**Universidade Aberta**


Mestrado em Estatística, Matemática e Computação
(Ramo – Estatística Computacional)

# Regressão Linear e Árvores de Regressão:

# Previsão do desempenho na disciplina de Matemática

**Marcolino José Ribeiro Sobral**

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

Orientador:

**Professor Doutor Amílcar Manuel do Rosário Oliveira**

(Professor Auxiliar da Universidade Aberta)

Lisboa 2014

**Agradecimentos**

Ao Professor Doutor Amílcar Manuel do Rosário Oliveira, meu orientador, agradeço pelo apoio, paciência, confiança e disponibilidade demonstrada ao longo deste caminho.

Aos professores da parte curricular do mestrado pela disponibilidade e incentivo na realização de todos os trabalhos propostos.

Aos meus filhos e à minha esposa, agradeço o seu apoio incondicional, a sua paciência e a compreensão demonstrada pela minha ausência em diversos momentos.

A toda a minha família, em especial à Lú.

À D.ra Teresa Ramos.

A todos os intervenientes nos questionários.

A todos os colaboradores na Linguagem R

**RESUMO**

A presente dissertação tem como objetivo primordial prever o desempenho da disciplina de Matemática, analisando dados recolhidos – em escolas secundárias - através de um questionário. Os métodos utilizados serão a regressão linear múltipla – método "clássico" e paramétrico – e as árvores de regressão binárias – método "contemporâneo" e não paramétrico.

Este estudo tem como objetivo dar a conhecer e compreender a capacidade preditiva das árvores de regressão - por si só ou em alternativa a outros métodos - e a eficiência computacional, aplicadas às classificações obtidas, na disciplina de Matemática, quando relacionadas com vários fatores. Considera-se como variável alvo e como variável dependente, a média na disciplina de matemática. As observações serão modeladas, explorando as técnicas de visualização gráficas e analíticas do software *open source* R, ao qual se faz uma introdução.

**Palavras-chave**: Desempenho na Matemática, Regressão Linear Múltipla, Árvores de Regressão Binárias, Software R

**ABSTRACT**

The primordial aim of the present dissertation is to predict the performance of the subject of Mathematics by analyzing collected data in secondary schools conducted through a questionnaire. The methods used will be the linear regression multiple - the “classic” method and parametric - and the binary regression trees - "contemporary" method and nonparametric.

This study aims to inform and understand the predictive ability of regression trees - by itself or as an alternative to other methods - and the computational efficiency, applied to the ratings obtained in Mathematics, when related to several factors. It is considered as a variable target and the dependent variable, the average in the subject of mathematics. The observations will be modeled by exploiting the techniques of graphical and analytical view of the open source software R, which is an introduction.

**Keywords**: Performance in Mathematics, Multiple Linear Regression, Binary Regression Trees, Software R.

# Índice

**Simbologia e notações**

| | |
|---|---|
| $\mathbf{Y}$ | Variável dependente ou explicada |
| $\mathbf{Y_i}$ | Valor observado da variável dependente |
| $\mathbf{X}$ | Variável independente ou explicativa |
| $\mathbf{X_i}$ | Valor observado da variável independente |
| $\mathbf{I_n}$ | Matriz identidade de ordem n |
| $\boldsymbol{\beta_1, \beta_2,...,\beta_k}$ | Parâmetros populacionais |
| $\hat{\beta}_j$ | Estimador de $\beta_j$ |
| **OLS** | Estimador dos mínimos quadrados |
| **E** | Esperança |
| **var** | Variância |
| **cov** | Covariância |
| $\Sigma(\hat{\beta})$ | Matriz de variância-covariância das estimativas dos parâmetros |
| $\boldsymbol{\sigma^2}$ | Variância das variáveis aleatórias residuais |
| $S^2$ | Estimativa de $\sigma^2$ |
| **S** | Erro padrão da regressão |
| $\mathbf{H_0}$ | Hipótese nula |
| $\mathbf{H_1}$ | Hipótese alternativa |
| **R.C.** | Região crítica (R.C.) |
| $\alpha$ | Nível de significância |
| $\mathbf{R^2}$ | Coeficiente de determinação $R^2$ |
| **TSS** | Soma dos quadrados totais |
| **ESS** | Soma dos quadrados dos resíduos |
| **RSS** | Soma dos quadrados da regressão |
| **ANOVA** | Análise de Variância |

| | |
|---|---|
| $\overline{R}^2$ | Coeficiente de determinação ajustado |
| **r** | Coeficiente de correlação parcial |
| **CART** | Classification and Regression Trees |
| ***d (X)*** | Preditor de *Y* |
| ***N (t)*** | Número de nós t |
| ***T, T'*** | Árvores |
| $\widetilde{T}$ | Conjunto de nós terminais *T* |
| **r(t)** | Critério de eficácia local |
| **R(t)** | Critério de eficácia global |
| $c_i^j$ | Corte tirado de $X_i$ |
| $t_e$ | Nó terminal esquerdo |
| $t_d$ | Nó terminal direito |
| $R_\alpha$ **(T),** | Custo de complexidade de uma árvore T |
| **cp** | Custo-complexidade |
| $\mu$ | Média |
| $\sigma$ | Desvio padrão |
| $\varepsilon$ | Resíduos |
| X' | Transposta de X |
| $\theta_n$ | Vetor nulo |

**Índice de tabelas**

**Índice de gráficos**

**Índice de figuras**

# INTRODUÇÃO

A importância da matemática no desenvolvimento dos indivíduos é amplamente aceite pela comunidade em geral. No entanto, o processo de ensino/aprendizagem da Matemática está por vezes associado a grandes percentagens de insucesso, que ocorrem nos vários graus de ensino. Esta é mesmo a disciplina que regularmente apresenta percentagens elevadas de baixas classificações, facto muitas vezes bastante focalizado até pelos órgãos de comunicação social.

Diversos estudos têm sido feitos com o intuito de compreender, interpretar e justificar os desempenhos dos alunos, assim como a relação dos alunos com a matemática e a sua aprendizagem ((Tavares, L. V., Graça, P.M. e Tavares, M. M.V., 2002) ou (Ramos,M., 2003), por exemplo).

Os métodos de regressão CART ("Classification and Regression Trees") apresentam formas alternativas para exploração dos dados e podem contribuir para uma melhor compreensão de alguns fenómenos nesta área.

Estes métodos devem ser vistos como uma nova ferramenta, flexível, não paramétrica, proporcionando ao investigador explorar conjuntos de dados numa perspetiva inovadora e têm sido utilizados em áreas como a biologia, a medicina, a biometria e também nas ciências sociais e ciências da educação. Para além da sua capacidade preditiva, têm um forte poder descritivo, o qual permite compreender quais das variáveis em estudo estão mais relacionadas com a génese da temática principal abordada.

Por outro lado apresenta a vantagem de não necessitar à priori do conhecimento de todos os atributos. Esta evidência tem vantagem, nomeadamente em problemas nos quais os valores dos atributos são difíceis de medir ou cuja medição acarreta custos elevados. Para prever o valor resposta de um caso temos apenas de recolher um a um os valores dos atributos que aparecem no seu percurso de descida na árvore (binária). A utilização e interpretação simples de certas árvores são outros dos atrativos da utilização das mesmas.

Este trabalho de dissertação tratará de expor dois métodos de regressão: regressão por árvore binária e regressão linear múltipla.

A estrutura é feita em duas partes - sendo a primeira, a apresentação dos conceitos teóricos básicos, e a segunda, a aplicação prática dos métodos expostos, tratando dados recolhidos duma situação real. Estas partes serão divididas em cinco capítulos:

No capítulo 1, descreve-se a fundamentação da regressão linear múltipla.

No capítulo 2, mostra-se a metodologia das árvores de regressão, focando-nos nas árvores binárias (recursivas) derivadas do procedimento CART.

No capítulo 3, apresenta-se a Linguagem R, que será o suporte do software utilizado.

No capítulo 4, aplica-se os métodos - enunciados nos capítulos 1 e 2 – com o software R, na avaliação do desempenho na disciplina de Matemática, analisando e discutindo os resultados. Serão selecionadas seis variáveis quantitativas – uma dependente ( ou explicada) e cinco independentes (ou explicativas) - e através dos métodos já referidos – tendo em conta a validade dos mesmos para as observações – procuram-se relações entre a variável dependente e as restantes, validadas ( ou não) por testes estatísticos. A validação é feita pela observação gráfica e pela análise de parâmetros.

No capítulo 5, tiram-se - breves – conclusões sobre a aplicação dos métodos referidos e trabalho futuro.

# PARTE I

## 1.REGRESSÃO LINEAR

### 1.1. Introdução

Muitos dos problemas estatísticos que se colocam nas aplicações, têm como objetivo estudar o relacionamento entre diversas variáveis, por hipótese de um dos dois tipos:

a) 1º tipo: supõe-se que existe uma forma de dependência de uma variável aleatória em relação a outra (ou a um conjunto de outras) que podem ser observadas sem erros (significativos). Logo, o tratamento do problema é efetuado sob a suposição de que há apenas uma variável aleatória dependendo de um conjunto de variáveis determinísticas (ou assumidas como tal, nesse tratamento).
b) 2º tipo: supõe-se que o objetivo do estudo é o relacionamento de diversas variáveis aleatórias entre si, sem especificar à partida uma forma de dependência de uma(s) variável (variáveis) em relação a outras.

Os modelos adequados ao estudo deste 2º tipo são chamados de modelos de correlação, enquanto que os modelos de regressão se utilizam no estudo de problemas do 1º tipo. O trabalho que apresentamos refere-se aos modelos de regressão.

Comecemos por referir que o termo **regressão** foi introduzido por Francis Galton no seu artigo *"Regression Towards Mediocrity in Herditary Stature"* publicado no *Journal of Antropological Institute* em 1885. Galton analisou a relação entre duas variáveis: a altura média dos pais e a média da altura dos filhos adultos e concluiu, como esperado, que pais altos tendem, em geral, a ter filhos altos e pais baixos tendem a ter filhos baixos. Contudo Galton também verificou que as alturas dos filhos se concentram mais em torno da altura média do que a altura dos pais em relação à respetiva média. Ou seja, em média, pais altos têm filhos altos, mas os filhos não são tão altos como os pais e pais baixos têm filhos baixos mas estes não são tão baixos como os pais. Galton designou a

linha que descreve a relação média entre as duas variáveis como linha de regressão e daí emergiu o método de análise de regressão. Este tipo de modelo é designado por modelo de **regressão linear simples** uma vez que define uma relação linear entre a variável dependente (altura dos filhos adultos) e uma variável explicativa (altura média dos pais). Se em vez de uma forem incorporadas várias variáveis explicativas, o modelo passa a designar-se modelo de **regressão linear múltipla**.

A explicação dada pelas variáveis independentes utilizadas num modelo de regressão não é normalmente completa, isto é, o conjunto das variáveis consideradas não dá conta de toda a variância da variável dependente. Para obviar a este problema, acrescentamos uma variável não observada, designada por variável residual e que representa a variância da variável dependente que não é explicada pelo conjunto das variáveis explicativas consideradas. O modelo deixa, assim, de ser um modelo determinístico passando a ser um modelo estocástico.

Tem-se assim:

$$Dados = Modelo + Resíduo$$

Em termos estatísticos, o objetivo da análise de regressão é a estimação dos parâmetros do modelo (representado por uma equação) que maximize a explicação da variação da variável dependente.

A estimação consiste em determinar os valores dos parâmetros do modelo: os coeficientes das variáveis explicativas e os parâmetros da variável aleatória residual. Estes últimos podem ser vários mas, normalmente, estamos interessados no conhecimento de dois parâmetros: a média ($\mu_\varepsilon$) e o desvio padrão ($\sigma_\varepsilon$). Vários métodos podem ser utilizados para estimar os parâmetros e a estimação de cada parâmetro, qualquer que seja o método utilizado, pressupõe sempre a recolha de dados sobre as variáveis em estudo (observações). À função que permite propor um valor para o parâmetro dá-se o nome de estimador desse parâmetro. O valor proposto pelo estimador, em relação a um certo conjunto de observações tem o nome de estimativa.
Após a estimação dos parâmetros, a equação da regressão pode ser utilizada, com várias finalidades, por exemplo para simulação e previsão.

## 1.2. A construção do modelo

Perante os dados observados, torna-se necessário construir o modelo que melhor os represente, segundo determinado critério, de modo que a partir desse modelo se possa, por exemplo, estabelecer resultados futuros (previsão).
Na análise de regressão, e no sentido de alcançar um "bom modelo", segue-se, em geral, a seguinte metodologia:

Etapa 1 – Formulação do modelo
Etapa 2 – Estimação
Etapa 3 – Avaliação da qualidade do modelo ajustado

- Formulação do modelo
A primeira etapa assume uma elevada importância, no sentido de que dela depende grande parte do sucesso ou não do problema a tratar. A construção de um modelo original depende, obviamente, de diversos fatores, tais como, o conhecimento do problema, estudo dos dados conhecidos, etc.

Face aos dados observados $(x_i,y_i)$, i=1,2,...,n podemos constatar, por exemplo, a existência, por exemplo, de um dos seguintes modelos:

$L(x)=\beta_1+\beta_1 x$ linear

$P(x) = \beta_1+\beta_1 x+\ldots+\beta_n x^p$ polinomial

(notando que x pode ser vetorial, o que nos conduz à análise multivariada. Y é a variável dependente ou explicada e X a variável independente ou explicativa).

O modelo pode assim ser: $G\in\{L, P,...\}$

Como referimos na introdução, normalmente Y é função do X mais uma perturbação, ε, pelo que podemos escrever o modelo da forma: $G(x)+\varepsilon$, com $x=x_1,x_2,...,x_n$.

Em termos de estudo de uma regressão veremos mais à frente que propriedades se costuma supor para o termo perturbação. Assumiremos que $\varepsilon \sim N(0,\sigma)$ – a variável aleatória $\boldsymbol{\varepsilon}$ tem distribuição normal com média **0** e desvio padrão $\sigma$.

Depois da formulação do modelo, devemos, ainda nesta fase, verificar se não existe qualquer inconsistência em relação com o conhecimento teórico e/ou empírico do problema e verificar se esse modelo poderá ou não contribuir para o objetivo da investigação.

- Estimação

Na etapa da estimação dos parâmetros, o analista tem, de uma forma geral e na ótica das aplicações práticas, pouca intervenção, dado que os pacotes informáticos fornecem sem grandes exigências as estimativas dos parâmetros.

- Avaliação da qualidade do modelo ajustado

Uma vez formulado o modelo e estimados os respetivos parâmetros passamos a última etapa que consiste em avaliar a qualidade do ajustamento do modelo, ou seja, avaliar as duas fases anteriores. Nesta etapa devemos decidir se o modelo formulado está ou não estatisticamente adequado e, se a decisão for negativa, o modelo deverá ser rejeitado, voltando-se novamente à fase de formulação, com o objetivo de selecionar outro modelo.

Após a obtenção das estimativas dos parâmetros do modelo é conveniente avaliarmos a qualidade estatística do modelo encontrado. Devemos analisar a significância estatística dos parâmetros estimados. Tendo presente o princípio da parcimónia, devemos analisar as estimativas obtidas para os parâmetros com o objetivo de eliminar aqueles que sejam desnecessários, isto é, os parâmetros que não possam considerar-se significativamente diferentes de zero. A aplicação deste princípio tem grande importância pois a prática tem demonstrado que, em geral, modelos parcimoniosos produzem melhores previsões.

Assim devemos proceder, para cada um dos parâmetros do modelo identificado, ao ensaio da hipótese de nulidade, que será descrita detalhadamente mais à frente. Com base no estabelecimento de ensaios de hipóteses, estudamos a adequação do modelo estimado aos dados através dos correspondentes resíduos.

Um modelo que não satisfaça aos critérios aplicados deve ser rejeitado e nesse caso as informações recolhidas durante o processo de avaliação podem sugerir indicações que orientem a formulação de um novo modelo e repetimos o ciclo.

O coeficiente de determinação $R^2$ e o coeficiente de determinação ajustado são duas grandezas que permite-nos avaliar o ajustamento do modelo de regressão, como veremos mais à frente.

Para além da análise do $R^2$, a avaliação da qualidade do ajustamento de um modelo estimado deve ser medida através da análise dos correspondentes resíduos. Uma das hipóteses que está subjacente a toda a teoria de regressão, e nomeadamente à qualidade do ajustamento, é que num "bom modelo" é de esperar que os resíduos sejam aleatórios e próximos de zero.

O analista deve sempre verificar a validade das hipóteses subjacentes ao modelo de regressão (mais à frente, referiremos, com maior detalhe, as hipótese subjacentes ao modelo de regressão), designadamente verificar se a hipótese dos erros serem variáveis aleatórias não correlacionadas de média nula e variância constante com distribuição Normal é uma hipótese correta.

### 1.3. O Modelo de regressão linear múltipla

Como referimos nos modelos de regressão admite-se que existe uma variável aleatória Y, chamada variável dependente, que se supõe depender de um conjunto de outras variáveis, ditas variáveis independentes. A forma mais simples de traduzir uma dependência deste tipo é através de um relacionamento linear, isto é, supor que existem constantes $\beta_1, \beta_2, \beta_3, \ldots \beta_k$ tais que:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k$$

Mas uma relação desta forma traduziria uma relação exata, que se sabe não corresponder, em geral, à realidade. Usualmente, o que se espera é que o segundo membro da equação anterior traduza o relacionamento em termos médios, ou seja, que exista uma variável aleatória com média nula, que permita formalizar a verdadeira relação através de:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon, \qquad \text{com } E[\varepsilon]=0$$

O que também pode ser traduzido por

$$E[Y|\underline{X}]= Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$$

onde $\underline{X}=(X_2,...,X_k)$.

Uma equação como a anterior é chamada equação de regressão linear, por ser linear nos parâmetros, e aos parâmetros $\beta_1, \beta_2, \beta_3, ... \beta_k$ dá-se o nome de coeficientes de regressão. A variável $\varepsilon$ é chamada erro ou desvio do modelo. A equação de regressão diz-se simples ou múltipla, consoante se tem uma ou mais do que uma, respetivamente, variáveis independentes, e a designação usa-se também para os correspondentes modelos.
Começaremos por apresentar o modelo de regressão linear múltipla, e posteriormente referiremos o modelo de regressão linear simples que é um caso particular daquele modelo.

### 1.3.1 Formalização do modelo, hipóteses subjacentes

O modelo de regressão linear múltipla postula a existência de uma relação funcional linear entre uma variável dependente, que vamos designar por Y, e k-1 variáveis explicativas, a qual pode ser traduzida pela seguinte expressão:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon$$

onde

$\beta_1$, $\beta_2$ , $\beta_3$ , ... $\beta_k$ são os coeficientes de regressão e $\varepsilon$ é a variável aleatória residual (chamada termo estocástico ou termo de perturbação do modelo) na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento das variáveis independentes.

Se tivermos n observações sobre cada variável, cada observação de ordem i obedece à seguinte relação:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_i \qquad (i=1,2,\ldots,n) \quad (1)$$

O conjunto das n observações conduz ao sistema:

$$\begin{cases} Y_1 = \beta_1 + \beta_2 X_{21} + ... + \beta_k X_{k1} + \varepsilon_1 \\ Y_2 = \beta_1 + \beta_2 X_{22} + ... + \beta_k X_{k2} + \varepsilon_2 \\ \qquad\qquad ... \\ Y_n = \beta_1 + \beta_2 X_{2n} + ... + \beta_k X_{kn} + \varepsilon_n \end{cases}$$

A notação matricial simplifica a maior parte dos resultados obtidos na regressão linear múltipla, podemos, então, representar o sistema em notação matricial:

$$Y = X\beta + \varepsilon \quad (2)$$

$$\underset{(n\times 1)}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix} \quad \underset{(n\times k)}{X} = \begin{bmatrix} 1 & X_{21} & ... & X_{k1} \\ 1 & X_{22} & ... & X_{k2} \\ 1 & ... & ... & ... \\ 1 & X_{2n} & ... & X_{kn} \end{bmatrix} \quad \underset{(K\times 1)}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix} \quad \underset{(N\times 1)}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

onde

Y = vetor coluna de dimensão (n×1) da variável dependente

X = matriz das variáveis explicativas, (n×k) onde n é o número de observações. (Cada coluna da matriz X corresponde às observações relativas a um mesmo parâmetro. A primeira coluna resulta de considerarmos o modelo com um termo independente ($\beta_1$).

$\beta$ = vetor coluna de dimensão (k×1) dos parâmetros a estimar;

$\varepsilon$ = vetor coluna (n×1) dos resíduos

O problema que se põe é o de encontrar estimativas para os parâmetros $\beta_1, \beta_2,...,\beta_k$.

Os estimadores daqueles parâmetros vão representar-se por $\hat{\beta}_1, \hat{\beta}_2 ... \hat{\beta}_k$.

As propriedades dos estimadores vão depender essencialmente do método utilizado para os obter e das características do modelo, isto é, das hipóteses que o modelo deverá verificar.

Vejamos quais as hipóteses que consideramos em relação ao modelo de regressão:

**Hipóteses do modelo de regressão**

**1. $E(\varepsilon_i)=0$ $i=1,2,....,n$**

**2. $Var(\varepsilon_i) = \sigma^2$** $i=1,2,...,n$

Esta hipótese é designada por homoscedasticidade.

