individuo com determinada idade não assentará somente em valores previstos a partir deste modelo ou de um outro semelhante. Essa análise, deverá também assentar em exames clinicos especificos.

A regressão logística não assenta em pressupostos tão rigidos quanto à assunção de normalidade dos resíduos e homogeneidade da variância dos mesmos, no entanto apresenta algumas limitações. É um modelo cuja variável indpendente é categórica e tal fato limita o número de casos sobre os quais seria possível aplicar um modelo MLG. Ou seja, casos em que duas variáveis, regressora e dependente, sejam continuas não são passíveis da aplicabilidade de modelos logisticos. O modelo logistico também requer independência de dados, o que torna muito limitada a sua utilização em casos de dados de caráter temporal.

A utilização de um MLG múltiplo para modelação na análise do risco foi realizada na ultima parte do capítulo 5. Os dados reais provém de um estudo em que se pretende verificar se houve ou não penetração do tumor na cápsula prostática a partir de um

conjunto de sete variáveis dictómicas, contínuas e discretas. A partir do output obtido em R verificou-se que, através do teste de Wald, algumas variáveis não apresentavam significância estatística no modelo produzido. A retirada de variáveis como idade e raça dependem certamente, não só de critérios matemáticos mas também de critérios clinicos que não são abordados neste trabalho. Assim, optou-se por manter o modelo completo. Os valores correspondentes às chances de penetração da cápsula prostática pelo tumor foram calculados e apresentados na Tabela 68. O aumento de uma unidade na idade (AGE) e no volume (VOL) diminui ligeiramente as chances de penetração. No entanto, os I.C. (Tabela 69) calculados incluem, em relação às variáveis anteriores, o valor um no respetivo I.C., que é indicador que não existe associação entre estas variáveis regressoras e a variável dependente. A mesma conclusão é aplicável à variável DCAPS2. Assim, todas as outras variáveis tem influência nas chances de penetração. Por exemplo, individuos com mais uma unidade no teste GLEASON aumentam as suas chances de risco em 2,6 (aprox). Para aferir a qualidade ou bondade de ajuste, realizaram-se os testes aos resíduos e também o teste de Hosmer e Lemeshow. Todos os testes realizados apontaram para a evidencia estatística de que o modelo está ajustado corretamente. Finalmete, foi avaliada a capacidade de previsão do modelo através da curva ROC. O valor de corte foi definido em 0,5. Isto é, os individuos cuja probabilidade prevista está acima de 0,5 são definidos como positivos (*Y=1*) e os restantes como negativos (*Y=0*). Ao obter o valor de 0,86 para AUC, concluí-se que a capacidade de previsão do modelo é muito boa.

A modelação de variáveis regressoras e dependente através da regressão logística múltipla é realizada através de um modelo que, por um lado, é menos rigido quanto a pressupostos mas, que por outro lado, pode ser mais complexo. Tal como todos os modelos matemáticos, a regressão logística múltipla também tem limitações e desvantagens. O tamanho da amostra pode influenciar a qualidade do modelo, em particular no cálculo das chances de ocorrência ou não de um dado evento. A diminuição do tamanho da amostra pode provocar o aumento do enviesamento das chances e assim fornecer uma fraca estimativa do efeito populacional (Phil Reed, 2013). Este problema é habitualmente denominado "sparse-data problem". Para contornar este problema, devem ser utilizadas amostras grandes. Um outro problema é a modelação logística de eventos raros ou muito pouco frequentes tais como um evento catastrófico ou uma condição clínica rara (Phil Reed, 2013). A seleção de variáveis regressoras a incluir num estudo e no modelo

pretendido é outro dos grandes problemas que um investigador pode enfrentar. A escolha sobre a inclusão, ou não, de um dado fator pode afetar o valor quantitativo de predictabilidade sobre a ocorrência ou não de um dado evento. Segundo alguns autores, não existe definição clara sobre o número de variáveis independentes a incluir num modelo. No entanto, existem algums conselhos e linhas orientadoras que podem ter alguma utilidade. Uma delas refere que, se a especificidade e sensibilidade do modelo têm ambas uma precentagem acima de 80% então é bastante provável que as variáveis incluidas sejam as corretas.