**3. $Cov(\varepsilon_i,\varepsilon_j)=0$** $i \neq j,\ i,j = 1, 2,\ldots,n$

Os termos estocásticos não estão correlacionados, isto é a sua covariância é nula.

Ou seja, duas variáveis aleatórias residuais associadas a duas quaisquer observações não estão correlacionadas.

**4. $\varepsilon_i \sim N(0,\sigma^2)$**

Isto é, as variáveis aleatórias residuais têm distribuição Normal com média zero e igual variância $\sigma^2$.

**5. Independência das variáveis explicativas: e**xiste independência das variáveis explicativas. Isto é, pretende-se que não exista multicolinearidade entre as variáveis explicativas (significando que, em termos matriciais, não existe uma relação linear exata entre quaisquer duas colunas da matriz X).

Em termos de notação matricial, as hipóteses enunciadas podem ser sintetizadas:

$$E(\varepsilon) = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} = 0$$

$$Var(\varepsilon)=E[(\varepsilon-E(\varepsilon))(\varepsilon-E(\varepsilon))']$$

$$=E(\varepsilon\varepsilon')$$

$$E(\varepsilon\varepsilon') = \begin{bmatrix} E(\varepsilon_1^2) & cov(\varepsilon_1\varepsilon_2) & \dots & cov(\varepsilon_1\varepsilon_n) \\ cov(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & & cov(\varepsilon_2\varepsilon_n) \\ \dots & \dots & \dots & \dots \\ cov(\varepsilon_n\varepsilon_1) & cov(\varepsilon_n\varepsilon_2) & \dots & E(\varepsilon_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Ao valor esperado $E(\varepsilon\varepsilon')$ designamos matriz de variância-covariância dos resíduos.

Tem-se assim que a variável aleatória residual deve satisfazer as seguintes condições:

$$E(\varepsilon) = 0$$
$$E(\varepsilon\varepsilon') = \sigma^2 \mathbf{I_n}$$
$$\varepsilon_i \cap N(0,\sigma^2)$$

onde $I_n$ é a matriz identidade de ordem n

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Tal implica que $E(Y)=X\beta$.

### 1.3.2 Estimação dos parâmetros

O problema que se põe é o de encontrar estimativas para os parâmetros populacionais $\beta_1$, $\beta_2$,...,$\beta_k$. Contudo, a inacessibilidade da população não permite, em geral, dispor senão de estimativas dos $\beta_j$ baseadas numa amostra de dimensão n.

Pelo facto de, em teoria, se poder dispor de muitas amostras de dimensão n, permite obter várias estimativas para o mesmo parâmetro $\beta_j$. À variável aleatória geradora das estimativas de $\beta_j$ designamos estimador de $\beta_j$ e representamos por $\hat{\beta}_j$.

Para estimar os $\beta_j$ na equação de regressão (1) vamos encontrar estimadores $\hat{\beta}_1, \hat{\beta}_2 ... \hat{\beta}_k$ que conduzam a

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki}$$

À diferença $Y_i - \hat{Y}_i$ dá-se o nome de erro ou resíduo da regressão, sendo dado por:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki})$$

Podemos propor vários métodos de estimação, mas os mais usuais levam a encontrar $\hat{\beta}_j$ que minimizem alguma função do vetor desses resíduos, como sejam:

- a soma dos valores absolutos dos resíduos

- a soma dos quadrados dos resíduos (método dos mínimos quadrados)

Nas hipóteses referidas, o último método assume uma elevada importância pelas propriedades que confere aos respetivos estimadores.

### 1.3.2.1 O estimador dos mínimos quadrados (OLS[1])

O problema que se põe é o de encontrar estimativas para os parâmetros $\beta_1$, $\beta_2$,...,$\beta_k$.
Como referimos, os estimadores daqueles parâmetros vão representar-se por: $\hat{\beta}_1, \hat{\beta}_2 ... \hat{\beta}_k$

Consideremos de novo o modelo de regressão múltipla

$$Y=X\beta+\varepsilon$$

Vejamos como encontrar os estimadores $\hat{\beta}_1, \hat{\beta}_2 ... \hat{\beta}_k$ que minimizem a soma dos quadrados dos resíduos: $S(\beta)=\sum_{i=1}^{n} e_i^2$ .

Tendo em atenção que $\sum_{i=1}^{n} e_i^2 = e'e$ , tem-se

[1] Está-se a referir ao método dos mínimos quadrados "ordinários" - Ordinary Least Squares, em terminologia inglesa

$$e'e = (Y - X\beta)'(Y - X\beta)$$

$$= (Y - X\beta)'(Y - X\beta)$$

$$= (Y' - \beta' X')(Y - X\beta)$$

$$= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta$$

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

(X'Y é um escalar e por isso igual à sua transposta)

Derivando esta última expressão em ordem a $\beta$ e igualando a zero, tem-se

$$-2X'Y + 2X'X\beta = 0$$

donde se obtém

$$(X'X)\beta = X'Y$$

Se X'X for não singular, isto é, se admitir inversa, tem-se, assim, uma solução única:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

O vetor preditor é

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

e o dos resíduos é

$$e = Y - \hat{Y} = Y - X\hat{\beta}$$

### 1.3.2.2 Distribuição de probabilidades dos estimadores. Valor esperado; Matriz de variâncias-covariâncias. Erro-padrão da regressão

A hipótese de que os erros do modelo, $\varepsilon_i$, têm uma distribuição gaussiana dá-nos propriedades distribucionais muito fortes. O vetor dos estimadores de mínimos quadrados pode escrever-se, alternativamente,

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \varepsilon)$$

$$= \beta + (X'X)^{-1}X'\varepsilon$$

Esta relação exprime o vetor $\hat{\beta}$ como uma função linear do vetor ε, o qual possui distribuição Normal, pelo que os estimadores $\hat{\beta}_j$ possuem uma distribuição Normal, cujo valor esperado é:

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(\varepsilon)$$

$$= \beta + \theta_n$$

$$= \beta$$

onde $\theta_n$ é um vetor nulo de ordem n×1.

Notamos, assim, que os estimadores de mínimos quadrados ($\hat{\beta}_j$) são estimadores não enviesados dos verdadeiros e desconhecidos valores $\beta_j$.

A matriz de variância-covariância das estimativas dos parâmetros

$$\Sigma(\hat{\beta}) = \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1\hat{\beta}_n) \\ \text{cov}(\hat{\beta}_2\hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_2\hat{\beta}_n) \\ \dots & \dots & \dots & \dots \\ \text{cov}(\hat{\beta}_n\hat{\beta}_1) & \text{cov}(\hat{\beta}_n\hat{\beta}_2) & \dots & \text{var}(\hat{\beta}_n) \end{bmatrix}$$

$$= E(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'$$

$$= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$$

$$= E\left[((X'X)^{-1}X'\varepsilon)((X'X)^{-1}X'\varepsilon)'\right]$$

$$= (X'X)^{-1}X'\sigma^2 I_n X'(X'X)^{-1}$$

toma a forma:

$$\Sigma(\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad (3)$$

onde $\sigma^2$ é a variância das variáveis aleatórias residuais, que assumimos ser constantes para todas as observações.

Embora o cálculo matricial simplifique consideravelmente a álgebra dos Mínimos Quadrados, quando o modelo envolve duas ou mais variáveis explicativas, o recurso a meios informáticos é, na prática, quase sempre imprescindível.

Para calcularmos os valores da matriz de variância-covariância apresentada em (3) precisamos de conhecer a variância da variável aleatória residual $\sigma^2$ que supomos constante para todas as observações. Como normalmente desconhecemos $\sigma^2$ temos de utilizar uma estimativa.
Um estimador não enviesado e consistente de $\sigma^2$ é:

$$S^2 = \frac{e'e}{n-k} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \ldots - \hat{\beta}_k X_{ki})^2}{n-k}$$

onde k=número de parâmetros a estimar

À raiz quadrada da variância dos resíduos **(S)** chamamos **erro padrão da regressão**:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki})^2}{n-k}}$$

### 1.3.2.3 Propriedades dos estimadores

Um "bom estimador" deve ser não enviesado, consistente, eficiente e gerar soma dos quadrados dos erros mínima. É possível provar que, se as hipóteses que atrás referimos a respeito do modelo forem cumpridas, o método dos mínimos quadrados no modelo de regressão linear produz estimadores com todas as propriedades estatísticas desejáveis (essas propriedades incluem a linearidade, a centricidade, a eficiência e a consistência).
Com efeito, os estimadores dos mínimos quadrados são estimadores lineares, decorrendo de:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$= (X'X)^{-1}X'(X\beta + \varepsilon)$$

$$= \beta + (X'X)^{-1}X'\varepsilon$$

Os estimadores dos mínimos quadrados são também estimadores centrados, ou não enviesados (um estimador é não enviesado se $E(\hat{\beta}) - \beta = 0$ ou seja, se o valor de cada parâmetro coincidir com o valor médio da sua estimativa). Como referimos anteriormente, os estimadores dos mínimos quadrados são estimadores centrados, esta propriedade decorre de:

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(\varepsilon)$$

$$= \beta + \theta_n$$

$$= \beta$$

onde $\theta_n$ é um vetor nulo de ordem $n \times 1$.

Um estimador não enviesado não fornece, porém, informação acerca da sua dispersão em torno do verdadeiro valor do parâmetro, daí que o estimador deva também ser eficiente[2]. Refira-se que quanto maior a eficiência do estimador maior rigor se obtém quando aplicamos testes às estimativas dos parâmetros. Um estimador deve ainda minimizar o valor médio dos quadrados dos erros, no sentido de maximizar a precisão do ajustamento para permitir melhores previsões. Em virtude do teorema de Gauss-Markov os estimadores de mínimos quadrados são estimadores eficientes, isto é, possuem variância mínima dentro da classe dos estimadores lineares e centrados.

É ainda desejável que as estimativas dos parâmetros se aproximem dos seus valores reais à medida que a dimensão da amostra aumenta, o que nos conduz à propriedade da consistência[3].
Os estimadores dos mínimos quadrados são estimadores consistentes, isto é, à medida que a dimensão da amostra tende para infinito, os estimadores tendem, respetivamente, para os verdadeiros valores dos parâmetros, pois:

$$\operatorname{plim}\hat{\beta} = \operatorname{plim}\left[(X'X)^{-1}X'Y\right]$$

$$= \operatorname{plim}\left[\beta + (X'X)^{-1}X'\varepsilon\right]$$

$$= \beta + \operatorname{plim}(X'X)^{-1}X'\varepsilon$$

$$= \beta + \theta_n \qquad (\theta_n \text{ vetor nulo})$$

$$= \beta$$

[2] Um estimador não enviesado diz-se eficiente se tiver variância mínima de entre os estimadores não enviesados, para a mesma amostra considerada.

[3] Um estimador $\hat{\beta}_j$ é consistente se a probabilidade de $|\hat{\beta}_j - \beta_j|$ ser um infinitésimo à medida que a dimensão da amostra se aproxima do infinito, for igual a um.

O **estimador dos mínimos quadrados** é o melhor estimador linear não enviesado, designamos este estimador por **estimador BLUE** (Best Linear Unbiased Estimator) [24], isto é, é o estimador não enviesado que possui variância mínima (este resultado é normalmente apresentado através do Teorema de Gauss-Markov).

### 1.3.2.4 Inferências sobre os parâmetros: testes de hipóteses e intervalos de confiança

Um dos primeiros elementos sobre os quais nos devemos interrogar após termos efetuado uma regressão é a pertinência dos coeficientes estimados. Os coeficientes estimados $\hat{\beta}_j$ através do método dos mínimos quadrados diferem dos verdadeiros valores de $\beta_j$ e variam de amostra para amostra. Assim, para sabermos se é possível inferir para a população os resultados obtidos, precisamos de conhecer o erro padrão de cada parâmetro $s_{\hat{\beta}_j}$. Estes valores são a raiz quadrada dos elementos que aparecem na diagonal da matriz de variância-covariância dos parâmetros estimada, apresentada em (3).

Enquanto que testes de hipótese sobre distribuições normais são possíveis quando $\sigma^2$ é conhecido, se não conhecermos a variância dos desvios do modelo usamos o estimador $S^2$ pelo que teremos de usar o teste t. O teste t de significância individual testa a hipótese *$H_0$: "o coeficiente é nulo"* contra a hipótese alternativa *$H_1$: "o coeficiente é significativamente diferente de zero"*, ou seja:

$$H_0: \beta_j=0$$
$$H_1: \beta_j\neq0$$

Para tal, calculamos a estatística t dada por:

$$t = \frac{\hat{\beta}_j - 0}{s_{\hat{\beta}_j}} \sim t_{n-k}$$

a qual, sob $H_0$, segue uma distribuição t de Student com (n-k) graus de liberdade.

Este teste permite-nos, assim, analisar se a influência da variável explicativa sobre a variável dependente é significativa. Se a estatística t for, em módulo, maior que o valor crítico $t_c$, rejeitamos a hipótese nula ($H_0$). Para grandes amostras e para um nível de significância de 5% tem-se $t_c$=1,96, pelo que podemos utilizar como regra indicadora que se $|t| > 2$ podemos rejeitar a hipótese nula ($H_0$).

Mais geralmente, podemos testar a hipótese nula de que o coeficiente seja igual a um determinado valor $\beta_j^*$, ou seja:

$$\begin{cases} H_0 : \beta_j = \beta_j^* \\ H_1 : \beta_j \neq \beta_j^* \end{cases} \quad (a)$$

A hipótese alternativa, $H_1$, pode ainda ser formulada do seguinte modo:

$$\begin{cases} H_0 : \beta_j = \beta_j^* \\ H_1 : \beta_j < \beta_j^* \end{cases} (b) \quad \text{ou ainda} \quad \begin{cases} H_0 : \beta_j = \beta_j^* \\ H_1 : \beta_j > \beta_j^* \end{cases} (c)$$

Para tal calculamos:

$$t = \frac{\hat{\beta}_j - \beta_j^*}{s_{\hat{\beta}_j}} \sim t_{n-k}$$

Para um nível de significância $\alpha$, rejeitamos $H_0$ se o valor observado de t pertence à região crítica (R.C.) dada, respetivamente, para os testes (a), (b) e (c) por:

R.C.=$\{t: |t| > t_{(n-k;\, \alpha/2)}\}$ se considerarmos o teste bilateral, isto é a alternativa (a),

R.C.=$\{t < t_{(n-k;\, \alpha)}\}$, se considerarmos a alternativa (b)

R.C.=$\{ t > _{(n-k;\ 1-\alpha)}\}$, se considerarmos a alternativa (c)

A partir da estatística t podemos construir um intervalo de confiança para $\beta_j$ a (1-$\alpha$) por cento de confiança. A um nível de significância de $\alpha$ por cento, tem-se que:

$$\text{Prob}(-t_c < t_{n-k} < t_c) = 1-\alpha$$

substituindo $t_{n-k}$, vem:

$$\text{Prob}(-t_c < \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} < t_c) = 1-\alpha$$

$$\text{Prob}(\hat{\beta}_j - t_c s_{\hat{\beta}} < \beta_j < \hat{\beta}_j + t_c s_{\hat{\beta}})$$

onde $t_c$ é tal que

$$\text{Prob}(-t_c < t_{n-k} < t_c) = 1-\alpha$$

o que nos permite obter um intervalo de confiança para o parâmetro $\beta_j$ do modelo a um nível de significância de $\alpha$%:

$$\left[\hat{\beta}_j - s_{\hat{\beta}_j} t_c ; \hat{\beta}_j + s_{\hat{\beta}_j} t_c\right]$$

### 1.3.3 Avaliação da qualidade da estimação

#### 1.3.3.1 O coeficiente de determinação $R^2$ e o teste F

Após a obtenção das estimativas dos $\beta_j$ e $\sigma^2$ é conveniente avaliar a qualidade do modelo encontrado. Como referimos, uma das hipóteses subjacentes a toda a teoria de regressão é que um bom modelo deverá ter resíduo pequeno. Então, um dos instrumentos que permitirá avaliar essa qualidade será através da soma dos quadrados dos resíduos, pois o princípio dos mínimos quadrados está, em geral, bem presente nos

fundamentos da teoria. Os resíduos da equação de regressão são, assim, muito úteis para testar se o modelo estimado se ajusta bem aos dados utilizados. Um bom ajustamento origina resíduos pequenos e explica uma proporção elevada da variação da variável dependente Y.

Comecemos por definir a variação de Y em torno da média como

$$\text{Variação (Y)} = \sum (Y_i - \overline{Y})^2$$

A variação de Y em torno da sua média pode ser decomposta em duas partes:

- a variação explicada pelo ajustamento, isto é, aquela que é devida à influência linear das variáveis independentes;

- a variação que o ajustamento não consegue explicar, isto é, a variação que é devida aos resíduos ou erros

Notemos então que:

$$\underbrace{Y_i - \overline{Y}}_{\substack{\text{Devio} \\ \text{total}}} = \underbrace{(\hat{Y}_i - \overline{Y})}_{\substack{\text{Desvio} \\ \text{da Regressão}}} + \underbrace{(Y_i - \hat{Y}_i)}_{\substack{\text{Desvio} \\ \text{Residual}}}$$

Elevando ao quadrado ambos os membros e somando para todas as observações:

$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{\substack{\text{TSS} \\ \text{Variação Total}}} = \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}_{\substack{\text{RSS} \\ \text{Variação Explicada} \\ \text{pela Regressão}}} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{\substack{\text{ESS} \\ \text{Variação Residual}}} \qquad (4)$$

ou seja, mais simplificadamente:

$$\text{TSS} = \text{ESS} + \text{RSS}$$

onde:

TSS = soma dos quadrados totais

ESS = soma dos quadrados dos resíduos

RSS = soma dos quadrados da regressão

Na relação (4) TSS possui n-1 fontes independentes (graus de liberdade), ESS possui n-k graus de liberdade e RSS possui k-1 graus de liberdade.

O coeficiente de determinação $R^2$ dá-nos a proporção da variável dependente que é explicada em termos lineares pelas variáveis explicativas, isto é pela regressão, pelo que,

$$R^2 = \frac{RSS}{TSS} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = 1 - \frac{\sum e_i^{\,2}}{\sum(Y_i - \overline{Y})^2} \qquad 0 \le R^2 \le 1$$

À raiz quadrada positiva de $R^2$ dá-se o nome de coeficiente de correlação múltipla R.

Para testarmos a significância do coeficiente de determinação $R^2$ utilizamos a estatística F com k-1 e n-k graus de liberdade, a qual nos permite testar a hipótese $H_0$ de que nenhuma das variáveis explicativas contribui para a explicação da variação em torno da média da variável dependente, ou seja, que não existe relação linear entre a variável dependente e o conjunto das variáveis independentes utilizadas. Isto é, o teste F de significância global testa:

$H_0$: $\beta_1 = \beta_2 = ... = \beta_k$

$H_1$: pelo menos um $\beta_j \neq 0$ (j=1,2,....,k).

A realização deste teste é frequentemente feita com base no quadro da Análise da Variância (ANOVA).

**Quadro ANOVA**

| Fontes de Variação | Soma dos Quadrados | Graus de liberdade | Média Quadrática |
|---|---|---|---|
| Regressão ($X_2,...,X_k$) | $RSS = \sum(\hat{Y}_i - \overline{Y})^2$ | k-1 | RSS/k-1 |
| Resíduos ($e_i$) | $ESS = \sum(Y_i - \hat{Y}_i)^2$ | n-k | ESS/n-k |
| Total | $TSS = \sum(Y_i - \overline{Y})^2$ | n-1 | |

A estatística do teste é:

$$\frac{RSS/k\ -1}{ESS/n\ -k} \sim F_{\ k\text{-}1,n\text{-}k}$$

Notemos que:

$$F = \frac{RSS/k\ -1}{ESS/n\ -k} = \frac{n-k}{k-1} \times \frac{RSS}{ESS} = \frac{n-k}{k-1} \times \frac{R^2}{1-R^2}$$

Se a hipótese nula do teste F de significância global é verdadeira, é de esperar que $R^2$ seja próximo de zero pelo que o valor de F também deve ser pequeno. Assim, grandes valores de F levam-nos a pensar que é de não aceitação da hipótese nula. A aceitação da hipótese nula significa que o conjunto das variáveis explicativas utilizadas no modelo contribui pouco para a explicação da variação da variável dependente, e portanto devemos procurar um conjunto alternativo de variáveis explicativas para Y.
A aceitação da hipótese alternativa no teste F sugere que, pelo menos, uma das variáveis explicativas é estatisticamente significante. Para decidir qual ou quais variáveis, devemos utilizar testes t de significância individual.

### 1.3.3.2 O coeficiente de determinação ajustado $\overline{R}^2$

Devemos notar que uma análise baseada no valor do $R^2$ deve ser cautelosa, pois é quase sempre possível aumentar $R^2$ se adicionarmos novas variáveis ao modelo, o que pode não traduzir necessariamente num modelo melhor que o anterior.

Com efeito, uma das deficiências do coeficiente de determinação enquanto medida da qualidade da estimação é a de que o seu valor aumenta sempre que acrescentam variáveis explicativas ao modelo. A utilização de $R^2$ como uma medida de boa adequabilidade do modelo utiliza variação explicada ou não em Y, não entrando em linha de conta com o número de graus de liberdade do problema. O uso de variâncias em vez de variações elimina a dependência do bom ou mau ajustamento em relação ao número de variáveis independentes do modelo. O coeficiente de determinação ajustado, que procura corrigir o valor de $R^2$ com vista a refletir melhor a qualidade do ajustamento do modelo, é dado por:

$$\overline{R}^2 = 1 - \frac{Var(e)}{Var(Y)}$$

ou seja:

$$\overline{R}^2 = 1 - \frac{s^2}{s_Y^2}$$

como a variância residual é dada por

$$Var(e) = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k} \qquad e \qquad Var(Y) = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}$$

resulta que

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

Como referimos, quando novas variáveis são acrescentadas a um modelo de regressão, $R^2$ aumenta sempre enquanto que o coeficiente de determinação ajustado $\overline{R}^2$ pode aumentar ou não. Se este último não aumentar, isso significa que a variável incluída no modelo não trouxe informação adicional para a explicação da variação da variável dependente pelo que a devemos eliminar.

Muitos analistas consideram que o modelo de regressão que possua o valor máximo de $\overline{R}^2$ será um modelo razoavelmente bom.

### 1.3.3.3 Os coeficientes Beta

Os coeficientes de regressão estimados dependem das unidades de medida das variáveis que lhes estão associadas, pelo que é incorreto interpretar os parâmetros do modelo como indicadores de importância relativa das variáveis independentes. Assim, só se as variáveis independentes estiverem expressas nas mesmas unidades é que é possível comparar os seus coeficientes.

Podemos ultrapassar esta limitação calculando os coeficientes Beta ($\beta_j^*$), ou seja, os coeficientes do modelo de regressão utilizando variáveis normalizadas.

Se

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_i$$

$$\frac{Y_i - Y}{S_Y} = \beta_2^* \frac{X_{2i} - X_2}{S_2} + ... + \beta_k^* \frac{X_{ki} - X_k}{S_k} + \varepsilon_i$$

pelo que

$$\beta_j^* = \beta_j \frac{S_j}{S_Y} \qquad (j=2,\ldots k)$$

onde $S_j$ é o desvio padrão da variável $X_j$.

Quando k=2, $\beta_2^*$ é igual ao coeficiente de correlação entre a variável dependente Y e a variável explicativa $X_2$. Quando k>2 β* não pode ser interpretado como simples coeficiente de correlação. Um coeficiente Beta de 0,65 é interpretado como sendo esta a variação provocada no desvio padrão da variável dependente por uma variação de uma unidade no desvio padrão da variável independente. Notemos, contudo, que os coeficientes Beta são também afetados por existir uma certa correlação entre as variáveis explicativas pelo que, em estrito senso, não representam corretamente a importância das variáveis independentes.

### 1.3.3.4 Coeficientes de correlação parcial e part

Uma outra maneira de medir a importância relativa das variáveis independentes é considerar o aumento em $R^2$ quando uma variável é incluída numa equação que já contenha as outras variáveis independentes. Um grande aumento no $R^2$ significa que essa variável contém informação acerca da variável dependente que não é explicada pelas restantes variáveis utilizadas no modelo. À raiz quadrada desse aumento provocado no $R^2$ designamos **coeficiente de correlação part**.

O **coeficiente de correlação parcial** entre uma variável independente e a variável dependente mede o efeito que essa variável independente tem na variável dependente e que não é explicado pelas outras variáveis independentes do modelo.
Se considerarmos o modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

O coeficiente de correlação parcial entre Y e $X_2$ é dado por:

$$r_{YX_2X_3} = \frac{r_{YX_2} - r_{YX_3} r_{X_2X_3}}{\sqrt{1 - r^2_{X_2X_3}} \sqrt{1 - r^2_{YX_3}}}$$

e podemos interpretar como a correlação entre $X_2$ e Y após remoção do efeito em Y provocado pela variável $X_3$. O quadrado do coeficiente de correlação parcial é dado por

$$r^2_{YX_2X_3} = \frac{R^2 - r^2_{YX_3}}{1 - r^2_{YX_3}}$$

O numerador desta última expressão indica o aumento em $R^2$ quando uma variável é incluída no modelo, neste caso, quando $X_2$ é incluída no modelo, o qual já contém a variável $X_3$. Este numerador é o quadrado do coeficiente de correlação part. O denominador da expressão mede a proporção da variação não explicada por todas as restantes variáveis, neste caso, pela variável $X_3$. Uma vez que o denominador da

expressão é sempre menor ou igual a 1, o coeficiente de correlação part nunca é maior, em valor absoluto, que o coeficiente de correlação parcial.

Se todas as variáveis independentes não estiverem correlacionadas, a variação em $R^2$, quando uma variável é incluída no modelo, não é mais do que o quadrado do coeficiente de correlação entre essa variável e a variável dependente.

O coeficiente de correlação parcial é utilizado no procedimento de **regressão Stepwise** (literalmente, "regressão por passos"), onde acrescentamos variáveis a um modelo de modo a maximizar o valor de $R^2$, ou seja, a minimizar a soma dos quadrados dos erros (ESS). Os coeficientes de correlação parcial entre cada uma das variáveis explicativas e a variável dependente tornam-se úteis para escolher qual a variável explicativa a incluir no modelo, elegendo como primeira variável a incluir no modelo aquela que possui maior coeficiente de correlação com a variável dependente. Em seguida devemos analisar os coeficientes de correlação parcial entre a variável dependente e cada uma das variáveis independentes não incluídas na equação, devendo a próxima variável a entrar no modelo aquela que possui maior coeficiente de correlação parcial. Contudo, esta variável só deve entrar no modelo se for significativa, isto é, se a estatística ***t*** correspondente for superior ao valor crítico. O processo continua até se chegar à situação em que nenhuma variável deva ser acrescentada à equação porque o valor F para entrar é inferior ao valor crítico da tabela e nenhuma deva ser eliminada da equação.

### 1.3.3. 5 Análise dos resíduos

Para além da análise do $R^2$, a avaliação da qualidade do ajustamento de um modelo estimado deve ser medida através da análise dos correspondentes resíduos.

Uma das hipóteses que está subjacente a toda a teoria de regressão, e nomeadamente à qualidade do ajustamento, é que num "bom modelo" é de esperar que os resíduos sejam aleatórios e próximos de zero.

O analista deve sempre verificar a validade das hipóteses subjacentes ao modelo de regressão, designadamente verificar se a hipótese de que os erros são variáveis aleatórias não correlacionadas de média nula e variância constante com distribuição Normal é uma hipótese correta.

### 1.3.4 Estimação: situações particulares

#### 1.3.4.1 Multicolinearidade

A hipótese de "ausência de multicolinearidade entre as variáveis explicativas", subjacente a uma boa utilização do modelo de regressão linear, requer que não exista relação linear entre as variáveis independentes do modelo. Com efeito, variáveis explicativas correlacionadas originam estimativas dos parâmetros do modelo e somas dos quadrados atribuídos a cada variável, dependentes de outras variáveis presentes na equação, o que faz com que as variâncias dos estimadores resultem maiores que as reais, originando coeficientes de determinação significativos sem que os coeficientes do modelo o sejam. Se existir perfeita correlação entre as variáveis independentes, a matriz de correlação é singular (isto é, não admite inversa) não existindo solução OLS para o problema da estimação dos parâmetros. Na prática não são muito usuais situações de perfeita singularidade, mas podemos estar frequentemente em presença de variáveis independentes com um elevado grau de multicolinearidade (situações em que as variáveis independentes sejam aproximadamente combinações lineares de outras variáveis independentes do modelo). Elevados coeficientes na matriz de correlação entre as variáveis independentes indica-nos que estamos em presença de multicolinearidade, podendo ser apenas um problema de amostra em que a inclusão de mais observações leva ao desaparecimento do problema.

O uso de determinados métodos e/ou técnicas permitem-nos resolver o problema da multicolinearidade, como sejam: i) o uso de componentes principais (é uma técnica frequentemente utilizada quando nos encontramos em presença de multicolinearidade e não podemos aumentar o número de variáveis explicativas); ii) a regressão Stepwise (é

um processo expedito para resolver o problema da multicolinearidade por eliminação de algumas variáveis explicativas da equação).

#### 1.3.4.2 Variáveis *Dummy*

As variáveis utilizadas no modelo de regressão múltipla são normalmente variáveis quantitativas. Porém, podemos ter interesse em incluir variáveis qualitativas como variáveis explicativas do modelo. Neste caso, criamos uma variável muda (dummy variable) que toma apenas dois valores (0 e 1) e que é utilizada para quantificar efeitos de ordem qualitativa sobre a variável dependente.
As variáveis *dummy* tornam-se, assim, particularmente úteis quando lidamos com variáveis nominais ou ordinais.

#### 1.3.4.3 Funções não lineares mas linearizáveis

A análise de regressão assume que as relações entre as variáveis são lineares, ou seja que a equação de regressão possa ser traduzida por uma linha reta. Contudo há situações em que os dados não correspondem a este tipo de formulação ( por exemplo potências, exponenciais,...). Para que os parâmetros de uma função não-linear sejam estimáveis pelo método dos mínimos quadrados, a função deve ser linearizável. Vejamos dois exemplos de funções não lineares que mediante transformações se tornam em funções lineares:

##### 1.3.4.3.1 Função potência

Trata-se de uma função do tipo: $Y = \alpha X^{\beta}$

Considerando então

$$Y_i = AX_{2i}^{\beta_2} X_{3i}^{\beta_3} \ldots X_{ki}^{\beta_k} . e^{\varepsilon_i}$$

(onde $\varepsilon_i$ possui as propriedades desejáveis e ***e*** é o número de Neper, e=2.71...)

A linearização da função potência faz-se por logaritimização,

$$\ln Y_i = \ln A + \beta_2 \ln X_{2i} + \ldots + \beta_k \ln X_{ki} + \varepsilon_i$$

e operando em seguida a transformação de variáveis:

$$Y_i^* = \ln Y_i$$

$$X_{ji}^* = \ln X_{ji} \qquad (j=2,\ldots,k)$$

o modelo pode então escrever-se na forma linear

$$Y_i^* = \beta_1 + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + \ldots + \beta_k X_{ki}^* + \varepsilon_i$$

onde $\beta_1=\ln A$

### 1.3.4.3.2 Função exponencial

A função exponencial é do tipo:

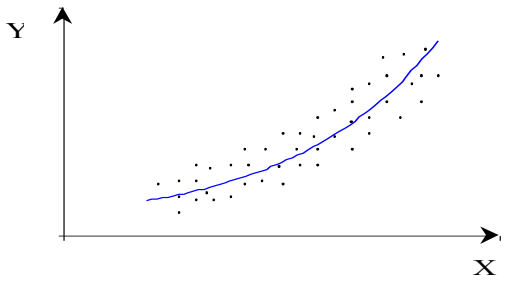



Figura 1 - Gráfico de função exponencial

Considerando então,

$$Y_i = e^{(\beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i)}$$

para linearizar esta função aplica-se logaritmos a ambos os membros:

$$\ln Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$$

e fazendo

$$Y_i^* = \ln Y_i$$

obtemos a versão linear

$$Y_i^* = \beta_1 + \beta_2 X_{2i}^* + ... + \beta_k X_{ki}^* + \varepsilon_i$$

### 1.3.5 Heteroscedasticidade e autocorrelação dos resíduos

Como referimos, a aplicação do modelo de regressão linear múltipla pressupõe como hipótese que os resíduos sejam independentes e tenham distribuição normal com média nula e variância constante.

Violar a hipótese de normalidade será muito grave pois nessas situações o estimador dos mínimos quadrados gera estimativas dos parâmetros consistentes e não enviesadas. Com o teorema do limite central podemos ter a garantia de que as estatísticas t e F podem ser utilizadas se a amostra for razoavelmente grande. A violação de que os resíduos têm valor esperado nulo não trará grande problema porque as estimativas dos coeficientes do modelo permanecerão inalteradas, apenas a constante absorve o efeito do enviesamento.

Quando os resíduos não são independentes, estamos em presença de autocorrelação e quando a variância dos resíduos não é constante, estamos na presença de heteroscedasticidade.

A violação das hipóteses "independência dos resíduos" e "variância dos resíduos constante" coloca reservas quanto à validade do modelo OLS estimado.

### 1.3.5.1 Heteroscedasticidade dos resíduos

Os resíduos são heteroscedásticos quando a variância não é constante, ou seja, quando:

$$\sigma^2_{\varepsilon i} \neq \sigma^2_{\varepsilon j} \qquad \text{com } i \neq j$$

e neste caso o estimador OLS é não enviesado mas não é eficiente.

Na prática, o gráfico dos resíduos permite, em certos casos, detetar a presença de heteroscedasticidade (resíduos crescentes em valor absoluto), existindo testes adequados para testar a hipótese nula de heteroscedasticidade, ou seja:

$$\sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_n$$

Um teste que pode ser aplicado é o Teste de Goldfeld e Quandt[4].

### 1.3.5.2 Autocorrelação dos resíduos

Uma das hipóteses formuladas aquando da aplicação dos mínimos quadrados OLS é a ausência de correlação entre as variáveis aleatórias residuais, quer quando dispomos de séries temporais, quer quando usamos dados seccionais. Muitas das vezes quando dispomos, por exemplo de séries temporais, essa correlação existe, por exemplo poderá

[4] A aplicação do teste de Goldfeld e Quandt resume-se às seguintes etapas:
1ª - Ordenam-se os dados em função da variável explicativa que supomos ser a que influencia o crescimento da variância dos resíduos. No caso de séries temporais, se pensarmos que as variâncias dos resíduos crescem com o tempo ou são função de uma variável explicativa, ela mesma crescente, é inútil alterar a ordenação dos dados da amostra.
2ª - Eliminamos d observações por forma a tornar mais explicito o fenómeno da heteroscedasticidade. O número de observações eliminadas deve ser escolhido por forma a permitir que a regressão se possa processar sobre as duas sub-amostras resultantes. As d observações devem ser retiradas do meio e o seu número não deve ser superior a 1/5 do total das observações da amostra (n).
3ª- Efetuam-se separadamente regressões sobre as duas sub-amostras. Cada regressão envolve um conjunto de (n-d)/2 observações. O número de observações eliminadas (d) deve ser suficientemente pequeno para permitir suficientes graus de liberdade para cada sub-amostra.
4ª - Calcula-se a soma do quadrado dos resíduos (ESS) associados a cada regressão.
5ª - Assumindo que os resíduos têm distribuição Normal e não são correlacionados, $ESS_2/ESS_1$ terá uma distribuição F com (n-d-2k)/2 graus de liberdade tanto para o numerador como para o denominador (onde k é o número de variáveis independentes, incluindo a constante). Se $ESS_2/ESS_1$ for maior que $F_{crítico}$ então concluímos que há heteroscedasticidade. Detetada a existência de heteroscedasticidade dos resíduos, devemos ponderar se a formulação do modelo está correta ou não. Por vezes a logaritmização das variáveis corrige a heteroscedasticidade evitando-se a recorrência a um método de estimação mais complicado. Quando necessário, e se apenas a heteroscedasticidade estiver presente, um caso particular do método dos mínimos quadradados generalizados (GLS), o método dos mínimos quadrados ponderado (WLS) é o aconselhado.

existir autocorrelação de primeira ordem, ou seja os erros num dado período estão correlacionados com os erros no período imediatamente anterior:

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t$$

(usamos o índice t em vez de i, uma vez que este tipo de situação é mais frequente quando trabalhamos com séries temporais)

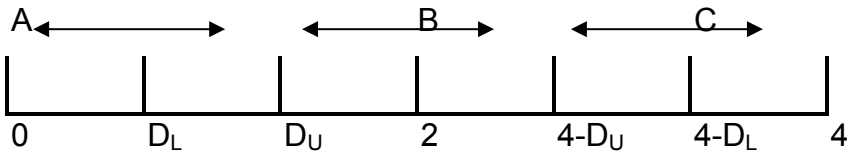
Assim, quando existe correlação entre as variáveis aleatórias residuais, o estimador dos mínimos quadrados é não enviesado mas não é eficiente. A autocorrelação pode ser positiva ou negativa, consoante o valor de ρ. Na presença de autocorrelação positiva, a qual ocorre frequentemente em estudos de séries temporais, a perda de eficiência resulta do facto de que as estimativas dos erros padrão obtidas pelo método OLS são inferiores aos valores reais desses erros, o que conduz a que as estimativas dos parâmetros β pareçam mais precisas do que na realidade o são. Para testar a presença de autocorrelação podemos usar o teste de Durbin-Watson, que envolve o cálculo da estatística D sobre os resíduos depois de aplicado o método dos mínimos quadrados OLS.

$$D = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})}{\sum_{t=1}^{n}(\varepsilon_t)^2}$$

É fácil verificar que:

$$-1 \le \rho \le 1 \rightarrow 0 \le D \le 4$$

Quando D está próximo de zero estamos em presença de autocorrelação positiva, quando D está próximo de 4 estamos na presença de autocorrelação negativa. Se D está próximo de 2 não há autocorrelação. A tabela da estatística DW dá-nos os valores críticos superiores e inferiores.



Se D se encontrar na zona marcada com B, aceitamos $\rho=0$ para um determinado nível de significância $\alpha$, ou seja, não há autocorrelação dos resíduos. Se D se encontrar na zona marcada com A, aceitamos $\rho>0$ ao nível de significância $\alpha$. Se D cai na zona marcada com C aceitamos $\rho<0$ ao nível de significância $\alpha$, existindo neste caso autocorrelação negativa. Fora das regiões o teste é inconclusivo. Notemos que a existência de autocorrelação negativa significa que os resíduos flutuam rapidamente entre positivo e negativo, o que é menos preocupante que a autocorrelação positiva. Assim devemos preocupar-nos mais com $D\cong 0$ do que com $D\cong 4$. Em geral, se $D>D_U$ não necessitamos de nos preocupar com a autocorrelação positiva. Se $D<D_U$ devemos ter a preocupação de fazer algo para removê-la. Refira-se que as tabelas da estatística de Durbin Watson que aparecem publicadas referem-se ao modelo de regressão linear na sua forma geral. Estas tabelas não devem ser utilizadas no caso de um modelo sem constante ou quando haja variáveis desfasadas no tempo como variáveis explicativas. Nestas circunstâncias devem utilizar-se tabelas de Durbin Watson corrigidas.

## 2.Árvores de Regressão

### 2.1.Métodos de segmentação (recursivos em árvore)

Tomando como referência o método de segmentação CART (trabalho de 1984) , há outros métodos que o precederam – técnicas iniciadas nas ciências sociais - tais como o método BELSON (Belson, 1959; Cailliez et al., 1976 ) onde a variável dependente Y é binária à custa de p caracteres explicativos $X_1,..,X_p$. ; o método ÉLISÉE ( Bouroche et al., 1970; Cailliez et al., 1976) com Y qualitativa e $X_i$ também qualitativas; segue-se o AID ( Automatic Interaction Detection) trabalho de Morgan C. e Sonquist J.N. em 1963, com Y quantitativa e $X_i$ qualitativas. THAID ( Theta Automatic Interaction Detection ) é uma expansão do AID devido a Morgan J. A. e Messenger R.C. em 1972 – válido para o caso de Y qualitativa - e CHAID ( Chi-square Automatic Interaction Detection ) proposto por Kass G. V. em 1980 é uma melhoria – fusão – dos métodos AID e THAID, tornando-o num método de classificação e regressão tal como o CART .

Depois de 1984 – ano do método CART – outros métodos apareceram. Há que referir o Exaustive CHAID proposto por Biggs et al. em 1991; neste mesmo ano J. H. Friedman – co-autor do CART – desenvolveu o MARS ( Multivariate Adaptive Regression Splines); o QUEST (Quick Unbiased Efficient Statistical Tree ) desenvolvido por Loh W. e Shih Y. em 1997; ID3 apresentado por J. R. Quinlan em 1986, C4.5 em 1993 e C5.0 em 1997, ainda deste último autor; …Estes e outros são os métodos que integram os principais pacotes estatísticos.

### 2.2. O método CART

O método CART foi desenvolvido nos anos 80 por Breiman, Freidman, Olshen e Stone no seu trabalho "Classification and Regression Trees " (1984) [4]. Consiste na construção de uma árvore, à custa de divisões sucessivas de um conjunto, em dois segmentos (chamados nós) homogéneos, tendo como referência uma variável Y e utilizando a informação de p variáveis $X_1, X_2, \ldots, X_p$.

A árvore resultante é uma árvore invertida tendo na raiz ( no topo ) a amostra total a segmentar quer em segmentos intermédios (nós intermédios) ou em outros terminais (nós terminais). Os nós terminais formam uma partição do conjunto em

classes homogéneas e distintas relativamente à variável Y. A árvore será chamada de árvore de regressão se Y é quantitativa e de classificação se Y é qualitativa.

### 2.2.1 Passo-a-passo de uma árvore CART

A análise CART consiste em quatro passos básicos. O primeiro passo é a *construção da árvore* (máxima), neste, a árvore binária é construída usando a divisão recursiva dos seus nós. A atribuição de uma classe a cada nó ocorre independente do nó ser dividido posteriormente. O segundo passo consiste em *parar de construir a árvore*. Neste momento a árvore está no seu tamanho máximo e, provavelmente, ajusta de modo exagerado as informações contidas no conjunto de aprendizado. O terceiro passo necessário será a *podagem da árvore* que resulta na criação de uma sequência de árvores menores e mais simples, através da remoção de nós com menor relevância. O último passo consiste da *seleção da árvore ideal* do conjunto de sequência de árvores podadas. Cada um destes passos será discutido a seguir.