Uma outra questão acerca dos modelos de regressão logística está relacionada com o conhecido e amplamente utilizado teste de Hosmer-Lemeshow para a qualidade de ajuste do modelo. Tal como referido na definição, este teste tem por base a utilização de grupos de probabilidades estimadas a partir do modelo. No entanto, não há um número definido para o número de grupos a utilizar e com a utilização de meios computacionais os investigadores notaram que a utilização de diferentes números de grupos afetava a estatística de teste que permite a conclusão sobre a qualidade de ajuste. Isto é, o valor de corte para a escolha dos grupos é uma limitação ao teste e para contornar esta limitação, vários autores propuseram uma série de testes baseados nos resíduos suavizados (D. W. Hosmer, 1997).

O estudo e modelação de dados de caráter temporal em que não há independência, é realizado através de conjuntos de observações ordenadas no tempo, habitualmente designadas de séries temporais. Esta abordagem permite que o construtor de um modelo económico use a inferência estatística para construir e testar equações que caracterizam as relações entre variáveis económicas e financeiras. Note-se que esta definição pressupõe que a série temporal contenha um componente estocástico (probabilístico) , o que acontece na maioria dos casos práticos.

Para além da regressão linear e regressão logistíca, e também das séries temporais, existe um conjunto de outras técnicas estatísticas utilizadas no âmbito da análise do risco. Uma das técnicas mais utilizadas atualmente são as simulações Monte Carlo. A simulação de Monte Carlo é uma metodologia estatística que se baseia na obtenção, por meios computacionais, de uma grande quantidade de amostragens aleatórias geradas para se obter resultados aproximados, o quanto possível, à realidade. Para além do ramo financeiro, as

simulações Monte Carlo de projetos em rede são cada vez mais utilizados por empresas de engenharia para analizar o risco de custo e cumprimento de prazos de uma dado projeto (Duffey, 1999). Na construção de modelos de avaliação de projetos de investimento, é frequente realizar a análise de vários cenários. Assim, o processo inicia-se com a definição de um "cenário-base". Tomando o este como ponto de partida, o processo continua com a criação de alternativas através do método de simulação de Monte Carlo, definindo um cenário pessimista – aquele que, verificando-se, afeta negativamente a viabilidade do projeto - e um cenário otimista que, caso venha a acontecer, cumprirá os objetivos inicialmente propostos.

No entanto, a regressão linear múltipla e regressão logística múltipla, e os testes associados a estes modelos, têm evoluido de forma a ultrapassar as principais limitações aqui mencionadas, proporcionando a modelação de situações cada vez mais complexas. As adaptações e alterações destes modelos e outras aplicações poderão ser analisadas num futuro trabalho.

Por último, refiro que os comandos e sintaxes utilizadas nos modelos de aplicação aos casos de análise do risco são análogos e muito semelhantes aos indicados nos capítulos anteriores. A utilização de uma ferramenta como o software R é indispensável para a realização de modelos de regressão linear e logístico e sua posterior análise. A constante atualização dos ficheiros estatisticos (*packages*) e matemáticos, em geral, fazem desta ferramenta um instrumento indispensável para a investigação científica, dado que se trata de uma linguagem utilizada de forma abrangente pela comunidade cientifica internacional.

Assim, concluo um trabalho que teve como propósito principal aplicar modelos de regressão linear e logística em problemas relacionados com a análise do risco através da utilização do software R e foi extremamente importante para o meu conhecimento, compreensão e aprofundamento neste ramo da Estatística Computacional.

Bibliografia

Agresti, A. (2007). *An Introduction to Categorical data Analysis.* New Jersey: Wiley.

Braga, A. C. (2000). Curvas ROC:Aspetos Funcionais e Aplicações. Braga: Universidade do Minho.

C.Montegomery, D. (2009). *Design and Analysis of Experiments.* Arizona state university: Wiley.

Christensen, R. (1997). *Log-Linear models and Logistic regression.* New York: Springer.

Crawley, M. J. (2007). *The R Book.* Willey.

D. W. Hosmer, T. H. (1997). A Comparison of Goodness-of-Fit Tests For The Logistic Rgression Model. *Statistics in Medicine, Vol. 16* , 965-980.

Dobson, A. J. (2002). *An introduction to Generalized Linear Models.* Boca Raton: Chapman&Hall/CRC.

Duffey, J. R. (1999). Statistical Dependence in Risk Analysis for project networks using Monte Carlo methods. *International Journal of Production Economics* , 17-29.