### 2.2.2. Construção da árvore

A construção de uma árvore começa com o seu nó raiz, que inclui todos os valores. Começando neste nó, o algoritmo que implementa o CART encontra a melhor variável para efetuar a divisão do nó em dois filhos. Para encontrar esta variável, o algoritmo terá que verificar dentre todas as possíveis variáveis de divisão (também conhecidas como variáveis splitter) e dentre todos os possíveis valores de cada uma destas variáveis, qual a melhor para dividir o nó.

Sendo a árvore construída por divisões sucessivas dos dados em dois conjuntos, o número de divisões possíveis depende das modalidades das variáveis explicativas. Assim uma binária B(0,1) origina uma divisão. Uma variável nominal N com k modalidades leva a $2^{k-1}$-1 divisões possíveis. Uma variável ordinal O com k modalidades admite k-1 divisões possíveis – o mesmo acontece com uma variável quantitativa com k valores distintos: k-1 divisões possíveis.

Por exemplo, 3 variáveis: B, N e O - com as respetivas modalidades (b1,b2), (n1,n2,n3) e (o1,o2,o3,o4) – utilizadas para dividir um nó t em dois nós $t_e$ – esquerdo - e $t_d$ - direito - , admitem as 7 divisões seguintes :

| $t_e$ | $t_d$ |
|---|---|
| (b1) | (b2) |
| (o1) | (o1,o2,o3) |
| (o1,o2) | (o3,o4) |
| (o1,o2,o3) | (o4) |
| (n1) | (n2,n3) |
| (n2) | (n1,n3) |
| (n3) | (n1,n2) |

Tabela 1 - Divisões possíveis de uma variável

### 2.2.3. Definições

$X = (x_1,x_2,\ldots,x_p)$ um vetor de variáveis aleatórias composto de $p$ variáveis explicativas.

$Y$ Variável para prever (resposta).

$d(X)$ um preditor de $Y$ dado $X$ definida por $d = E(Y|X)$.

$N(t)$ o número de nós t.

$n$ total.

$T, T'$ árvores .

$\tilde{T}$ o conjunto de nós terminais $T$.

Uma árvore de regressão constrói-se à custa de divisões sucessivas de subconjuntos de amostra de treinamento em dois descendentes. A ideia é selecionar uma divisão de modo que os nós descendentes forneçam uma melhor previsão do que o nó pai. Segundo este princípio, uma grande árvore é construída, que, por razões de generalização, é de seguida podada.

A construção de uma árvore de regressão segue dois objetivos:

1. predizer a variável Y tão eficazmente quanto possível.
2. estudar a relação entre a variável resposta Y e variável X .

A construção é baseada em divisões sucessivas. Mas para alcançar essas divisões, é preciso avaliar, utilizando critérios, o desempenho preditivo de um nó e os de toda a árvore, sabendo que eles são reduzidos para o desempenho de nós terminais.

Para medir o desempenho de um preditor do nível de uma amostra de tamanho n, a medida seguinte é usada:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - d(X_i))^2 \qquad (*)$$

com *d (Xi) = E(Y|Xi)* preditor de Yi pode-se escrever da seguinte forma:

$$E(Y\text{-}d(X))^2$$

Uma vez definida a medida de eficácia do preditor, resta medir a eficácia dos preditores ao nível da árvore. Para este fim, são utilizados os critérios relativos aos nós terminais da árvore.

Seja r(t) **o critério de eficácia local** de um preditor d para um nó terminal t ponderado segundo o efetivo N (t) do nó.

$$r(t) = \frac{1}{N(t)}\sum_{i=1}^{N(t)}(Y_i - d(X_i))^2$$

Seja R(t) **o critério de eficácia global** de um preditor *d* para um nó *t* ponderado segundo o efetivo total.

$$R(t) = p(t)r(t) = \frac{1}{n}\sum_{i=1}^{N(t)}(Y_i - d(X_i))^2$$

Com $p(t) = \frac{N(t)}{n}$ a proporção de individuos de T no nó t.

Seja **R(T) o critério de eficácia de uma árvore T –** valor esperado da soma dos erros quadráticos da regressão. Esta eficácia é medida a partir da eficácia dos preditores presentes nos nós terminais ou folhas da árvore.

$$R(T) = \sum_{t\in\widetilde{T}} p(t)r(t) = \sum_{t\in\widetilde{T}} R(t)$$

Todos esses critérios fazem apelo a um preditor específico para cada nó. Este preditor não é aleatório. Com efeito, de todos os possíveis preditores para um nó, só o preditor $d^*$ definido como a esperança condicional de Y para uma medida *x* dada, é utilizada.

$$d(x) = E(Y \mid X = x)$$

No caso de uma regressão linear, *E (Y / X = x)* escreve-se:

$$E\ (Y / X = x) = X^t\ b$$

com $b = (X^t X)^{-1} X^t Y$ solução dos mínimos quadrados de $\frac{1}{n}\sum_{i=1}^{n}(Y_i - d(X_i))^2$.

$X$ é a matriz das variáveis explicativas

$Y$ o vetor das variáveis resposta.

As figuras que se seguem ilustram a estrutura de uma árvore de regressão:

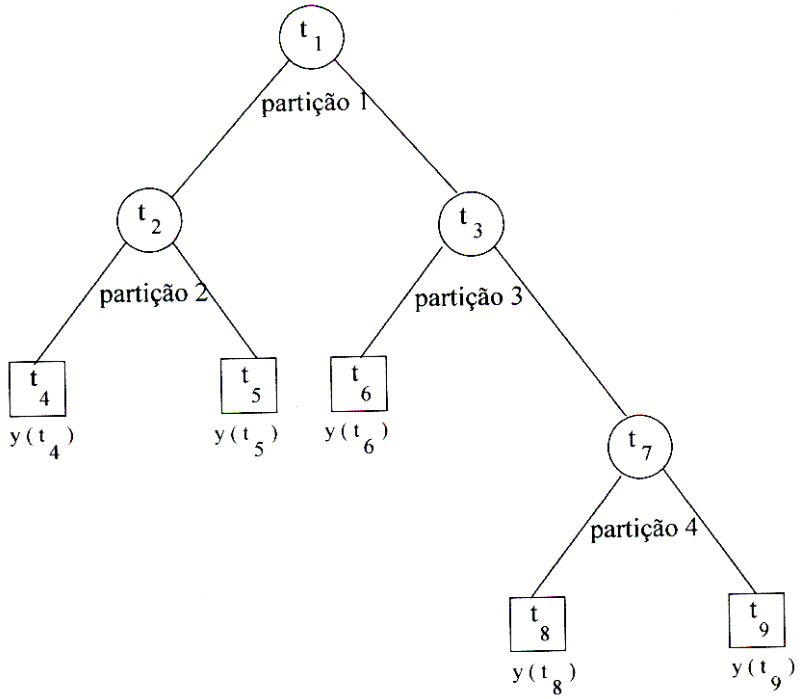



Figura 2 - Estrutura de uma árvore de regressão

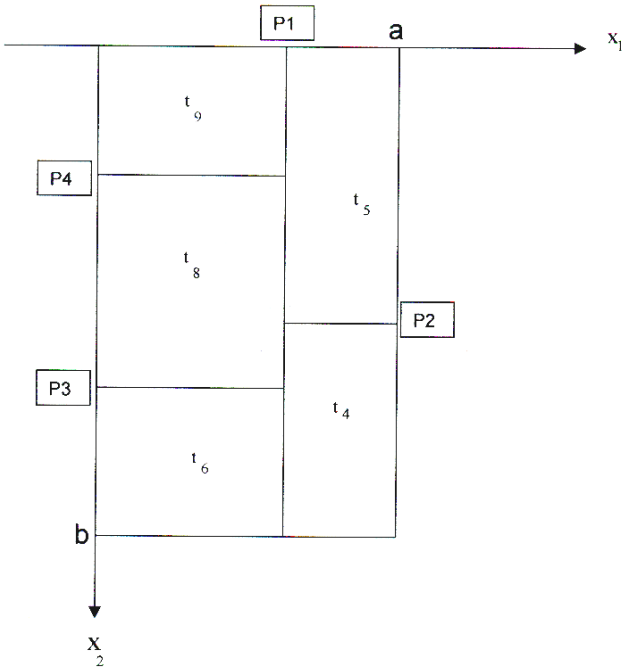



Figura 3 - Regiões determinadas no espaço das variáveis explicativas

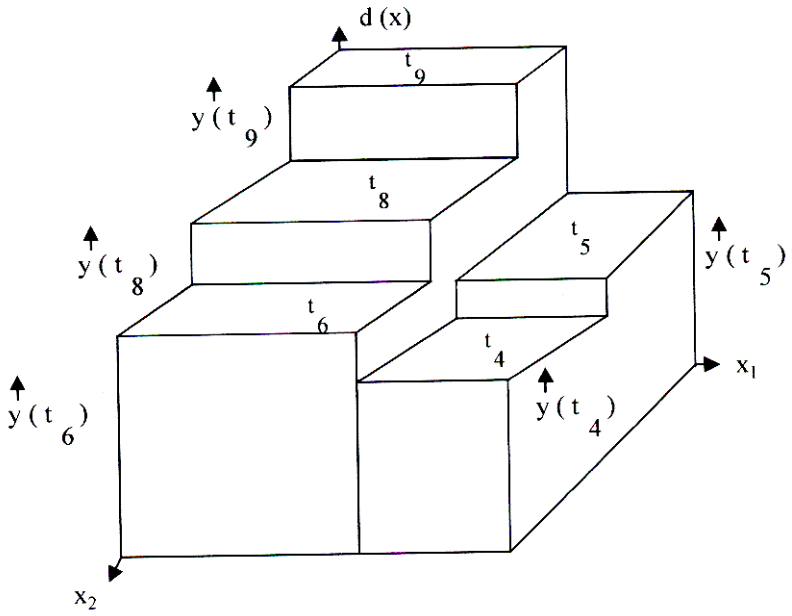



Figura 4 - Gráfico da superfície de resposta

### 2.2.4 Critérios divisão

A árvore de regressão é construída usando divisões supostamente para melhorar a eficiência de predição da árvore. A divisão é uma segmentação do conjunto de aprendizagem através de perguntas binárias sobre as variáveis $X_i$, i = 1, ..., p.

As questões são da forma:

$$\boldsymbol{X_i} \leq c_i^j$$

Com $X_i$ uma variável de corte, e $c_i^j$ um corte tirado de $X_i$ .

São examinadas, uma por uma, todas as variáveis explicativas. O critério da "melhor" divisão de um nó baseia-se na minimização da variância de Y nos segmentos descendentes.

Cada divisão separa a amostra em segmentos descendentes: $t_e$ que contem os dados que verificam $\boldsymbol{X_i} \leq c_i^j$ e $t_d$ os restantes ( $\boldsymbol{X_i} > c_i^j$).

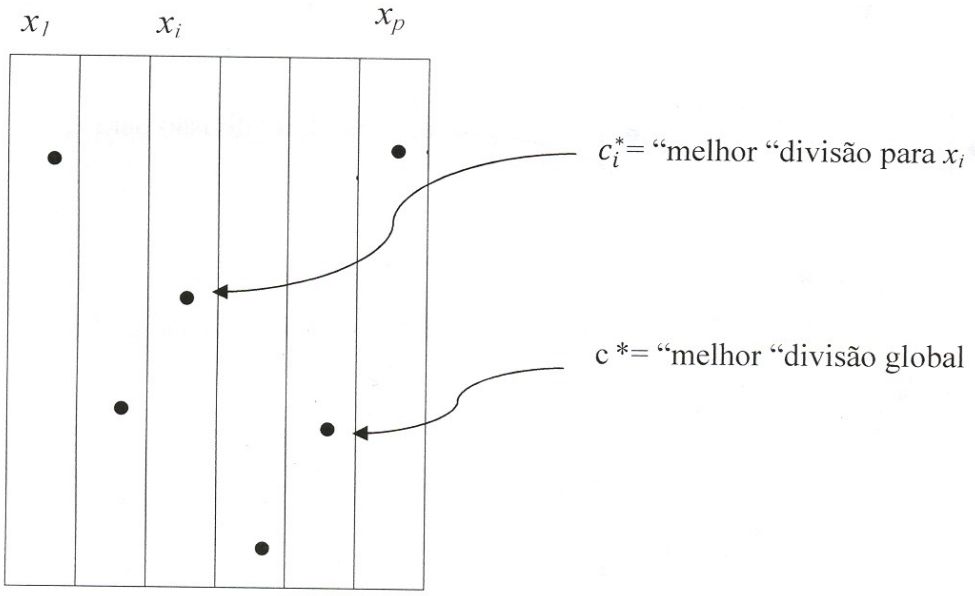



Figura 5 - Melhores divisões para as variáveis

A divisão de um nó t de uma árvore T com um corte *s* conduz a uma nova árvore T'. Esta divisão de um nó terminal t em dois novos nós terminais $t_e$ e $t_d$ tem como objetivo a minimização de um critério, ou seja, reduzir o erro realizado em T' relativamente ao realizado sobre T.

A redução do erro associado a uma divisão T ', escreve-se:

$$\Delta R(s,t) = R(T) - R(T')$$

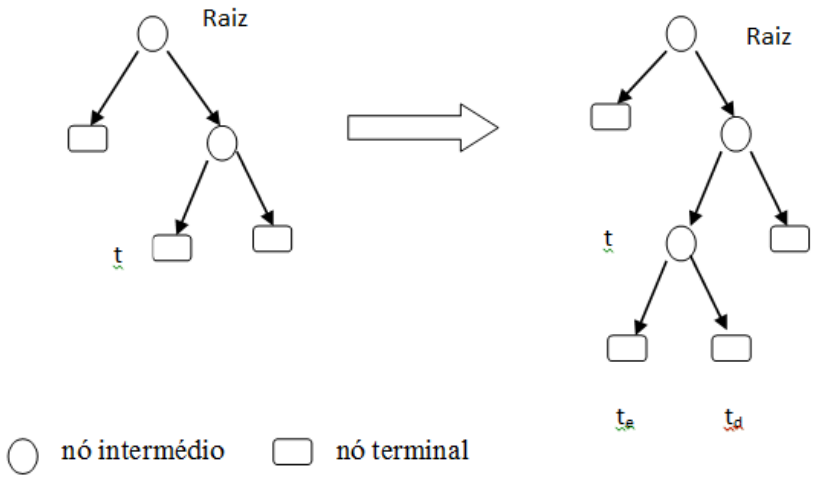



Figura 6 - Desenvolvimento de um nó terminal t em dois nós filhos $t_e$ e $t_d$

De todas as divisões possíveis, só a divisão que maximiza o critério $\Delta R$, é retida.

Seja $s^*$ essa divisão tal que:

$$\Delta R(s^*,t) = \max_{s \in S} \Delta R(s,t)$$

Por causa do critério de aditividade R(T**) -** critério de eficácia de uma árvore T- apenas os nós terminais t, $t_e$, $t_d$ intervêm. O critério ΔR é reescrito:

$$\Delta R(s,t) = R(t) - R(t_e) - R(t_d)$$

Da definição **R(t)** - critério de eficácia global – os critérios de eficácia podem ser decompostos, o que permite a simplificação de ΔR:

$$\Delta R(s,t) = p(t)[r(t) - p_e r(t_e) - p_d r(t_d)] \qquad \text{com} \quad p_e = \frac{p(t_e)}{p(t)} \quad \text{e} \quad p_d = \frac{p(t_d)}{p(t)}$$

### 2.2.5. Poda de custo mínimo

A fase de construção da árvore leva a uma grande árvore, que por razões de generalização deve ser podada. A poda é equivalente a encontrar uma pequena árvore com a máxima eficiência em um ou mais conjuntos diferentes do que a utilizada para a construção.

Poda, consiste em manter a melhor árvore entre todas as árvores de m nós e a melhor árvore entre todas as árvores de m-1 nós, … , até à árvore com um nó, ou seja, a raiz. Entre o conjunto das melhores árvores, fica aquela que obteve uma taxa de erro dos mais baixos num conjunto de teste.

A desvantagem dessa abordagem reside no número de subárvores a serem consideradas para construir a sequência de árvores podadas {To, $T_1$, ...,$T_{H-1}$}. Além disso, a árvore podada $T_{k+1}$ não é, necessariamente, derivada de $T_k$. Isto significa que o nó podado em $T_k$ pode reaparecer numa árvore $T_j$ para j> k.

Para evitar a busca de muitas subárvores, Breiman propôs uma alternativa para uma abordagem de poda recursiva que constrói uma sequência de árvores podadas. Cada árvore podada é obtida a partir do corte de um ramo da subárvore podada antes. A poda leva a construir uma sequência $\{ T_m^*,\ldots, T_2^*, \{t_1^*\}\}$ de acordo com o processo resumidos a seguir.

A cada subárvore $T \leq T_{max}$ é associada uma complexidade $|\tilde{T}|$, representando o número de nós terminais de T. Quanto ao custo de complexidade, é definido um parâmetro $\alpha$,tendo em conta, a altura da árvore e sua eficácia.

O custo de complexidade de uma árvore T é denotada por $R_\alpha$ (T), e é definido por:

$$R_\alpha (T) = R (T) + \alpha |\tilde{T}|$$

De acordo com os valores de α, a árvore podada é de tamanho diferente.

### 2.2.6. Validação do Modelo

Para um modelo a estimativa da qualidade do mesmo pode ser obtida através de *estimadores de amostra independentes* – quando a dimensão dos dados é grande - aplicando o modelo a novos dados, que resultam da divisão aleatória dos dados originais em duas amostras: normalmente a proporção é de 2:1. Uma amostra – dita de treino, a de dimensão superior – para desenvolver o modelo e a outra para testar o modelo.

O outro método, *a validação cruzada* – aplicado quando a dimensão dos dados é pequena e aplicado neste trabalho – é uma alternativa, uma vez que a mesma amostra é usada para construir a árvore e para o cálculo do erro. Usando a validação cruzada, o cálculo do erro, faz-se da seguinte maneira:

- Subdivide-se a amostra inicial em V subamostras, $l_1, l_2, \ldots, l_v$ de dimensões aproximadamente iguais – normalmente V=10 – com as variáveis explicativas consideradas a terem uma distribuição semelhante.
- São construídas V árvores diferentes, utilizando (V-1) /V das observações, sendo as restantes 1/V observações utilizadas para avaliar o erro.

Produzem-se árvores $T^{(v)}(\alpha)$ que são de custo-complexidade ( cp) mínimo para o valor do parâmetro α , aplicando o processo a todos os dados, obtemos as sequencias $\{T_k\}$ e $\{\alpha_k\}$. Defina-se $\propto'_k = \sqrt{\propto_k \propto_{k+1}}$ . Seja $d_k^{(v)}(x)$ o preditor de correspondente à árvore $T^{(v)}(\propto'_k)$ . As estimativas de validação cruzada, $R^{CV}(T_k)$ e $RE^{CV}(T_k)$ são dadas por:

$$R^{CV}(Tk) = \frac{1}{N}\sum_{V=1}^{V}\sum_{(x_n,y_n)\epsilon l_v}(y_n - d_k^{(v)}(x_n))2$$

e

$$RE^{CV}(T_k) = \frac{RCV(Tk)}{R(\bar{y})}$$

( Ver Breiman et all (1983, p. 234)

O $T_k$ selecionado é a menor árvore que satisfaz :

$R^{CV}(T_k) \leq R^{CV}(T_{k0}) + SE\ (R^{CV}(T_{k0}))$ ( *regra 1SE* , Breiman *et al.* (1983, p. 237) )

onde $R^{CV}(T_{k0}) = \min R^{CV}(T_k)$ e SE o desvio padrão.

A proporção de variação explicada resultante de validação cruzada permite-nos obter uma estimativa da capacidade preditiva do modelo proposto:

$$1-\sum_{v=1}^{v} \frac{n_v}{n} \left( \frac{\sum_{i=1}^{n_v} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_v} (y_i - \bar{y}_v)^2} \right)$$

## 3.A Linguagem R

### 3.1 Um pouco da história do R

" O R é ao mesmo tempo uma linguagem de programação e um ambiente para computação estatística, modelação e visualização de dados." (Torgo,2009). O seu desenvolvimento deve-se aos professores Robert Gentleman e Ross Ihaka do Departamento de Estatística da Universidade de Auckland na Nova Zelândia, que durante cinco a seis anos, trabalharam inseparáveis: um digitava e o outro pensava… (www.nytimes.com/2009/01/07/technology/businesscomputing/07program.html), (publicação do jornal New York Times em 7 de janeiro de 2009, da entrevista ao professor R. Gentleman). Conhecidos por "R & R"- origem do nome para o programa R - tinham como objetivo inicial, em 1991, produzir um software para as suas aulas de laboratório, baseado na linguagem S, que é um pacote estatístico utilizado pelo software comercial S-Plus criado por Jonh M. Chambers da AT&T (American Telephone and Telegraph) no final dos anos 80 .

Em 1993 distribuiu-se pela primeira vez o R, sendo as cópias disponibilizadas no StatLib - um sistema de Dados, Software e Notícias da Comunidade Estatística.

Em 1995, por incentivo de Martin Mächler - um dos primeiros utilizadores do programa R- do ETH Zürich (Instituto Federal Suíço de Tecnologia de Zurique) "R & R", lançaram o código fonte do R, disponível por ftp (uma forma de se transferir dados pela internet), sobre os termos de GNU General Public License (Licença Pública Geral), GNU GPL ou simplesmente GPL, é a designação da licença para software livre idealizada por Richard Matthew Stallman em 1989, no âmbito do projeto GNU da Free Software Foundation (FSF) ( www.gnu.org). Devido à sua compatibilidade com vários sistemas operacionais, o R, "corre" em MacOS, em Windows (a partir da versão Windows 95), em UNIX e sistemas similares como Linux e FreeBSD.

Desde 1997, um grupo de profissionais têm acesso ao código fonte do R, possibilitando assim a atualização mais rápida do software – projeto colaborativo com muitos contribuidores. (www.r-project.org/contributors.html).

Devido a vários fatores: totalmente gratuito, ser código aberto ( *open source,* software também conhecido por software livre), possibilidade de desenvolver novos modelos e personalizá-los, de fácil uso,…, esta linguagem – e ambiente – é cada vez mais um software que tem cativado muitos adeptos, destronando softwares "rivais" ou, em alguns casos, estes últimos integram o R no seu ambiente.

Do que foi dito sobre a linguagem, o ambiente, o projeto,…, R, a notícia publicada no jornal New York Times em 6 de janeiro de 2009, vem dar uma visão – ainda atual – do que é este "dialeto" do S:

"R também é o nome de uma linguagem de programação popular usada por um número crescente de analistas de dados dentro de empresas e universidades. Está-se a tornar a língua franca, em parte, porque a mineração de dados entrou numa era de ouro, seja sendo usada para definir os preços dos anúncios, encontrar novas drogas mais rapidamente ou modelos financeiros afinar. Empresas tão diversas como a Google, Pfizer, Merck, Bank of America, o InterContinental Hotels Group e Shell, usam-no".

"R é similar a outras linguagens de programação, como C, Java e Perl, na medida em que ajuda as pessoas a executar uma grande variedade de tarefas de computação, dando-lhes acesso a vários comandos. Para os estatísticos, no entanto, R é particularmente útil porque contém uma série de mecanismos internos para organizar dados, executar cálculos sobre a informação e criar representações gráficas de conjuntos de dados".

"O que torna a R tão útil - e ajuda a explicar a sua rápida aceitação - é que os estatísticos, engenheiros e cientistas podem melhorar o código do software ou escrever variações para tarefas específicas".

"Perto de 1.600 pacotes (packages) diferentes - em 2009 – encontram-se em apenas um dos muitos sites dedicados a R, …,um pacote, chamado BiodiversityR, oferece uma interface gráfica que visa tornar o cálculo das tendências ambientais mais fáceis,…, outro pacote, chamado Ema, analisa os padrões de fala, enquanto GenABEL é usado para estudar o genoma humano, …"

“A comunidade de serviços financeiros tem demonstrado uma afinidade particular para com o R “.

“A grande beleza de R é que se pode modificá-lo para fazer todo tipo de coisas", disse Hal Varian, economista-chefe do Google.

E ainda: “R é uma demonstração real do poder da colaboração, e eu não acho que se poderia construir algo parecido com isso de outra maneira", disse R. Ihaka , um dos criadores do R.

### 3.2 O ambiente R

A linguagem de programação R é disponibilizada no site *www.r-project.org*. [26] Para fazer o *download* – depois de ligado à internet e aberto o site - seguir o *link* CRAN (*Comprehensive R Archive Network*) ou CRAN mirror, escolher um servidor – aconselha-se o mais próximo: http://cran.dcc.fc.up.pt/ em Portugal – e seguindo várias etapas (semelhantes a outros “*downloads*”) instala-se o programa R. No site *www.r-project.org* há muita informação sobre o programa R.
Ao abrir o programa – não necessitando de ligação à internet e se estiver em ambiente *Windows* – aparecerá uma *interface* semelhante à da seguinte figura – R Gui (*graphical user interface*) e R Console.

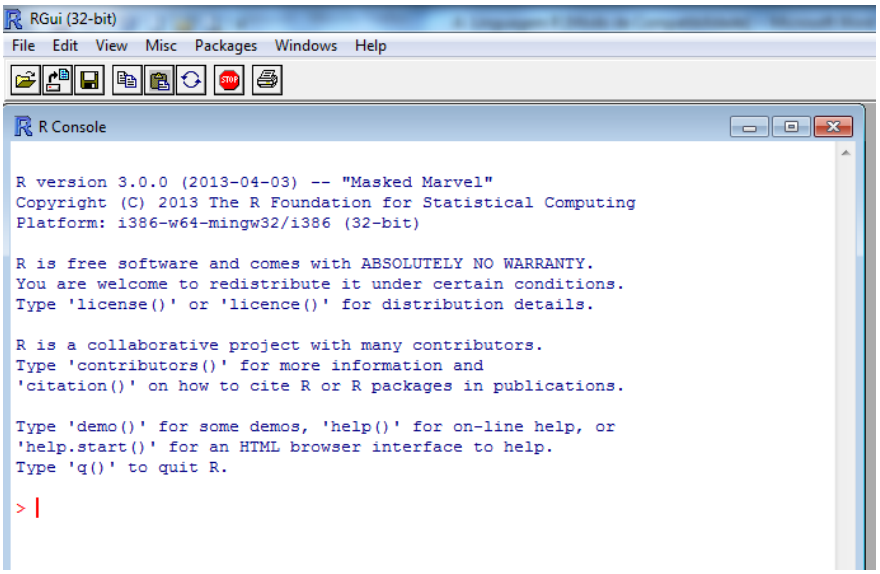



Figura 7 - Interface gráfica do R

Nesta última, pode ler-se: a versão – 3.0.0 a desta - "R é um software livre e vem SEM NENHUMA GARANTIA ", o que não é totalmente verdade, percebendo-se que "sem garantia " vem do facto de ser *open source* mas como projeto colaborativo que é, e dos seus contribuidores, não está em causa o seu desempenho, …

### 3.2.1 Começar a usar / terminar o R

Depois de aberto, aparece o símbolo " > ", chamado de *prompt* do R, com o cursor. O cursor será substituído por linhas de comando – para o R executar- finalizadas com a tecla ENTER.

```
> version
               _
platform       i386-w64-mingw32
arch           i386
os             mingw32
system         i386, mingw32
status
ma
minor          0.0
year           2013
month          04
day            03
svn rev        62481
language       R
version.string R version 3.0.0 (2013-04-03)
nickname       Masked Marvel
```

Para terminar uma sessão, utiliza-se a função "**q( )** " , e uma mensagem (questão) aparece " save workspace image? " com as possíveis respostas [ **yes , no ,cancel** ]:

- se **sim,** o R cria dois ficheiros - **.Rhistory**, com os comandos executados e **.RData**, com os objetos da sessão – no diretório de trabalho. Para saber este diretório usa-se "**getwd( )**" e para uma possível alteração " **setwd( )** ", especificando o caminho:

> getwd()

[1] "C:/Users/… /Documents"

> setwd('C:/Users/… /Documents') ou > setwd('C:\\Users\\… \\Documents')

com "/ " ou " \\ " para separar os diretórios.

- se **não**, abandonamos o R sem guardar o estado atual, sendo salvaguardados os ficheiros.

### 3.2.2 Ajuda

O modo - mais simples - de aceder à ajuda do R será através do menu Help, que se encontra na aplicação. Aí encontram-se ajudas para a consola – atalhos,… - "as perguntas mais frequentes" (FAQ), manuais, ligações para as páginas principais do Projeto R e do CRAN,… O mesmo se encontra com > help.start( )

Para obter ajuda sobre uma função em particular do R, ( e.g. média ) podemos mandar executar os comandos: > help(mean) ou > ?mean e uma janela se abrirá com muita informação. Quando não sabemos o nome concreto da função, podemos efetuar a procura por texto, e agora as funções serão > apropos( ) ou >help.search( ), colocando o texto no argumento das funções. Se a dúvida é complexa, o uso da poderosa função RSiteSearch( ) ou do site http://finzi.psych.upenn.edu/search.html é uma boa escolha.

### 3.2.3 Importação / Exportação / Salvar dados

Os dados constituem o ponto de partida da análise estatística. O R guarda-os na memória do computador sob a forma de *objetos* (vetores, matrizes, fatores, listas, funções, data-frames, …). A importação pode ser feita " à mão " – introdução via teclado – ou colocá-los, no R, se armazenados noutras plataformas.

Os *objetos* – de que se falará mais adiante – criam-se começando por lhes dar um nome e ligando-os ao seu conteúdo através dos operadores " <- " , " -> " ou " = " ( este último só aceite nas versões mais recentes do R), assim:

```
> a<- 5 # criado o objeto a com o valor 5
> x<-a  # x recebe o valor a
> a->x  # x recebe o valor a
> x=a   # x recebe o valor a
```

( O símbolo " # " serve para colocar comentários e estes são ignorados, pelo R , depois do # e até ao fim da linha.)

Se os dados a importar estão armazenados num ficheiro de texto, então a função *read.table( )*:

*> dados<-read.table('dados.txt',header=T,sep=';',dec=',')*

No argumento da função podem-se escrever mais parâmetros – ver: >?*read.table( )* – mas os que foram escritos nesta são:

- file: *'dados.txt'*, local onde se encontra o ficheiro 'dados' com a extensão .txt , estes estão no ambiente do R, podendo ser procurado o caminho do diretório com a função 'getwd()' – como já visto na secção **3.2.1**.

- *header*: o T – de TRUE, em letras maiúsculas, pois o R é "sensível" a maiúsculas/minúsculas – indica que o ficheiro vem com a primeira linha dos dados, o que corresponde ao nome das colunas ou seja (quase sempre) o nome das variáveis.
- *sep*: permite indicar o separador de valores usado – ';'(ponto e virgula), ' '(um ou dois caracteres de espaço), '\t' (por uma tabulação).
- *dec*: indica o caracter para separar as casas decimais nos números reais – ','(vírgula) ou '.'(ponto).

Para importar ficheiros da internet, conhecido o URL, o processo é idêntico:

*> dados<-read.table('http:// www. …/dados.txt',header=T,sep=';',dec=',')*

Se os dados estão armazenados em tabelas do *Excel*, há a possibilidade de gravar esses ficheiros com extensão *.csv* (*comma separated values* –valores separados por vírgulas) e depois escrever uma das funções *read.csv( ), read.csv2( ), read.delim( ), read.delim2( ),* …

Uma outra possibilidade – com ficheiros do *Excel* - será utilizar a função *read.xls ( )* :

*> dados<-read.xls('dados.xls', sheet=2)*

Esta função permite explicitar o nome da folha e o número da *worksheet*, mas a sua disponibilidade encontra-se no *package* - libraria com funções - *gdata*. Outros formatos – de outros programas de "computação estatística" - também é possível a sua importação, sendo que para tal é preciso instalar *packages* ( o que veremos mais adiante).

Para versões *Windows* do R podemos, ainda, fazer o usual " *copy / paste* " : selecionamos a tabela de dados , fazemos *edit* + *copy*, e já no R, mandamos executar a instrução

*> dados<-read.table('clipboard ', header=T')*

Depois da importação dos dados e de realizarmos as tarefas a que nos propusemos, os dados podem ser exportados, podemos salvar o trabalho e sair do R.

**Exportar**:

*> write.table(nome.dos.dados,file=' C:\\Users\\… \\Documents', …)*

*# nome.dos.dados terá que ser uma "palavra" sem espaços – nomedosdados - ou em alternativas: como está - com pontos - ou nome_dos_dados – com 'underscore' .*

**Salvar**:

> savehistory(*´C:\\Users\\… \\Document\\.historial.txt')*

Guardamos os comandos introduzidos numa sessão.

> loadhistory(´*C:\\Users\\… \\Documents\\historial.txt'*)

Importamos para o R, a sessão especificada.

O comando rm () , remove objetos:

>rm(list = ls()).

### 3.2.4 R como máquina de calcular

O R também calcula o valor de expressões aritméticas – simples ou complexas – salvaguardando a prioridade das operações:

> 1+3*2**10  # $1+3x2^{10}$

[1] 3073

> x<-3; y<-4; z<-log(x^y,9); z  # valor da expressão: $\log_9 (3^4)$

[1] 2

A seguinte tabela tem algumas das funções matemáticas (aritméticas):

| Função | Descrição |
|---|---|
| sqrt(x) | raíz quadrada de x |
| abs(x) | valor absoluto de x |
| exp(x) | exponencial de x $$e^x$$ |
| log10(x) | logaritmo de x na base 10 |
| log(x) | logaritmo de x na base $$e$$ |
| log(x,n) | logaritmo de x na base n |
| factorial(x) | $$x!$$ |
| sin(x) cos(x) tan(x) | funções trigonométricas |
| asin(x) acos(x) atan(x) | funções trigonométricas inversas |
| runif(n) | geração de números pseudo-aleatórios entre 0 e 1 com distribuição uniforme |

Tabela 2 - Algumas das funções matemáticas

### 3.2.5 Objetos em R

Na secção **3.2.3** já se referiu que os dados eram guardados, no R, sob a forma de *objetos* (vetores, matrizes, fatores, listas, funções, datadas frames, …) e também se abordou o modo de como estes se criam. Acrescentar que os nomes dos objetos começam com uma letra e além desta, podem conter mais letras, pontos e números.

### 3.2.5.1 Vetores

Vetor – objeto mais básico para guardar dados – é uma estrutura de dados que permite armazenar um conjunto de valores do mesmo tipo: conjuntos de caracteres, números reais ou complexos e valores lógicos.

Para criar vetores com mais do que um elemento, usa-se a função **c ( )**, separando os seus elementos por vírgulas:

```
> v<-c(4,2.5,7); v
[1] 4.0 2.5 7.0
>  length(v)   # nº de elementos
[1] 3

> mode(v)   # tipo: null,logical,numeric,complex ou character
[1] "numeric"
```

Algumas das funções vetoriais usadas no R, aplicadas em Estatística:

| Operação | Descrição |
|---|---|
| choose(n,k) | Calcula $\binom{n}{k}$ |
| min(x) | valor mínimo do vector x |
| sum(x) | somatório dos valores de x |
| mean(x) | média aritmética dos valores de x |
| median(x) | mediana dos valores de x |
| range(x) | Indica os valores máximo e mínimo do vector x |
| var(x) | variância |
| sd(x) | Desvio padrão dos elementos do vector x |
| summary(x) | Calcula os extremos, os quartis e a média do vector x |
| cor(x,y) | correlação entre x e y |
| quantile(x) | vector contendo o mínimo, primeiro quartil, mediana, terceiro quartil e o máximo de x |
| IQR(x) | Amplitude inter-quartil dos elementos do vector x |
| quantile(x,probs=p) | Quantil de ordem p dos elementos do vector x |

Tabela 3 - Funções vetoriais usadas em R

### 3.2.5.2 Matrizes

Matriz são vetores com duas dimensões – com mais de duas chamam-se *arrays* – que, tal como os vetores, armazenam dados do mesmo tipo.

A criação das matrizes é semelhante à dos vetores, precisando, agora, de organizar os dados em linhas e colunas:

```
> a<-matrix(c(1,2,3,4,5,6),2,3,byrow=TRUE)  # preenchimento por linhas
> rownames(a)<-c("L1","L2")  # nome das linhas
> colnames(a)<-c("C1","C2","C3")  # nome das colunas
> a
   C1 C2 C3
L1  1  2  3
L2  4  5  6
```

Para aceder a elementos da matriz:

```
> a[2,3]  # elemento na linha 2 e coluna 3
[1] 6
```

Resolver sistemas de equações lineares com a função **solve( )** :

$$\begin{cases} 23\,x + 31\,y = 1 \\ 34\,x + 46\,y = 2 \end{cases}$$

```
>  A<-matrix(c(23,34,31,46),2,2); A
    [,1] [,2]
[1,]   23   31
[2,]   34   46
> B<-c(1,2); solve(A,B)
[1] -4  3
```

A solução é $x = -4$ e $y = 3$

### 3.2.5.3 Fatores

Fatores são vetores para a manipulação de dados categóricos ( ou nominais).

```
> b<-factor(c(0,0,1,0,1,1,0),labels=c("f","m")) ; b
[1] f f m f m m f
Levels: f m
> table(b) # conta o nº de cada nível (valor)
b
f m
4 3
```

### 3.2.5.4 Listas

Lista é um conjunto ordenado de objetos – não necessariamente do mesmo tipo, nem do mesmo comprimento. Os objetos são chamados de componentes – podendo ter um nome na sua ordenação – tendo os atributos dos vetores: comprimento e tipo (length e mode). A importância das listas deve-se ao facto do resultado de vários objetos serem sob a forma de listas.

```
> vetor<- 1:5 ; matriz<-matrix(10:15,2,3);fator<-factor(c("M","F","M","M"))
> lista<-list(vetor,matriz,fator)
> names(lista)<-c("vec","mat","sexo")
> mode(lista); length(lista) # tipo e comprimento da lista
[1] "list"
[1] 3
> lista[[2]]
     [,1] [,2] [,3]
[1,]   10   12   14
[2,]   11   13   15
> lista$sexo  # ou lista[["sexo"]]
[1] M F M M
Levels: F M
```

### 3.2.5.5 Data frames

Um *data frame* é um objeto do R que pode ser visto como uma matriz com vetores de tipos diferentes – numéricos e de caracteres – ou como uma lista , onde as componentes têm o mesmo comprimento. As tabelas de dados - muito utilizadas em Estatística – são consideradas *data frames* : as linhas são os indivíduos e as colunas são vetores com os nomes das variáveis.

```
> pauta<- data.frame(nºproc=c(223,564,123),nome=c("Ana","Carlos","Rui"),notas=c(
+ 15,17,16))
> pauta
  nºproc   nome  notas
1    223    Ana     15
2    564 Carlos     17
3    123    Rui     16
```

A criação dum *data frame* pode ser feito através da função **data.frame,** ou das funções **read.table** , convertendo a tabela com a função **as.data.frame,** …

Para manipular os *data frames* podem-se utilizar os métodos já descritos quer para matrizes quer para listas.

### 3.2.5.6 Funções

Função é um objeto de R. Há várias funções pré-definidas - como as que se têm vindo a utilizar - ou podemos criá-las. A criação de uma função consiste na atribuição do conteúdo – a um nome, tal como noutros objetos de R - que depois de transformado – por ordens, no corpo da função – é fornecido como resultado: número, gráfico, lista, …

argumento(s) → ordem(ns) → resutado

```
> x<-c(30,40,50)
> mean(x)  #  função  média pré-definida em R
Ou
> média <-function(x)
+ {
+ soma=sum(x)
+ nºobs=length(x)
+ média=soma/nºobs
+ return(média)
+ }
> média(x)  # função média criada
[1] 40
```

### 3.2.6 Packages

Packages são bibliotecas de programas, num domínio de aplicações, que foram criadas e disponibilizadas de forma gratuita. Ao instalar o R, um conjunto de packages são também instaladas: packages base – automaticamente carregadas e de uso mais comum pela sua importância. Para instalar outras packages, fazemo-lo através da função **install.packages( )**, desde que tenhamos uma ligação à internet :

```
> install.packages( lm )   #  lm - package de regressão linear
```

Para saber quais as packages que estão disponíveis no sistema:

```
>  library( )
```

Para utilizá-las:

```
>  library(rpart )   #  rpart -  package base, para árvores de regressão
```

### 3.2.7 Gráficos / Visualização de dados

Numa análise estatística, os gráficos são uma das componentes a ter em conta pois eles dão-nos uma visão da distribuição dos mesmos. No R, as capacidades gráficas são um componente muito importante e extremamente versátil. É possível criar uma grande

variedade de gráficos, assim como, definir novos tipos de gráficos. Os gráficos em R estão organizados em 2 tipos:

1. O sistema tradicioanl implementado na *package* graphics
2. O sistema de gráficos Trellis - gráfico de rede que combina vários gráficos simples, adequados para a visualização de dados multidimensionais – implementado na *package* grid e disponibilizado na *package* lattice.

As funções gráficas podem ser classificadas em 3 tipos:

- ✓ Funções gráficas de *alto- nível*, que cria novos gráficos na janela gráfica.
- ✓ Funções gráficas de *baixo nível*, que permite adicionar novas informações a gráficos já criados, tal como novos dados, linhas e etiquetas.
- ✓ Funções gráficas *iterativas*, que permitem adicionar ou remover informações aos gráficos, empregando um dispositivo apontador.