Fawcett, T. (2005). *An Introduction to ROC analysis.* Elsevier.

Ferreira, M. C. (2013). Modelos de Regressão: Uma aplicação em Medicina Dentária. Lisboa.

Figueira, C. V. (2006). *Modelos de Regressão Logistíca.* Porto Alegre.

H. Stryhn, J. C. (s.d.). Confidence intervals by the profile likelihood method, with applications in veterinary epidemiology. Charlottetown, Canada.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* Wiley.

John Neter, W. W. (1983). *Applied Linear Regression Models.* Illinois: Richard D. Irwin, Inc.

John O. Rawlings, S. G. (1998). *Applied Regression Analysis: A Research Tool, Second Edition.* New York: Springer.

Ken Kleinman, N. J. (2014). *SAS and R, Data Manegement, Statistical Analysis and Graphics.* CRC Press.

Landim, P. M. (2003). *Análise Estatística de Dados Geológicos,2ª edição revista e ampliada.* São Paulo: Editora UNESP.

Ligo, A. K. (2003). *A Simulação de Monte Carlo Como Instrumento de Análise do riscos e Selecção de Projectos.* São Paulo: Fundação Getúlio Vargas.

Loveridge, M., Rahman, A., & Babister, M. (2013). Probabilistic flood hydrographs using Monte Carlo simulation: potential impact to flood inundation. *20th International Congress on Modelling and Simulation.* Adelaide, Austrália.

McHugh ML.(2009). *The odds ratio: calculation, usage, and interpretation*. Biochem Med. 19:120–6. http://dx.doi.org/10.11613/BM.2009.011 .

Oliveira, S. d. (2013). Inferência e Análise de Resíduos e de Diagnóstico em Modelos Lineares Generalizados. Juiz de Fora.

Paula, G. A. (2010). *Modelos de Regressão com apoio computacional.* São Paulo: Instituto de Matemática e Estatística-Universidade de São Paulo.

Phil Reed, Y. W. (2013). Logistic regression for risk factor modelling in stuttering research. *Journal of Fluency Disorders* , 88-101.

Provete, D. B., Silva, F. R., & Souza, T. G. (Abril de 2011). Estatística aplicada à ecologia usando o R. São Paulo: UNESP.

Silva, A. C. (2011). *Análise Estatística de Inquéritos Online.* Braga: Universidade do Minho.

Simonovich, M. (1997). *Um Caso Prático de Uso de Ferramentas de Simulação de Risco na Subscrição de Apólices de Seguro.* São Paulo.

Souza, É. C. (2006). Análise de Influência local no modelo de regressão logística. *Dissertação (Mestrado)* . Piracicaba, São PAulo.

Sperandei, S. (2014). *Understanding logistic regression analysis.* Biochem Med (Zagreb). 2014 Feb; 24(1): 12–18. Published online 2014 Feb 15. doi: 10.11613/BM.2014.003.

Team, T. R. (2016). *R: A Language and Enviroment for Statistical Computing.* R Foundating for Statistical Computing.

Trevor Hastie, R. T. (2008). *The elements of Statistical Learning-Data Mining,Inference and Prediction.* Stanford: Springer.

Weisberg, S. (2005). *Aplied Linear Regression* (3ª edição ed.). Wiley.

Xuezheng Sun, Z. Y. (2008). *Generalized McNemar's Test for Homogeneity of the Marginal Distributions.*

Y-K.Tu, M. K. (2005). Problems of correlations between explanatory variables in multiple regression analyses in the dental literarture. *British Dental Journal* , 457-461.

ANEXOS

ANEXO I

Coeficiente de correlação de Pearson

O coeficiente de correlação de Pearson é uma medida que permite avaliar o grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor 0 (zero) significa que não há relação linear, o valor 1 indica uma relação linear perfeita e o valor -1 também indica uma relação linear perfeita mas inversa, ou seja quando uma das variáveis aumenta a outra diminui. Quanto mais próximo estiver de 1 ou -1, mais forte é a associação linear entre as duas variáveis.