Para visualizar alguns destes gráficos, numa janela gráfica, digite-se o comando:

```
> demo(graphics)
```

Há funções para gráficos univariados – **hist( )**, **boxplot( )**, **barplot( )**, .. - com 3 variáveis – **persp( )**, **countour( )**, **image( )**, … - multivariados – **matplot( )**, **pairs( )**, …- sendo a função **plot( )** a mais usada pois é uma função genérica. A todas elas é possível acrescentar linhas, texto, títulos , legendas, fórmulas, setas, cor,…

```
> par(mfrow=c(1,2))  # divide a janela de visualização : 1 linha por 2 colunas
 > x <- rnorm(200)
> plot(x)
> hist(x, main = "histograma ", axes = F, xlab = "dados", ylab = "frequências absolutas")
> axis(1, at = seq(-2.5, 3.5, by = 0.5), pos = 0)
> axis(2, at = seq(0, 50, by = 10), pos = -2.5)
```

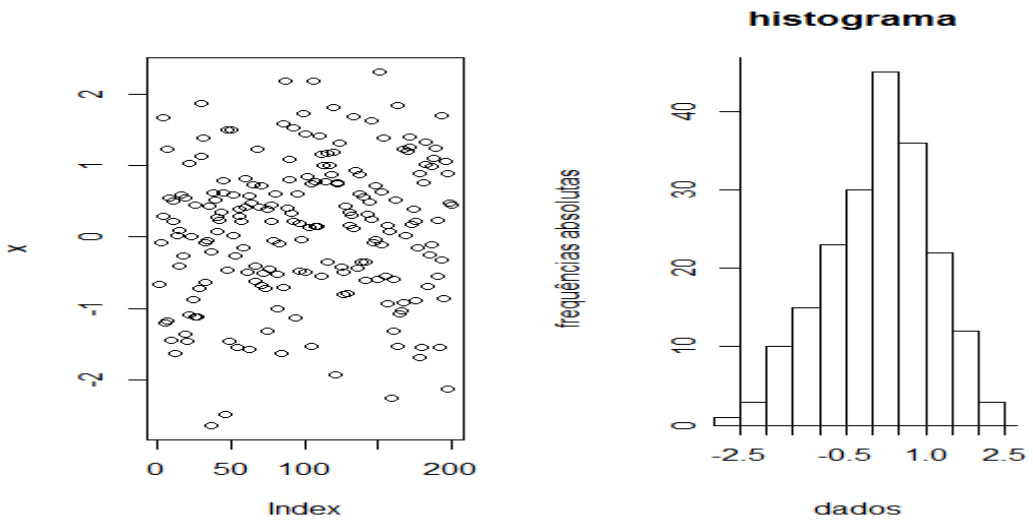
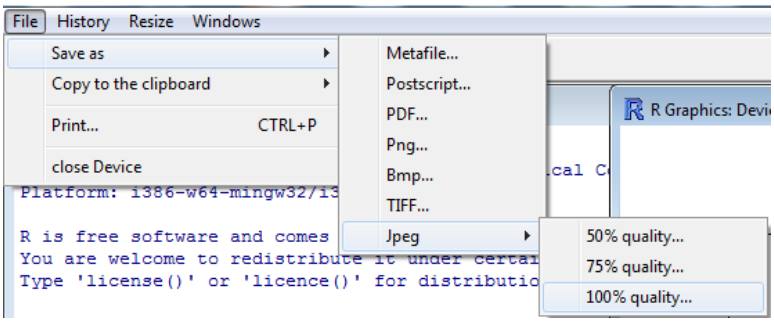



Gráfico 1 - Gráficos do R

Os principais parâmetros das funções "gráficas" são: *col* – cor , *main* – título do gráfico, *xlab,ylab* – título do eixo dos X's, Y's (respetivamente), xlim,ylim – 2 valores ( mínimo e máximo) para os eixos, *lty* – tipo de linhas e *pch* – símbolos para desenhar os pontos do gráfico.

Os gráficos podem ser salvos em vários formatos:




### 3.2.8 Programação em R

A programação na linguagem R é também possível – como já foi visto na construção de novas funções e nas pré-definidas – utilizando instruções de controlo de execução, que se apresentam a seguir de uma forma muito resumida e só as principais:

- **for** ( variável in sequência ) expressão
- **if** ( condição) expressão
- **if** ( condição) expressão **else** expressão alternativa
- **while** ( condição) expressão

- **repeat** expressão
- **break**
- **next**
- **ifelse** (teste, yes,no)

### 3.2.9 Packages (pacotes) utilizados

Além dos packages que foram mencionados em "**3.2 O ambiente R**" – constando uma pequena explicação depois do símbolo # - há ainda que referir os seguintes, que serão utilizados neste trabalho:

- abind Combina arrays
- lattice Gráficos Lattice
- MASS Funções de Suporte e conjuntos de dados de Venables and Ripley
- rpart Recursive Partitioning (particionamento recursivo)
- lm Modelo de Regressão Linear

# PARTE II

## 4. Aplicação na avaliação do desempenho na disciplina de Matemática

### 4.1. Recolha e descrição dos dados.

O questionário - que se apresenta em anexo – foi utilizado no estudo e avaliação do desempenho na disciplina de Matemática. Para tal, foram preenchidos 85 questionários – sendo todos validados -, por alunos que frequentam o 12º ano de escolaridade das escolas secundárias de Bragança no ano letivo de 2010/ 2011 – população do estudo - e tratadas algumas das questões, formuladas nos mesmos.

A construção do questionário seguiu uma linha onde se começa por conhecer a situação dos pais – e do agregado familiar - para depois chegar ao aluno. Dos pais pretende-se saber: as habilitações literárias, a situação laboral, as relações no seio do agregado familiar, o rendimento, … Do aluno: o tempo dedicado ao estudo, os meios postos ao se dispor, … Resumindo: tentar verificar se o estatuto socioeconómico do agregado familiar, fatores demográficos, fatores específicos do aluno, explica o desempenho, deste último, na disciplina de Matemática.

### 4.2.Análise dos dados

Os dados retirados do questionário serão analisados, recorrendo ao programa estatístico R.

Inicia-se por uma análise descritiva, seguindo-se os dois métodos de Regressão: Regressão Linear Múltipla – método paramétrico – e Árvore de Regressão Binária – método não paramétrico.

#### 4.2.1. Análise Descritiva

A análise descritiva dos resultados dos inquéritos será feita recorrendo a tabelas, gráficos, medidas de tendência central e de dispersão, …, formas usadas na Estatística

Descritiva para uma primeira abordagem ao estudo, dando assim a conhecer as principais características.

No estudo, responderam ao inquérito 85 indivíduos, sendo 34 (40%) do sexo masculino e 51 (60%) do sexo feminino, como se mostra no gráfico seguinte:

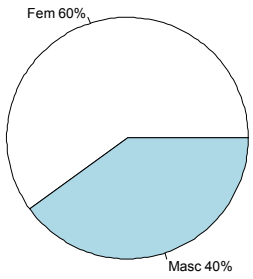



Gráfico 1 - Distribuição dos inquiridos por género

Passando agora para uma descrição dos pais, começamos por verificar (Gráficos 2 e 3) que os pais têm habilitações inferiores às mães: 28.2% dos pais tiraram um curso superior enquanto que as mães estão nos 43.55%; com o 12ºano e Curso Superior, os pais totalizam 52.9% , as mães 72.9% , .

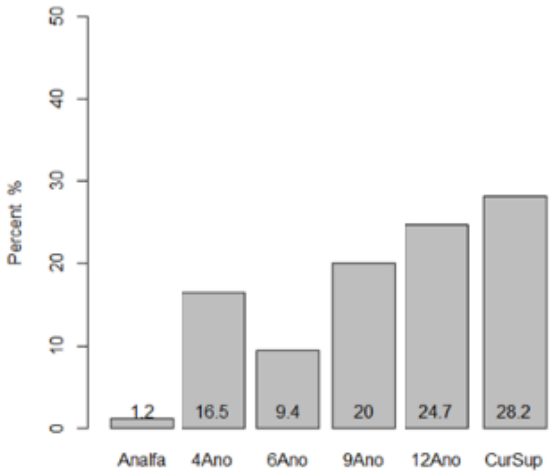



Gráfico 2 - Habilitações Literárias do Pai

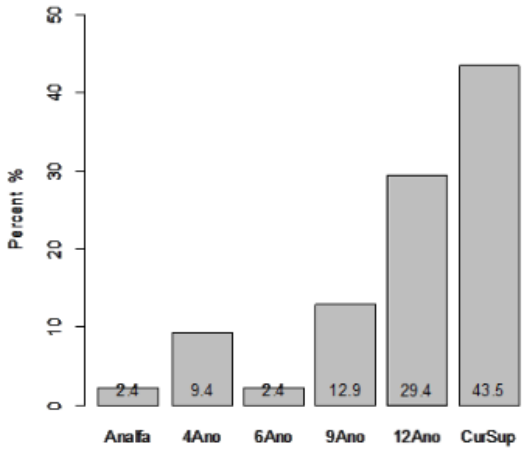



Gráfico 3 - Habilitações Literárias da Mãe

A Situação Laboral no agregado familiar apresenta as seguintes distribuições (Gráficos 4 e 5 ) : tanto os pais como as mães trabalham maioritariamente por conta de outrem ( 63.5 % dos pais e 70.6 % das mães) ; na situação de temporariamente desempregados, os pais apresentam um valor mais baixo ( 3.6 % contra 15.3 %) ; …

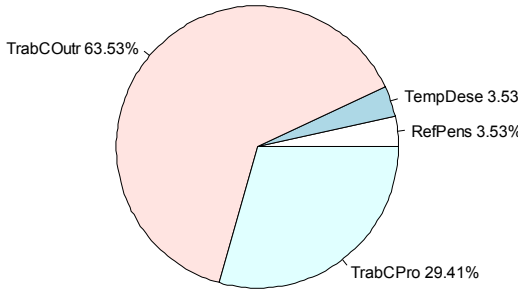



Gráfico 4 - Situação Laboral do Pai

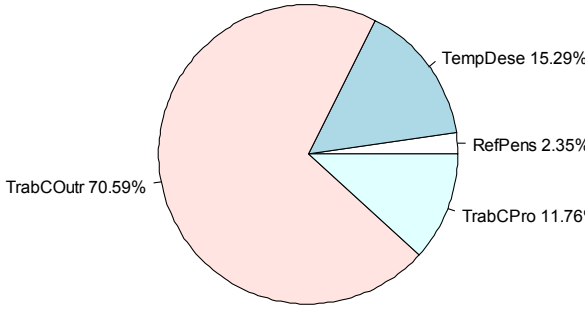



Gráfico 5 - Situação Laboral da Mãe

O Estado Civil dos pais dos inquiridos é de 91.8 % Casados, 5.9% Divorciados/Separados e as categorias de União de facto e Viúvo/a repartem a restante percentagem com valores muito baixos (Gráfico 6)

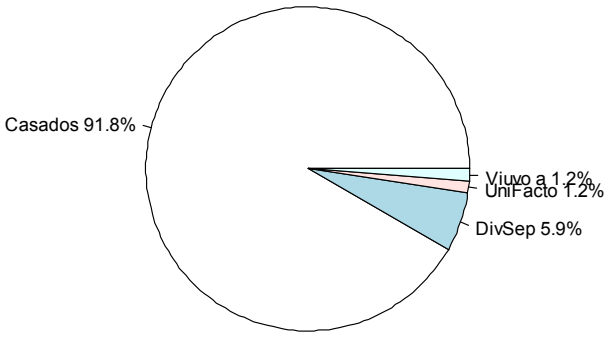



Gráfico 6 - Estado Civil dos pais

Na distribuição dos Encarregados de Educação, podemos observar e destacar as mães com 68.2%, os pais representam 25.9 %, com 2.4 % aparecem os próprios e em outras situações além das indicadas 3.5 % (Gráfico 7)

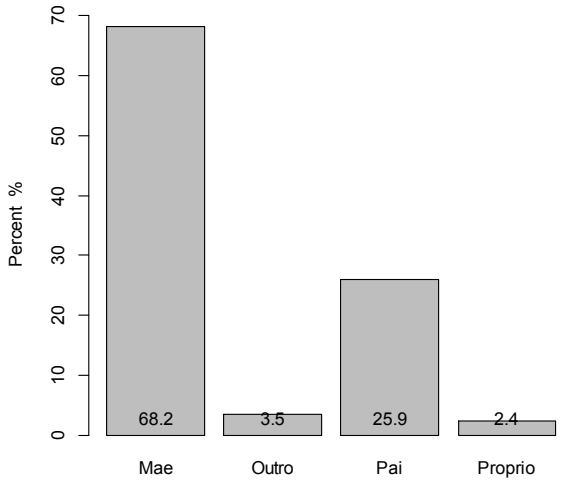



Gráfico 7 - Encarregado de Educação

O número de pessoas do agregado familiar (Gráfico 8) varia entre 2 e 6, sendo 51.8 % dos casos com 4 pessoas e 36.5 % com 3 pessoas, …

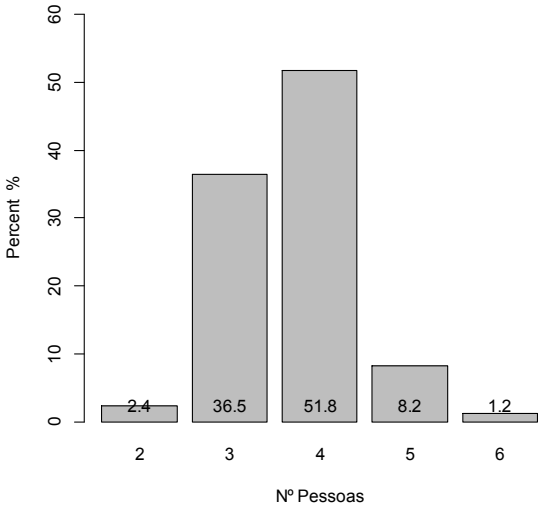



Gráfico 8 - Número de pessoas do agregado familiar

Ainda no que respeita ao agregado familiar, e questionados com quem viviam, 37.6 % moram com os pais, 51.8 % com irmãos, 4.8 % vive com um dos pais (2.4% com pai e 2.4% com a mãe), 5.9 % encontra-se noutra situação não descrita anteriormente (Gráfico 9)

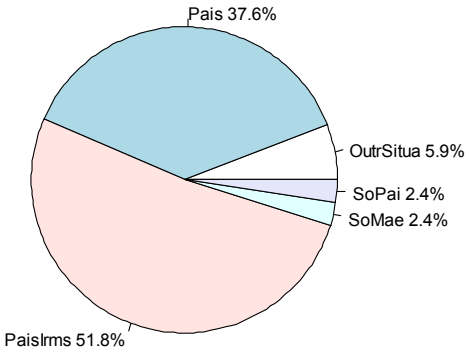
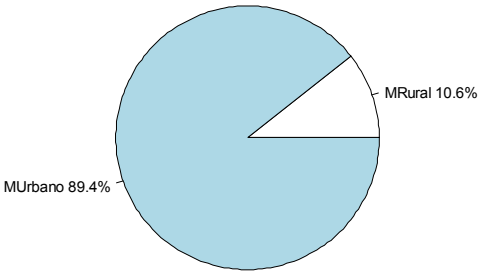



Gráfico 9 - Pessoas com quem vivem




Questionados sobre o seu local de residência - Meio Urbano ou Meio Rural - 89.4% respondeu "Meio Urbano" e 10.6% "Meio Rural" (Gráfico 10)

Gráfico 10 - Local de Residência

Quanto à mobilidade dos inquiridos, o tempo (médio) que gastam no percurso de casa à escola é apresentado no gráfico 11 e constata-se que mais de 2/3 demora aproximadamente 8 minutos, cerca de 1/4 gasta 22.5 minutos ,…

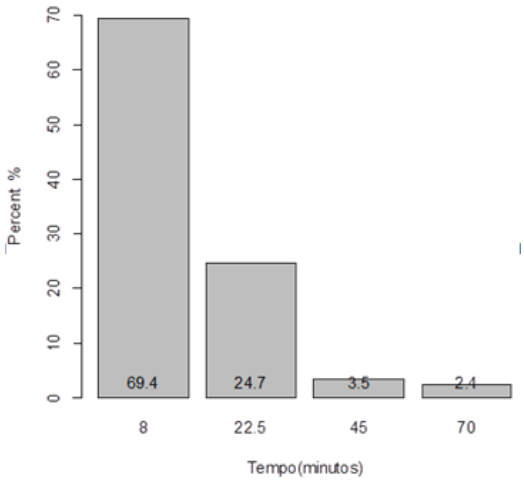



Gráfico 11 - Tempo Casa /Escola

À questão "Como classificas as relações familiares em casa?", numa escala de 1 - Conflituosas – a 5 – Excelentes – responderam 41.2 % 4, 37.6 % Excelentes, 16.5 % 3,…(Gráfico 12 )

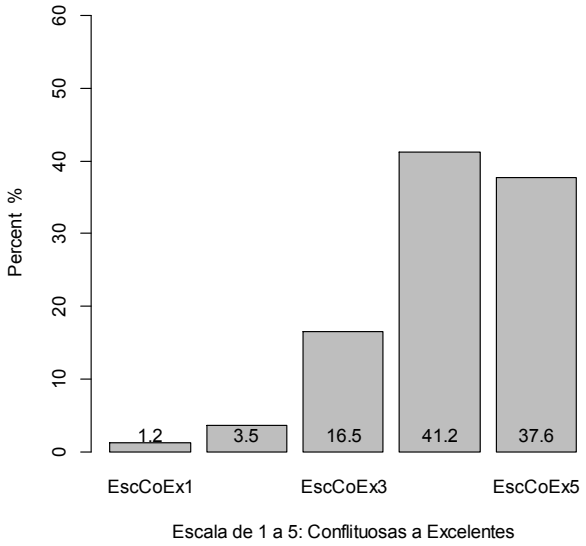



Gráfico 12 - Relações familiares em casa

O rendimento mensal líquido do agregado familiar dos inquiridos ( Tabela 4 e Gráficos 13 e 14) varia entre 500 € e 6000 €, com uma média de 1788 € - muito próxima do valor mediano que é de 1750 € - apresentando frequências muito semelhantes para os valores entre 750 e 2500, …

(Considerando o valor de 6000 € um caso isolado, aberrante,… - um outlier – as medidas supra mencionadas não seriam muito diferentes , excetuando que a variação seria de 500 a 4000 – média passaria para 1738, … )

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 500 | 1250 | 1750 | 1788 | 2500 | 6000 |

Tabela 4 - Rendimento Mensal líquido do Agregado Familiar

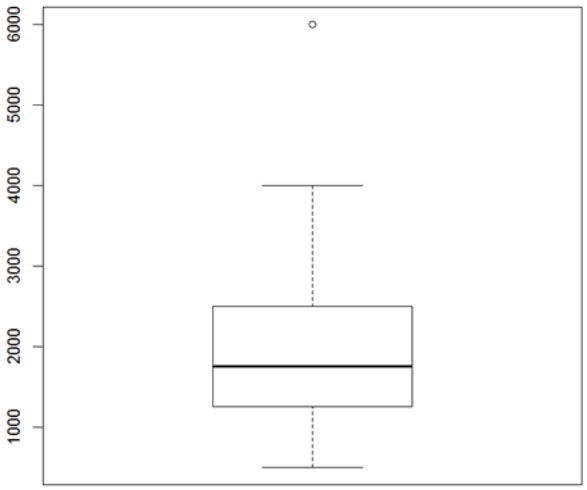



Gráfico 13 - Caixa de bigodes para Rendimento Mensal líquido do Agregado Familiar

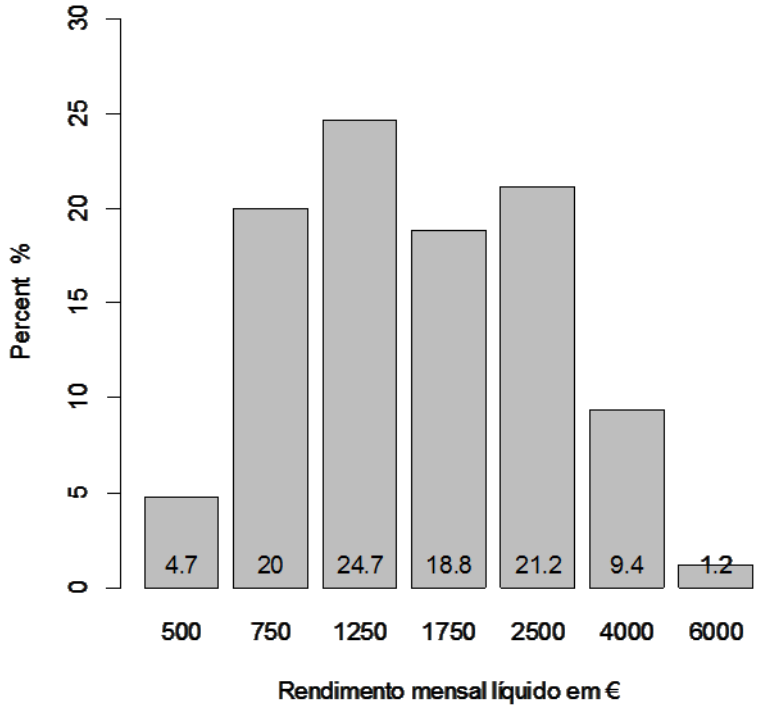



Gráfico 14 - Rendimento Mensal líquido do Agregado Familiar

Sobre o seu percurso académico, foram questionados quanto ao número de reprovações e o resultado é apresentado no Gráfico 14: 90.6 % nunca reprovou, 8.2 % reprovou uma vez, 1.2 % reprovou duas vezes, ou seja o número de reprovações é inferior a 10 %.

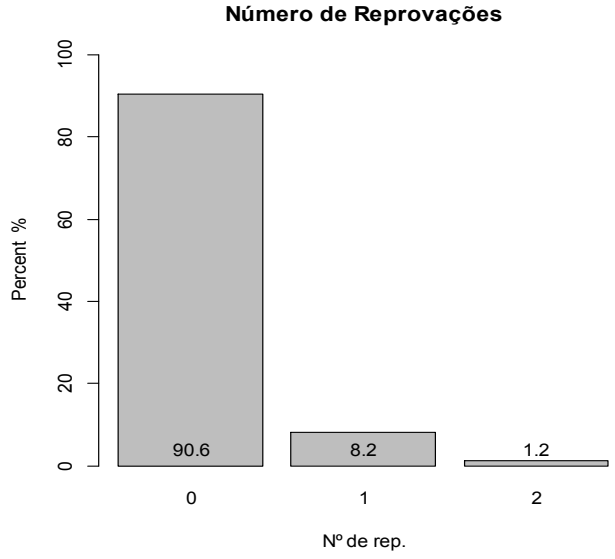



Gráfico 15 - Número de Reprovações

Questionados sobre a "Média obtida, na disciplina de Matemática, no 10º e 11º ano" ( Questão 16) , os resultados são os apresentados na tabela 5 e nos gráficos 15 e 16:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 8.00 | 11.00 | 14.00 | 14.04 | 16.00 | 19.00 |

Tabela 5 - Média obtida, na disciplina de Matemática, no 10º e 11ºano

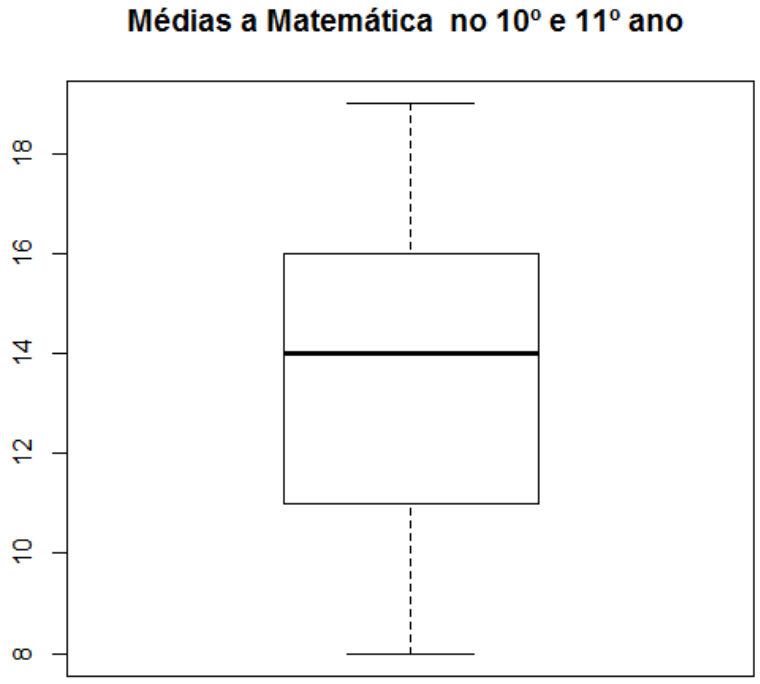



Gráfico 16 - Caixa de bigodes para as Médias a Matemática no 10º e 11º ano

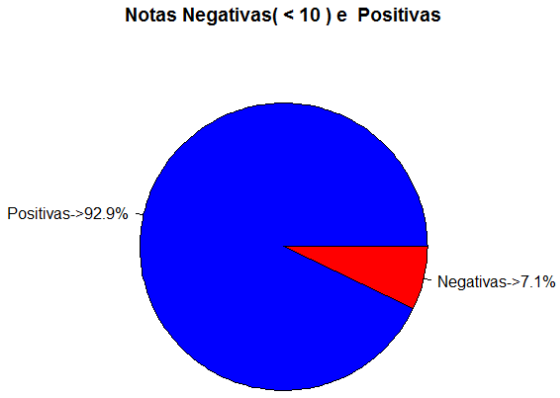



Gráfico 17 - 16 Positivas e Negativas da Questão 16

Verifica-se que as médias variam entre o mínimo de 8.0 e um máximo de 19.0, com valores muito próximos para a média e mediana (que são de 14.0), com 7.1 % de valores inferiores a 10 – "Negativas" – e 92.9 % de valores superiores ou iguais a 10 – "Positivas" - , …

A "Média obtida, nas restantes disciplinas, no 10º e 11º ano" (Questão 17), (Tabela 6 e Gráfico 17), teve valores entre 10.5 e 19.2, média de 15.5, …

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 10.50 | 14.00 | 15.00 | 15.48 | 17.00 | 19.20 |

Tabela 6 - Média obtida, nas restantes disciplinas, no 10º e 11ºano

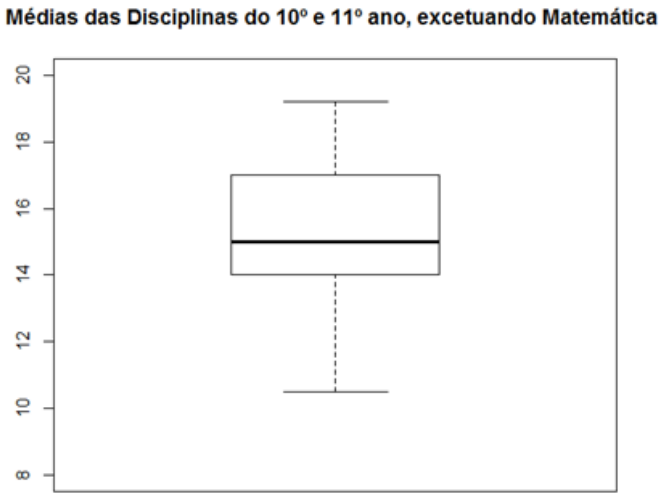



Gráfico 18 - Caixa de bigodes para as Médias obtidas, nas restantes disciplinas, no 10º e 11ºano

Relativamente às horas dedicadas semanalmente ao estudo (Gráfico 18), os dados apontam para que quase metade dos inquiridos (48.2 %) se ocupem nesta tarefa entre 2 a 5 horas, seguindo-se 23.5 % com 5 a 10 horas, empreendem-se 12.9 % em mais de 10 horas e os restantes (15.3 %) em menos de 2 horas.

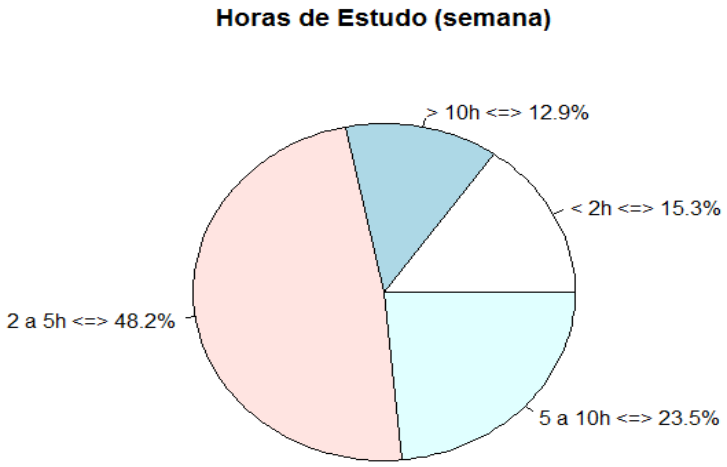



Gráfico 19 - Horas de Estudo por semana

Quanto a ter Computadores e Internet em casa, todos responderam "Sim", (Gráficos 19 e 20).


Gráfico 20 - Computador em casa

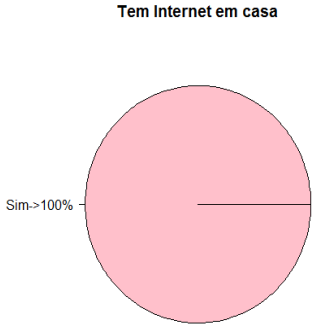



Gráfico 21 - Internet em casa

Sobre o "Local de Estudo"(Gráfico 21) , 97.7 % responderam "Em casa" e 2.4 % em outro local que não "Em casa" ou "Em casa de amigos" ou "Na escola".

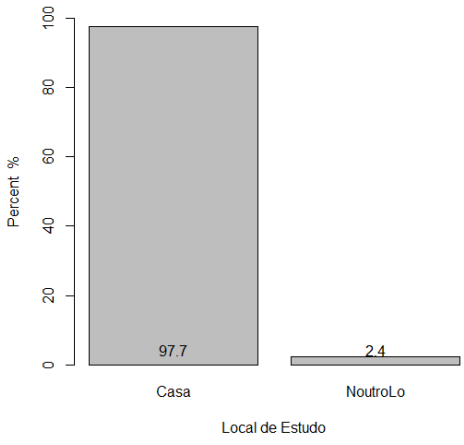



Gráfico 22 - Onde estudas habitualmente?

Ainda no contexto do trabalho escolar realizado em casa, e questionados sobre esses mesmos trabalhos (TPC's): "Fazes os TPC?" e "Tens ajuda nos TPC", apura-se que 2.4 % nunca faz os TPC's, embora haja 21.2 % que só os faz "Às vezes" e 40.0 % "Muitas Vezes" , restando 36.5% ( aproximadamente 1 em cada 3) que faz "Sempre" os TPC's (Gráfico 22). Em resposta à segunda questão, 72.9 % não tem ajuda nos TPC's e dos sobrantes, 8.2 % tem efetivamente ajuda (Gráfico 23), …

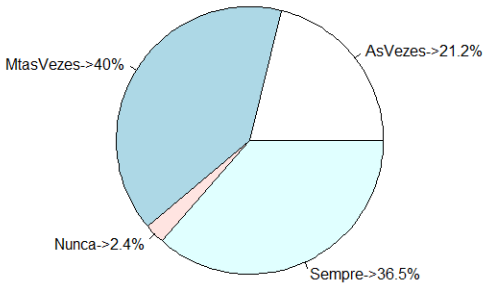



Gráfico 23 - Fazes os TPC?

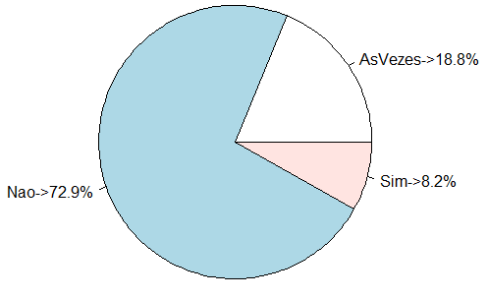



Gráfico 24 - Tens ajuda nos TPC?

À questão específica " Usas recursos informáticos nas aulas de Matemática?", 50.6 % admite nunca usar, 43.5% usa "Às vezes" e uma minoria (5.9 %) usa com frequência (Gráfico 24).

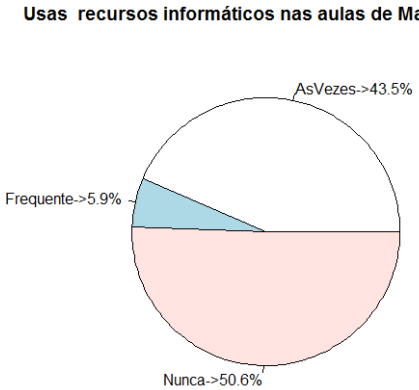



Gráfico 25 - Uso de recursos informáticos nas aulas de Matemática?

À última questão do questionário: “Desejas frequentar um curso superior?”, a resposta foi unânime, 100% de “Sim” (Gráfico 25)


Gráfico 26 - Desejo de frequência de Curso Superior

### 4.2.2. Análise pelas Regressões

Nas duas Regressões tenciona-se explicar se há relação entre a variável **MDM** – questão 16: Média obtida, na disciplina de Matemática, no 10º e 11º ano (variável dependente) – e as variáveis ( questões) que a seguir se enunciam:

- **NAF**- questão 9: Número de pessoas do agregado familiar
- **TCE** – questão 12: Quanto tempo demoras de casa à escola
- **RMAF** – questão 14: Rendimento mensal líquido do agregado familiar
- **MRD** – questão 17: Média obtida, nas restantes disciplinas, no 10º e 11º ano
- **HE** – questão 18: Quantas horas estudas (por semana)?

#### 4.2.2.1. Regressão Linear Múltipla

O estudo começa com a importação dos dados de um ficheiro, previamente guardado com a extensão .csv (comma separated values), com o nome das variáveis no cabeçalho e a vírgula como indicador de decimais

```
> QCt<-read.table('QCi.csv', header=TRUE, sep=';',dec=',')
> QCt
> cor(QCt)  # matriz de correlação
```

| | NAF | TCE | RMAF | MDM | MRD | HE |
|---|---|---|---|---|---|---|
| NAF | 1.000000000 | 0.03617395 | 0.1407873 | 0.1687991 | 0.1815595 | -0.007021078 |
| TCE | 0.036173952 | 1.00000000 | -0.1986068 | -0.1289540 | -0.1344208 | -0.140320100 |
| RMAF | 0.140787265 | -0.19860680 | 1.0000000 | 0.3249249 | 0.3211324 | 0.136564033 |
| MDM | 0.168799107 | -0.12895404 | 0.3249249 | 1.0000000 | 0.8686211 | 0.323414748 |
| MRD | 0.181559549 | -0.13442075 | 0.3211324 | 0.8686211 | 1.0000000 | 0.306288274 |
| HE | -0.007021078 | -0.14032010 | 0.1365640 | 0.3234147 | 0.3062883 | 1.000000000 |

Dos valores da matriz, pode-se observar que variável MRD é a que está mais correlacionada com a variável MDM (variável resposta) e as outras variáveis preditoras não apresentam multicolinearidade.

Com histogramas investiga-se a distribuição das variáveis:

```
> hist(QCt$MRD)
```

…

Por análise dos histogramas – que se apresentam no anexo2 – podemos verificar que as variáveis não se distribuem da mesma forma e algumas não apresentam distribuição "próxima" da normal. As varáveis **TCE**, **RMAF** e **HE**, apresentam distribuição com forte assimetria à direita, sendo objeto de transformações futuras.

Para analisar o padrão de relacionamento entre a variável resposta (**MDM**) e as variáveis preditoras, constroem-se diagramas de dispersão.

```
> plot(QCt$NAF,QCt$MDM)   # diagrama de dispersão entre as var. MDM e NAF
```

…

Da análise dos diagramas – anexo3 - nota-se que, excetuando a relação entre MDM e MRD, o relacionamento não é linear, o que reforça a ideia já mencionada anteriormente de que haverá lugar a transformações.

Continuando a análise, iremos estimar a regressão, com as variáveis no escala original, mas extraímos uma amostra aleatória de dados para no final validarmos o modelo.

Como dispomos de 85 observações e as candidatas a preditoras são 5, dividimos os dados em duas amostras: uma – a de estimação – com 60 observações e a de validação com as restantes 25 observações. Este critério aqui adotado - ainda que com alguma dose de subjetividade – teve em conta outros que se encontram na literatura consultada, e que está em "Referências Bibliográficas": há recomendações para que a razão, entre as variáveis preditoras e as observações, seja de 1 para 20 [3], 1 para 15, 1 para 10, nunca inferiores a 1 para 6 ou 5 [16]. As amostras – de estimação e validação – serem do mesmo tamanho.

A amostra de estimação é extraída no R:

```
>QCEst<- data.frame(QCt$MDM,QCt$NAF,QCt$TCE,QCt$RMAF,QCt$MRD,QCt$HE)
>  QCEst <- sample(nrow(QCt),60)  # extração da amostra de estimação
>  QCEst  # registos da amostra de estimação
 [1] 49  8 17 55  2 43 35 67  3 21 76 77 57 12  1 75 82 15  7  9 66 28 48 64 14
[26] 56 10 24 33 74 11 16 20 38 46 27 69 23 84 29 70 41 36 50 30 71 54 73  6 42
[51] 51 53  5 19 34 83 47 80 32 39
```

( Linhas dos dados originais que fazem parte da amostra de estimação, as restantes vão para a de avaliação.)

O modelo estimado – com as variáveis na escala original e a amostra de estimação ( QCEst ) – será:

```
> RM <- lm(MDM ~ NAF + TCE+ RMAF+ MRD+ HE,data=QCt [QCEst,])
> summary(RM)
```

---

```
Call:
lm(formula = MDM ~ NAF + TCE + RMAF + MRD + HE, data = QCt[QCEst,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-4.8734 -0.6726 -0.0624  0.8365  5.4597

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.313e+00  2.106e+00  -3.472  0.00102 **
NAF          2.310e-01  3.130e-01   0.738  0.46373
TCE         -7.382e-03  2.033e-02  -0.363  0.71788
RMAF         6.448e-05  2.569e-04   0.251  0.80280
MRD          1.310e+00  1.437e-01   9.114 1.64e-12 ***
HE           3.720e-02  9.034e-02   0.412  0.68217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   Residual standard error: 1.779 on 54 degrees of freedom
   Multiple R-squared:  0.7253,    Adjusted R-squared:  0.6999
   F-statistic: 28.52 on 5 and 54 DF,  p-value: 5.027e-14
```

Do sumário supra, vê-se que a distribuição dos erros residuais não está centrada no valor 0 – a mediana deveria estar perto de 0. Os parâmetros estimados (coeficientes: $\beta$ 's ) estão com níveis de confiança "muito baixos"- exceção para os dois que têm asteriscos . O grau de ajustamento é de aproximadamente 70%, por fim, a estatística F permite concluir que as variáveis preditoras são significativas para a explicação da variabilidade em MDM.

Da análise de resíduos – gráficos de resíduos versus valores ajustados e de resíduos versus variáveis preditoras, normalidade dos resíduos – estudam-se os seguintes pressupostos para o modelo de regressão:

1. linearidade da função de regressão;
2. homocedasticidade (homogeneidade de variância dos erros);
3. independência dos erros;
4. Normalidade dos erros.

Dos gráficos:

> plot (fitted (RM) ,residuals (RM) ) # gráfico de resíduos vs valores ajustados

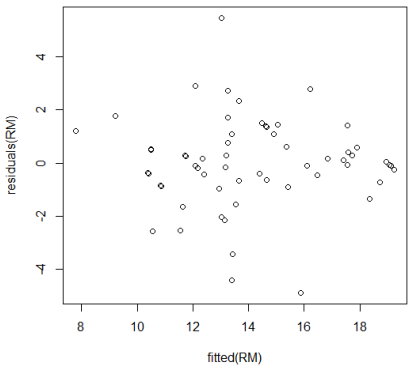



Gráfico 27 – Gráfico dos Valores Ajustados versus Resíduos

do gráfico que se apresentou, pode- se ver que há homogeneidade de variância dos erros – ainda que não muito forte.

> plot(QCt[QCEst, ]$NAF ,residuals(RM)) # gráfico de resíduos versus preditora (NAF)

…

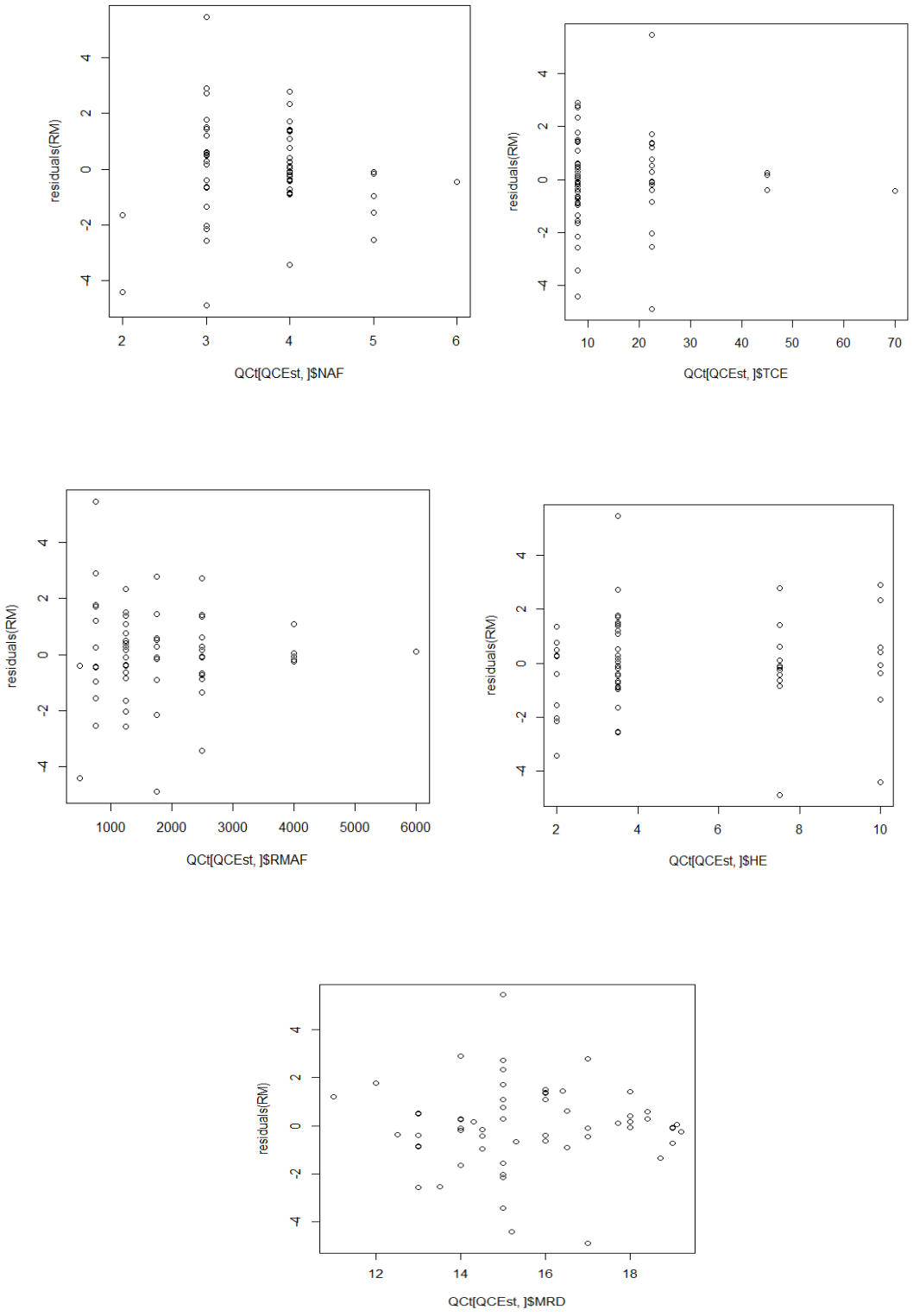



Gráfico 28 - Gráficos de Resíduos versus preditoras

Observando o gráfico de normalidade dos resíduos:

```
> qqnorm(residuals(RM))          # o gráfico de normalidade de resíduos
```

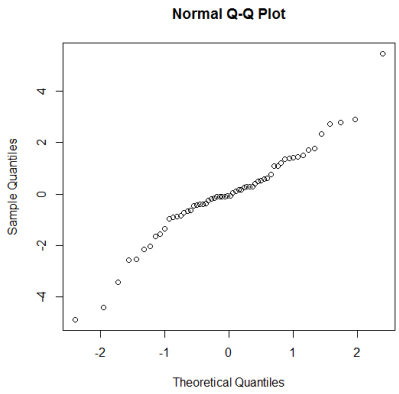



Gráfico 29 - Gráfico de normalidade dos resíduos

Este mostra a não-linearidade, indicando a não sustentação da suposição de normalidade dos erros da regressão.