O coeficiente de correlação de Pearson é normalmente representado pela letra **r** e a sua fórmula de cálculo é: $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

Coeficiente de correlação	*correlação*
$r = 1$	Perfeita positiva
$0{,}8 \leq r < 1$	Forte positiva
$0{,}5 \leq r < 0{,}8$	Moderada positiva
$0{,}1 \leq r < 0{,}5$	Fraca positiva
$0 < r < 0{,}1$	Ínfima Positiva
$r = 0$	Nula
$-0{,}1 < r < 0$	Ínfima negativa
$-0{,}5 < r \leq -0{,}1$	Fraca Negativa
$-0{,}8 < r \leq -0{,}5$	Moderada Negativa
$-1 < r \leq -0{,}8$	Forte Negativa
$r = -1$	Perfeita Negativa

ANEXO II

Comandos utilizados em R

Exemplo 5.1

Distribuição das variáveis X e Y

```
hist(DadosSeg$X,main="Histograma sobre Número de pedidos de indemnização",xlab="Pedidos de
Indemnização",ylab="freq. absoluta",col="lightblue",cex.main=2,cex.lab=1.5,ylim=c(0,40))

summary(DadosSeg$X)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

    0.0     7.5    14.0    22.9    29.0   124.0

hist(DadosSeg$Y,main="Histograma sobre montantes pagos por região",xlab="Coroas Suecas(em milhares)",ylab="freq.
absoluta",col="lightgreen",cex.main=2,cex.lab=1.5,xlim=c(0,500),ylim=c(0,25))

summary(DadosSeg$Y)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

   0.00   38.85   73.40   98.19  140.00  422.20
```

Coeficiente de correlação de Pearson

```
cor(DadosSeg$Y,DadosSeg$X)
```

Diagrama de dispersão X vs Y

```
plot(DadosSeg$Y~DadosSeg$X,main="Pedidos de Indemnização vs Montantes pagos",sub="(em 63
regiões)",xlab="Pedidos de indemnização",ylab="Montantes pagos (em
milhares)",cex.main=2,cex.lab=1.5,pch=19,col="orange")

mtext("Coef.Pearson r=0,91",line=-18,adj=1)
```

Modelo de regressão linear simples

```
> rl<-lm(DadosSeg$Y~DadosSeg$X)

> summary(rl)

Call:

lm(formula = DadosSeg$Y ~ DadosSeg$X)

Residuals:

    Min      1Q  Median      3Q     Max

-86.561 -24.051  -0.347  23.432  83.977

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  19.9945     6.3678    3.14   0.0026 **

DadosSeg$X    3.4138     0.1955   17.46   <2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.94 on 61 degrees of freedom

Multiple R-squared:  0.8333,    Adjusted R-squared:  0.8306
```

F-statistic: 305 on 1 and 61 DF, p-value: < 2.2e-16

Teste de Breusch-Pagan

studentized Breusch-Pagan test

data: rl

BP = 1.7109, df = 1, p-value = 0.1909

Teste de Normalidade Shapiro Wilk

> shapiro.test(res)

Shapiro-Wilk normality test

data: res

W = 0.97496, p-value = 0.2262

Intervalos de confiança para os parâmetros

> confint(rl,interval="confidence")

2.5 % 97.5 %

(Intercept) 7.261369 32.727602

DadosSeg$X 3.022966 3.804681

Análise de resíduos

> summary(res)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-86.5600 -24.0500 -0.3465 0.0000 23.4300 83.9800

par(mfrow=c(2,2))

> hist(res,bg="lightblue",main="Histograma dos Resíduos",xlab="resíduos",ylab="freq.absoluta",cex.main=1.7)

> qqPlot(rl,main="Gráfico QQ",ylab="Resíduos",xlab="Quantis da normal",col="blue",cex.main=1.7)

> par(mfrow=c(2,2))

> hist(res,bg="lightblue",main="Histograma dos Resíduos",xlab="resíduos",ylab="freq.absoluta",cex.main=1.7)

> plot(res~fit,main="Resíduos vs Valores ajustados",ylab="Resíduos",xlab="Valores ajustados",col="blue",cex.main=1.7)

> abline(h=0)

> qqnorm(res,main="Normal QQ",ylab="residuos padronizados",xlab="Quantis Teóricos",cex.main=1.7)

> qqline(res)

> plot(leverage.varname,main="Leverage",cex.main=1.7)

Exemplo 5.2

Correlação de variáveis

```
> library(psych)

> pairs.panels(dados5.2)
```

Modelo "completo"

```
> modelo5.2<-lm(VOLUME~AVGT+NTRAN+PRICE+SHARE+VALUE+DEBEQ,data=dados5.2)

> summary(modelo5.2)

Call:

lm(formula = VOLUME ~ AVGT + NTRAN + PRICE + SHARE + VALUE +

   DEBEQ, data = dados5.2)

Residuals:

    Min      1Q  Median      3Q     Max

-12.5106  -2.6490  -0.5479   1.8021  15.8238

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  6.1570658  1.9567026   3.147  0.00210 **

AVGT        -0.4436086  0.1498631  -2.960  0.00373 **

NTRAN        0.0013765  0.0001748   7.873 2.03e-12 ***

PRICE       -0.0137975  0.0227635  -0.606  0.54562

SHARE        0.0090415  0.0073415   1.232  0.22060

VALUE        0.0811754  0.1134958   0.715  0.47591

DEBEQ        0.0608825  0.0592697   1.027  0.30646

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 4.237 on 116 degrees of freedom

Multiple R-squared:  0.849,     Adjusted R-squared:  0.8412

F-statistic: 108.7 on 6 and 116 DF,  p-value: < 2.2e-16
```

ANOVA sobre modelo completo

```
> anova(modelo5.2)

Analysis of Variance Table

Response: VOLUME
```

```
          Df Sum Sq Mean Sq  F value  Pr(>F)
AVGT       1 6273.9  6273.9 349.5187 < 2e-16 ***
NTRAN      1 5336.7  5336.7 297.3094 < 2e-16 ***
PRICE      1    0.5     0.5   0.0272 0.86940
SHARE      1   71.1    71.1   3.9636 0.04884 *
VALUE      1    8.7     8.7   0.4847 0.48767
DEBEQ      1   18.9    18.9   1.0552 0.30646
Residuals 116 2082.2    18.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo “reduzido”

```
> modelo5.2.1<-stepAIC(modelo5.2)
Start:  AIC=361.97
VOLUME ~ AVGT + NTRAN + PRICE + SHARE + VALUE + DEBEQ
       Df Sum of Sq    RSS    AIC
- PRICE  1      6.59 2088.8 360.36
- VALUE  1      9.18 2091.4 360.51
- DEBEQ  1     18.94 2101.1 361.08
- SHARE  1     27.23 2109.4 361.56
<none>               2082.2 361.97
- AVGT   1    157.28 2239.5 368.92
- NTRAN  1   1112.68 3194.9 412.63
Step:  AIC=360.36
VOLUME ~ AVGT + NTRAN + SHARE + VALUE + DEBEQ
       Df Sum of Sq    RSS    AIC
- VALUE  1      3.50 2092.3 358.56
- DEBEQ  1     19.06 2107.9 359.47
<none>               2088.8 360.36
- SHARE  1     40.31 2129.1 360.71
- AVGT   1    150.79 2239.6 366.93
- NTRAN  1   1203.97 3292.8 414.34
```

```
Step:  AIC=358.56
VOLUME ~ AVGT + NTRAN + SHARE + DEBEQ

       Df Sum of Sq    RSS    AIC
- DEBEQ  1     18.63 2110.9 357.65
<none>               2092.3 358.56
- SHARE  1     71.64 2163.9 360.70
- AVGT   1    150.98 2243.3 365.13
- NTRAN  1   1466.93 3559.2 421.91
Step:  AIC=357.65
VOLUME ~ AVGT + NTRAN + SHARE

       Df Sum of Sq    RSS    AIC
<none>               2110.9 357.65
- SHARE  1     70.55 2181.5 359.70
- AVGT   1    146.43 2257.4 363.90
- NTRAN  1   1463.03 3574.0 420.42
> modelo5.2.1<-lm(VOLUME~AVGT+NTRAN+SHARE,data=dados5.2)
> summary(modelo5.2.1)
Call:
lm(formula = VOLUME ~ AVGT + NTRAN + SHARE, data = dados5.2)
Residuals:
    Min      1Q  Median      3Q     Max
-12.9730  -2.5172  -0.5541   1.7875  15.4196
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.276477   1.357014   3.888 0.000167 ***
AVGT        -0.396753   0.138090  -2.873 0.004814 **
NTRAN        0.001426   0.000157   9.082 2.79e-15 ***
SHARE        0.011908   0.005971   1.994 0.048402 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 4.212 on 119 degrees of freedom

Multiple R-squared: 0.8469, Adjusted R-squared: 0.8431

F-statistic: 219.5 on 3 and 119 DF, p-value: < 2.2e-16

ANOVA sobre modelo reduzido

> anova(modelo5.2.1)

Analysis of Variance Table

Response: VOLUME

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AVGT	1	6273.9	6273.9	353.6793	<2e-16	***
NTRAN	1	5336.7	5336.7	300.8485	<2e-16	***
SHARE	1	70.6	70.6	3.9773	0.0484	*
Residuals	119	2110.9	17.7			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Intervalos de confiança para os coeficientes

> confint(modelo3.2.1)

	2.5 %	97.5 %
(Intercept)	2.589454e+00	7.963499560
AVGT	-6.701844e-01	-0.123321113
NTRAN	1.114937e-03	0.001736686
SHARE	8.495984e-05	0.023731533

Resíduos do modelo reduzido

Res5.2.1<-residuals(modelo5.2.1)

> hist(res5.2.1)

Teste de normalidade sobre resíduos

> shapiro.test(res5.2.1)

Shapiro-Wilk normality test

data: res5.2.1

W = 0.9481, p-value = 0.0001287

Testes Breusch- Pagan e Durbin–Watson

```
> library(lmtest)

Loading required package: zoo

Attaching package: ‘zoo’

The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric

> bptest(modelo3.2.1)

        studentized Breusch-Pagan test

data:  modelo5.2.1

BP = 20.909, df = 3, p-value = 0.00011

> dwtest(modelo3.2.1)

        Durbin-Watson test

data:  modelo5.2.1

DW = 1.8916, p-value = 0.2738

alternative hypothesis: true autocorrelation is greater than 0
```

Gráficos sobre os resíduos

```
par(mfrow=c(2,2))

> plot(modelo5.2.1)
```

Exemplo 5.4

Introdução dos dados em ambiente R e formatação dos mesmos

```
>getwd()

>dados<-read.csv("prost.csv",header=TRUE,sep=";")

> dados$RACE<-factor(dados$RACE)

> dados$DPROS<-factor(dados$DPROS)

> dados$DCAPS<-factor(dados$DCAPS)

> dados1<-na.omit(dados)# Anulamento das linhas que contém "células" vazias1
```

Obtenção do modelo logístico múltiplo

```
>modelo3.4<glm(CAPSULE~AGE+RACE+DPROS+DCAPS+PSA+VOL+GLEASON,family=binomial(logit),data=dados1)

> summary(modelo3.4)#sumário de informações relevantes sobre o modelo logistico

Call:
glm(formula = CAPSULE ~ AGE + RACE + DPROS + DCAPS + PSA + VOL +
```

```
   GLEASON, data = dados1)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.8290 -0.3139 -0.1165  0.3962  0.9764

Coefficients:
             Estimate    Std. Error  t value  Pr(>|t|)
(Intercept) -0.566600   0.261584   -2.166    0.03095 *
AGE         -0.001813   0.003436   -0.528    0.59806
RACE2       -0.087278   0.075140   -1.162    0.24618
DPROS2       0.107704   0.057083    1.887    0.05998 .
DPROS3       0.253526   0.062047    4.086 5.40e-05 ***
DPROS4       0.218618   0.077717    2.813  0.00517 **
DCAPS2       0.106174   0.076554    1.387  0.16631
PSA          0.004156   0.001263    3.289  0.00110 **
VOL         -0.002006   0.001203   -1.668  0.09616 .
GLEASON      0.144381   0.022438    6.435 3.88e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1766695)

    Null deviance: 90.359  on 375  degrees of freedom
Residual deviance: 64.661  on 366  degrees of freedom
AIC: 427.12

Number of Fisher Scoring iterations: 2
```

Retirada de variáveis ao modelo completo

```
> library(MASS)
> step <- stepAIC(modelo)

Start:  AIC=427.12
CAPSULE ~ AGE + RACE + DPROS + DCAPS + PSA + VOL + GLEASON

           Df  Deviance    AIC
- AGE       1    64.710  425.41
- RACE      1    64.899  426.50
- DCAPS     1    65.001  427.09
<none>              64.661  427.12
- VOL       1    65.153  427.97
- PSA       1    66.573  436.07
- DPROS     3    67.992  440.00
- GLEASON   1    71.976  465.42

Step:  AIC=425.41
CAPSULE ~ RACE + DPROS + DCAPS + PSA + VOL + GLEASON

           Df  Deviance    AIC
- RACE      1    64.938  424.73
- DCAPS     1    65.046  425.35
<none>              64.710  425.41
- VOL       1    65.248  426.52
- PSA       1    66.651  434.51
- DPROS     3    68.152  438.89
- GLEASON   1    71.984  463.46

Step:  AIC=424.73
CAPSULE ~ DPROS + DCAPS + PSA + VOL + GLEASON

           Df  Deviance    AIC
- DCAPS     1    65.260  424.59
```

```
<none>              64.938   424.73
- VOL          1    65.535   426.17
- PSA          1    66.743   433.04
- DPROS        3    68.315   437.79
- GLEASON      1    72.345   463.34

Step:  AIC=424.59
CAPSULE ~ DPROS + PSA + VOL + GLEASON

              Df   Deviance    AIC
<none>              65.260    424.59
- VOL          1    65.957    426.58
- PSA          1    67.368    434.54
- DPROS        3    68.937    439.20
- GLEASON      1   73.409 466.83
```

Obtenção do modelo reduzido

```
> summary(modelo3.4.1)

Call:
glm(formula = CAPSULE ~ DPROS + PSA + GLEASON, family = binomial(logit),
    data = dados1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3539  -0.7585  -0.4466   0.9134   2.4599

Coefficients:
                Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)    -8.129843   1.057489    -7.688    1.50e-14 ***
DPROS2          0.768634   0.356055     2.159    0.03087 *
DPROS3          1.548911   0.371510     4.169    3.06e-05 ***
DPROS4          1.425491   0.449167     3.174    0.00151 **
PSA             0.027324   0.009398     2.907    0.00364 **
GLEASON         0.993644   0.161176     6.165    7.05e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 506.59  on 375  degrees of freedom
Residual deviance: 381.12  on 370  degrees of freedom
AIC: 393.12

Number of Fisher Scoring iterations: 5
```

Tabelas ANOVA sobre os modelo completo (3.4) e incompleto (3.4.1)

```
> anova(modelo3.4,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: CAPSULE

Terms added sequentially (first to last)


          Df   Deviance Resid. Df   Resid. Dev  Pr(>Chi)
NULL      375  506.59
AGE       1      0.613      374     505.97      0.4336454
RACE      1      0.040      373     505.93      0.8421371
DPROS     3     39.607      370     466.33      1.291e-08 ***
```

```
DCAPS     1    12.934      369    453.39    0.0003227 ***
PSA       1    34.428      368    418.97    4.423e-09 ***
VOL       1     3.542      367    415.42    0.0598176 .
GLEASON   1    40.888      366    374.54    1.612e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(modelo3.4.1,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: CAPSULE

Terms added sequentially (first to last)


          Df  DevianceResid. Df  Resid. Dev  Pr(>Chi)
NULL      375     506.59
DPROS       3     39.538      372    467.05    1.335e-08 ***
PSA         1     38.464      371    428.59    5.579e-10 ***
GLEASON     1     47.463      370    381.12    5.605e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

“Coeficientes de determinação”

```
> library(pscl)
Loading required package: lattice
Classes and Methods for R developed in the

Political Science Computational Laboratory

Department of Political Science

Stanford University

Simon Jackman

hurdle and zeroinfl functions by Achim Zeileis

> pR2(modelo3.4)
       llh         llhNull          G2         McFadden       r2ML       r2CU
-202.5600020 -265.4708555 125.8217070    0.2369784    0.2843988    0.3760079

> pR2(modelo3.4.1)
       llh         llhNull          G2         McFadden       r2ML       r2CU
-190.5612441 -253.2936721 125.4648559    0.2476668    0.2837193    0.3833730
```

Teste de Verossimilhança

```
logLik(modelo3.4)
'log Lik.' -187.2676 (df=10)

> logLik(basico)
'log Lik.' -253.2937 (df=1)

> logLik(modelo3.4.1)
'log Lik.' -190.5612 (df=6)

> G.2=-2*((-190.5612)-(-187.2676))

> 1-pchisq(G.2,df=6)

0.3607134
```

Valor das *Odds* e respetivos I.C. (95%)

```
 > exp(cbind(OR = coef(modelo3.4), confint(modelo3.4)))
Waiting for profiling to be done...
                    OR          2.5 %          97.5 %
(Intercept) 0.0009429248 3.610543e-05     0.02098054
AGE         0.9882626559 9.506081e-01     1.02723645
RACE2       0.5213321247 2.004140e-01     1.29034335
DPROS2      2.0754679391 1.039749e+00     4.27298011
DPROS3      4.5244022085 2.195562e+00     9.68418384
DPROS4      4.0035231965 1.633040e+00    10.06392121
DCAPS2      1.6362212815 6.696848e-01     4.18783042
PSA         1.0303387699 1.011255e+00     1.05222986
VOL         0.9886034646 9.732039e-01     1.00355112
GLEASON     2.6182586155 1.910451e+00     3.68074375
```

Gráfico sobre resíduos

```
 > plot(res.3.4,main="Resíduos",ylab="Resíduos",type="pearson")
Warning message:
In plot.xy(xy, type, ...) :
  plot type 'pearson' will be truncated to first character
> abline(h=0,col="red")
```

Gráficos sobre resíduos

```
> library(car)

Attaching package: ‘car’

The following object is masked from ‘package:psych’:

    logit

> residualPlots(modelo3.4)
            Test stat Pr(>|t|)
AGE             0.098    0.754
RACE               NA       NA
DPROS              NA       NA
DCAPS              NA       NA
PSA             0.069    0.793
VOL             0.006    0.936
GLEASON         0.312    0.577
```

Testes de Pearson e *Deviance* sobre os resíduos

```
> 1-pchisq(sum((res.3.4)^2),df=380-(9+1))#Teste para os resíduos Deviance
[1] 0.4244495

> res.3.4Pearson<-residuals(modelo3.4,type="pearson")
> 1-pchisq(sum((res.3.4Pearson)^2),df=380-(9+1))#Teste para os resíduos de Pearson
[1] 0.6799036
```

Teste de Hosmer-lemeshow

```
> library(ResourceSelection)

ResourceSelection 0.2-6        2016-02-15
> hl<-hoslem.test(dados1$CAPSULE,fitted(modelo3.4),g=10)
> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  dados1$CAPSULE, fitted(modelo3.4)

X-squared = 5.165, df = 8, p-value = 0.7398
```

Tabela de Contingência 2X2

```
> library(SDMTools)

> confusion.matrix(dados1$CAPSULE, pred3.4, threshold = 0.5)
           obs
pred     0   1
   0   190  54
   1    35  97

attr(,"class")
[1] "confusion.matrix"
```

Parâmetros associados à capacidade de previsão

```
caps.pred<-predict(m.completo,type="response")
> caps.v<-ifelse(dados.1$CAPSULE,1,0)
> roc.plot(caps.v, caps.pred)
Error: could not find function "roc.plot"
> confusionMatrix(caps.predt(0.5), caps.v)
Confusion Matrix and Statistics

              Reference
Prediction   0     1
        0      190  54
        1       35  97

            Accuracy : 0.7633
              95% CI : (0.717, 0.8054)
    No Information Rate : 0.5984
    P-Value [Acc > NIR] : 1.111e-11

              Kappa : 0.4971
 Mcnemar's Test P-Value : 0.05639

          Sensitivity : 0.8444
          Specificity : 0.6424
       Pos Pred Value : 0.7787
       Neg Pred Value : 0.7348
           Prevalence : 0.5984
       Detection Rate : 0.5053
   Detection Prevalence : 0.6489
      Balanced Accuracy : 0.7434

       'Positive' Class : 0
```

Gráfico e cálculo da área sob a curva ROC

```
> roc3.4<-plot.roc(dados1$CAPSULE,fitted(modelo3.4))

> plot(roc3.4, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2),xlab="Especificidade",cex.lab=1.5,
+ grid.col=c("green", "red"), max.auc.polygon=TRUE,ylab="Sensibilidade",
+ auc.polygon.col="lightgreen", print.thres=TRUE)

Call:
plot.roc.default(x = dados1$CAPSULE, predictor = fitted(m.completo))

Data: fitted(m.completo) in 225 controls (dados1$CAPSULE 0) < 151 cases (dados1$CAPSULE 1).
Area under the curve: 0.826
```