De outro tipo de gráficos – de resíduos semi-studentizados – que a seguir se apresenta um:

```
> residuo <-RM$residuals
> sigma <-summary(RM)$sigma
> residuosemistud <-residuo/sigma
> plot(QCt[QCEst, ]$MDM,residuosemistud)   #  gráfico de resíduos semi-studentizados
...
```

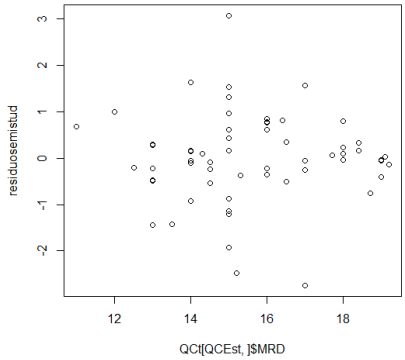



Gráfico 30 - Gráfico de resíduos semi-studentizados

Podemos retirar as seguintes conclusões: há *outliers* e pela sua dispersão nota-se que o ajustamento à superfície de regressão estimada não é muito bom.

Após a análise exploratória e diagnóstica feita – e porque alguns dos pressupostos não se verificaram - seguem-se umas transformações de variáveis,

construindo novos dataframes de estimação, para reestimar a regressão e haver lugar a novo diagnóstico.

```
> LnMDM<-log(QCt[QCEst, ]$MDM)
> LnNAF<-log(QCt[QCEst, ]$NAF)
> LnTCE<-log(QCt[QCEst, ]$TCE)
> LnRMAF<-log(QCt[QCEst, ]$RMAF)
> LnHE<-log(QCt[QCEst, ]$HE)
> MRD<-(QCt[QCEst, ]$MRD)  # não transformada

> DFE2<- QCt[QCEst, ]  # dataframe com variáveis transformadas
> DFE2<- cbind(DFE2,LnMDM)    # Combine R Objects by (Rows) or Columns
> DFE2<- cbind(DFE2,LnNAF)
> DFE2<- cbind(DFE2,LnTCE)
> DFE2<- cbind(DFE2,LnRMAF)
> DFE2<- cbind(DFE2,LnHE)
> DFE2<- cbind(DFE2,MRD)

> RMT<-lm(LnMDM~ LnNAF+LnTCE+LnRMAF+LnHE+MRD,data=DFE2)      # regressão com var transformadas
> summary(RMT)
```

---

```
Call:
lm(formula = LnMDM ~ LnNAF + LnTCE + LnRMAF + LnHE + MRD, data = DFE2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34400 -0.04910  0.00107  0.07051  0.38185

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.939746   0.299316   3.140  0.00274 **
LnNAF        0.109404   0.086952   1.258  0.21372
LnTCE       -0.003418   0.032387  -0.106  0.91635
LnRMAF       0.008648   0.038644   0.224  0.82376
LnHE         0.013538   0.036278   0.373  0.71049
MRD          0.094161   0.011191   8.414 2.13e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1368 on 54 degrees of freedom
Multiple R-squared:  0.7024,    Adjusted R-squared:  0.6748
F-statistic: 25.49 on 5 and 54 DF,  p-value: 4.184e-13
```

> RMT2<-lm(LnMDM~ LnNAF +LnRMAF+LnHE+MRD,data=DFE2)  # regressão com var transformadas semTCE

> summary(RMT2)

```
Call:
lm(formula = LnMDM ~ LnNAF + LnRMAF + LnHE + MRD, data = DFE2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34676 -0.04827  0.00203  0.07100  0.38060

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.922252   0.246960   3.734 0.000448 ***
LnNAF       0.108650   0.085875   1.265 0.211128
LnRMAF      0.009932   0.036349   0.273 0.785697
LnHE        0.014165   0.035465   0.399 0.691137
MRD         0.094138   0.011088   8.490  1.4e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1356 on 55 degrees of freedom
Multiple R-squared:  0.7023,    Adjusted R-squared:  0.6807
F-statistic: 32.44 on 4 and 55 DF,  p-value: 6.858e-14
```

> summary.aov(RMT2)  # sumário da análise da variância do modelo

```
            Df Sum Sq Mean Sq F value   Pr(>F)
LnNAF        1 0.2129  0.2129   11.58  0.00125 **
LnRMAF       1 0.5914  0.5914   32.16 5.43e-07 ***
LnHE         1 0.2563  0.2563   13.94  0.00045 ***
MRD          1 1.3253  1.3253   72.08 1.40e-11 ***
Residuals   55 1.0113  0.0184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> plot (fitted (RMT2) ,residuals (RMT2) )  # gráfico de resíduos  vs  valores ajustados

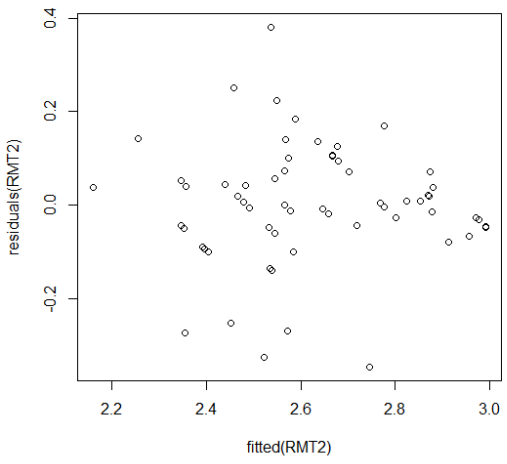



Gráfico 31 - Gráfico de resíduos vs valores ajustados

```
> qqnorm(residuals(RMT2))
```

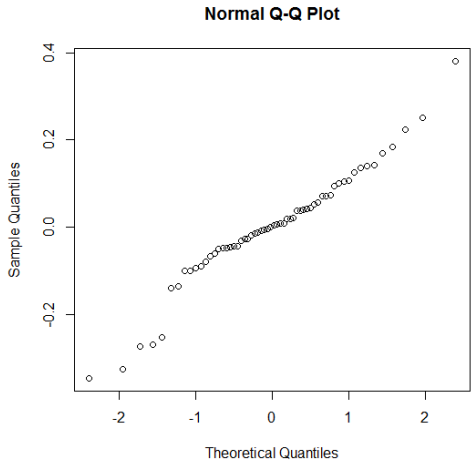



Gráfico 32 - Gráfico Normal Q-Q Plot

Desta última análise foi retirada a variável TCE, tendo em conta que por observação da matriz inicial de correlação, ela era a menos correlacionada com a variável MDM e depois de transformada – por observação do teste t – pouco explica, pouco significativa para o modelo…

A validação deste modelo – depois de transformar variáveis – também não é feita devido aos valores da coluna P( > $|t|$ ) serem muito superiores a 0.05; do gráfico 31, concluiu-se que a variância não é constante; do gráfico 32, observa-se que não há linearidade – pois não é possível ajustar os valores médios de Y a uma reta,… Por estas razões e por outras já vistas no modelo anterior – e que se repetem – e também por razões que se enunciam no Capítulo5, o método de regressão linear múltipla foi "abandonado", dando lugar ao método por Árvores de Regressão Binárias.

#### 4.2.2.2. Árvores de Regressão Binárias

As Árvores de Regressão Binárias têm sido aplicadas em diversas áreas: medicina, financeiras, ambientais, … Elas têm sido reconhecidas como uma ferramenta de modelação útil entre os estatísticos e os não-estatísticos, pois elas produzem um modelo que é muito fácil de interpretar.

Estas árvores têm uma série de características interessantes, que iremos resumir:

- Usam técnicas não-paramétricas e, por isso, não requerem suposições de normalidade dos dados.
- É possível manipular dados de diferentes tipos: contínuo, categórico, ordinal, binário. Transformações, dos dados, não são necessárias.
- As árvores podem ser úteis para a deteção de variáveis importantes, interações e identificar outliers. Isto pode ser útil na fase exploratória de modelagem.
- Lidam bem com dados faltantes, identificando um substituto.

…

Esta segunda análise - com a técnica das árvores de regressão binárias - iremos realizá-la com a library ***rpart*** [1], desenvolvida por Beth Atkinson e Therneau Terry, que implementa uma metodologia próxima da tradicional versão CART, devido a Breiman *et al.* (1984).

```
> dados<-read.table('QCi.csv', header=TRUE, sep=';',dec=',')
> summary(dados)
      NAF              TCE             RMAF            MDM             MRD
 Min.   :2.000   Min.   : 8.00   Min.   : 500   Min.   : 8.00   Min.   :10.50
 1st Qu.:3.000   1st Qu.: 8.00   1st Qu.:1250   1st Qu.:11.00   1st Qu.:14.00
 Median :4.000   Median : 8.00   Median :1750   Median :14.00   Median :15.00
 Mean   :3.694   Mean   :14.35   Mean   :1788   Mean   :14.04   Mean   :15.48
 3rd Qu.:4.000   3rd Qu.:22.50   3rd Qu.:2500   3rd Qu.:16.00   3rd Qu.:17.00
 Max.   :6.000   Max.   :70.00   Max.   :6000   Max.   :19.00   Max.   :19.20
       HE
 Min.   : 2.000
 1st Qu.: 3.500
 Median : 3.500
 Mean   : 5.053
 3rd Qu.: 7.500
 Max.   :10.000
>require(MASS)
> library(rpart) # rpart – Recursive PARTitioning - package base, para árvores de regressão # Criação da árvore maximal
```

>arM<-

rpart(MDM~.,data=dados,method='anova',control=rpart.control(minsplit=4,cp=0.0001

))

> summary(arM)

```
      NAF              TCE              RMAF             MDM              MRD
Min.   :2.000   Min.   : 8.00   Min.   : 500   Min.   : 8.00   Min.   :10.50
1st Qu.:3.000   1st Qu.: 8.00   1st Qu.:1250   1st Qu.:11.00   1st Qu.:14.00
Median :4.000   Median : 8.00   Median :1750   Median :14.00   Median :15.00
Mean   :3.694   Mean   :14.35   Mean   :1788   Mean   :14.04   Mean   :15.48
3rd Qu.:4.000   3rd Qu.:22.50   3rd Qu.:2500   3rd Qu.:16.00   3rd Qu.:17.00
Max.   :6.000   Max.   :70.00   Max.   :6000   Max.   :19.00   Max.   :19.20
       HE
Min.   : 2.000
1st Qu.: 3.500
Median : 3.500
Mean   : 5.053
3rd Qu.: 7.500
Max.   :10.000
```

>print(arM)

```
n= 85

node), split, n, deviance, yval
      * denotes terminal node

   1) root 85 823.8941000 14.03529
     2) MRD< 15.65 46 218.5924000 11.88043
       4) MRD< 13.75 17  23.5294100 10.20588
         8) MRD< 11.5 2   0.5000000  8.50000 *
         9) MRD>=11.5 15  16.4333300 10.43333
          18) RMAF< 2125 11  10.0454500 10.13636
            36) NAF>=3.5 4   0.6875000  9.62500
              72) HE< 5.5 2   0.1250000  9.25000 *
              73) HE>=5.5 2   0.0000000 10.00000 *
            37) NAF< 3.5 7   7.7142860 10.42857
              74) RMAF>=1000 5   6.8000000 10.20000
               148) RMAF< 1500 4   6.0000000 10.00000
                 296) HE>=2.75 2   4.5000000  9.50000 *
                 297) HE< 2.75 2   0.5000000 10.50000 *
               149) RMAF>=1500 1   0.0000000 11.00000 *
              75) RMAF< 1000 2   0.0000000 11.00000 *
          19) RMAF>=2125 4   2.7500000 11.25000
            38) TCE< 39 3   2.0000000 11.00000 *
            39) TCE>=39 1   0.0000000 12.00000 *
       5) MRD>=13.75 29 119.4483000 12.86207

                                    ...

        119) RMAF< 4250 6   0.8750000 18.25000
          238) RMAF< 1500 3   0.0000000 18.00000 *
          239) RMAF>=1500 3   0.5000000 18.50000 *
     15) MRD>=18.85 7   0.9285714 18.78571
       30) MRD< 19.05 5   0.8000000 18.70000
         60) HE< 5.5 2   0.5000000 18.50000 *
         61) HE>=5.5 3   0.1666667 18.83333 *
       31) MRD>=19.05 2   0.0000000 19.00000 *
```

Com 41 nós terminais a árvore é muito grande …

> summary(arM)

```
Call:
rpart(formula = MDM ~ ., data = dados, method = "anova", control = rpart.control(minsplit = 4,

    cp = 1e-04))
  n= 85

             CP nsplit  rel error    xerror       xstd
1  0.5650392273      0 1.00000000 1.0167687 0.10338771
2  0.0917772103      1 0.43496077 0.5263764 0.09049713
3  0.0854071555      2 0.34318356 0.4729713 0.08331713
4  0.0299729445      3 0.25777641 0.3758443 0.07738939
5  0.0263023138      5 0.19783052 0.3895626 0.07590496
6  0.0122183985      6 0.17152820 0.3407537 0.07246474
7  0.0091165976      8 0.14709141 0.3742397 0.06578644
8  0.0080059783      9 0.13797481 0.4509090 0.07521259
9  0.0079867872     10 0.12996883 0.4575382 0.07511801
10 0.0077924320     11 0.12198204 0.4560186 0.07521499 <<-- 1 SE
11 0.0073560497     12 0.11418961 0.4561905 0.07515976

12 0.0067514743     13 0.10683356 0.4547646 0.07521605
13 0.0053506066     15 0.09333061 0.4557179 0.07541832
14 0.0044154688     16 0.08798001 0.4543631 0.07666250
15 0.0040458273     17 0.08356454 0.4598230 0.07712630
16 0.0036513592     18 0.07951871 0.4610047 0.07710000
17 0.0034774849     19 0.07586735 0.4512188 0.07465053
18 0.0026334001     20 0.07238987 0.4504585 0.07466435
19 0.0021240593     21 0.06975647 0.4448220 0.07456949
20 0.0019950001     23 0.06550835 0.4353854 0.07449635      Mínimo
21 0.0019824554     24 0.06351335 0.4441851 0.07797294
22 0.0018206223     25 0.06153089 0.4483960 0.07797830
23 0.0013871408     26 0.05971027 0.4378517 0.07745874
24 0.0011097126     27 0.05832313 0.4412203 0.07736560
25 0.0010923734     28 0.05721342 0.4412203 0.07736560
26 0.0009103111     30 0.05502867 0.4434287 0.07734806
27 0.0006827334     31 0.05411836 0.4409455 0.07740366
28 0.0005852000     32 0.05343563 0.4428277 0.07742250
29 0.0005689445     33 0.05285043 0.4418628 0.07746071
30 0.0004551556     34 0.05228148 0.4422923 0.07743815
31 0.0003792963     35 0.05182633 0.4437252 0.07749906
32 0.0002275778     36 0.05144703 0.4434682 0.07752569
33 0.0001589432     38 0.05099187 0.4432786 0.07752056
34 0.0001000000     40 0.05067399 0.4425124 0.07747918
```

As linhas da tabela supra – tabela de complexidade - são as 34 árvores geradas pela função ***rpart***.

Na coluna *rel error* temos as estimativas medidas em função da árvore inicial. A coluna *xerror* são as estimativas na mesma medida mas obtidas por validação cruzada com 10 partições – no caso da função ***rpart***. A *xstd* mostra o erro estimado desta estimativa.

Olhando para a tabela de complexidade, vemos que a árvore 20 – com 23 divisões e 24 nós terminais – é a que tem uma menor taxa de erro: 0.4353854 – não observandos as iniciais pois essas têm poucas divisões e o "poder explicativo" seria reduzido - mas, outras há que não obstante terem menos divisões contêm um desvio padrão menor. Pela regra 1-SE (ver secção 2.2.5 e 2.2.6 ou Breiman *et al.* (1983, p.

237)), `0.4560186 < 0.4353854 + 0.07449635,` escolhendo desta forma uma árvore com 11 divisões e 12 nós terminais – linha 10. A escolha feita também se podia fazer observando o próximo gráfico - que nos dá os erros da validação cruzada versus valores de complexidade (***cp***).Assim para podar a árvore a escolha de ***cp*** recai entre os valores das linhas 9 e 10: `0.0077924320<` ***cp*** ≤ `0.0079867872,` ***cp*** = 0.0078

> plotcp(arM)

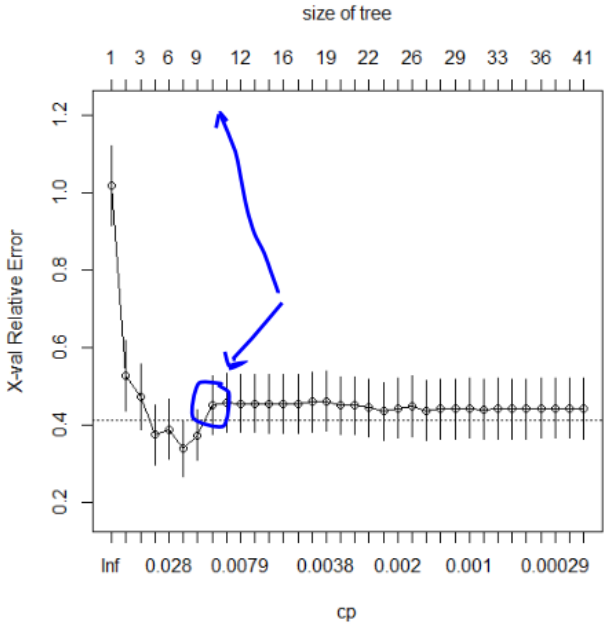



Figura 8 - Gráfico dos erros por validação cruzada vs valor de complexidade

**A poda da árvore:**

> arMp<-prune(arM,cp=0.0078)  # poda da árvore com12 nós terminais

> print(arMp)

```
n= 85

node), split, n, deviance, yval
      * denotes terminal node

 1) root 85 823.8941000 14.03529
   2) MRD< 15.65 46 218.5924000 11.88043
     4) MRD< 13.75 17  23.5294100 10.20588
       8) MRD< 11.5 2   0.5000000  8.50000 *
       9) MRD>=11.5 15  16.4333300 10.43333 *
     5) MRD>=13.75 29 119.4483000 12.86207
      10) NAF< 2.5 2   0.5000000  9.50000 *
      11) NAF>=2.5 27  94.6666700 13.11111
        22) HE< 2.75 8   9.8750000 11.62500 *
        23) HE>=2.75 19  59.6842100 13.73684
          46) MRD< 14.75 9   7.8888890 12.61111 *
          47) MRD>=14.75 10  30.1250000 14.75000 *
   3) MRD>=15.65 39 139.7692000 16.57692
     6) MRD< 17.25 21  48.1666700 15.33333
      12) NAF< 3.5 6  21.2083300 14.58333
        24) TCE>=15.25 1   0.0000000 11.00000 *
        25) TCE< 15.25 5   5.8000000 15.30000 *
      13) NAF>=3.5 15  22.2333300 15.63333
        26) MRD< 16.9 9   7.2222220 15.05556 *
        27) MRD>=16.9 6   7.5000000 16.50000 *
     7) MRD>=17.25 18  21.2361100 18.02778
      14) MRD< 18.85 11  13.7272700 17.54545 *
      15) MRD>=18.85 7   0.9285714 18.78571 *
```

>plot(arMp,uniform=T,branch=0.2)

>text(arMp,pretty=1,use.n=T)

> arMp.int <- snip.rpart(arMp)

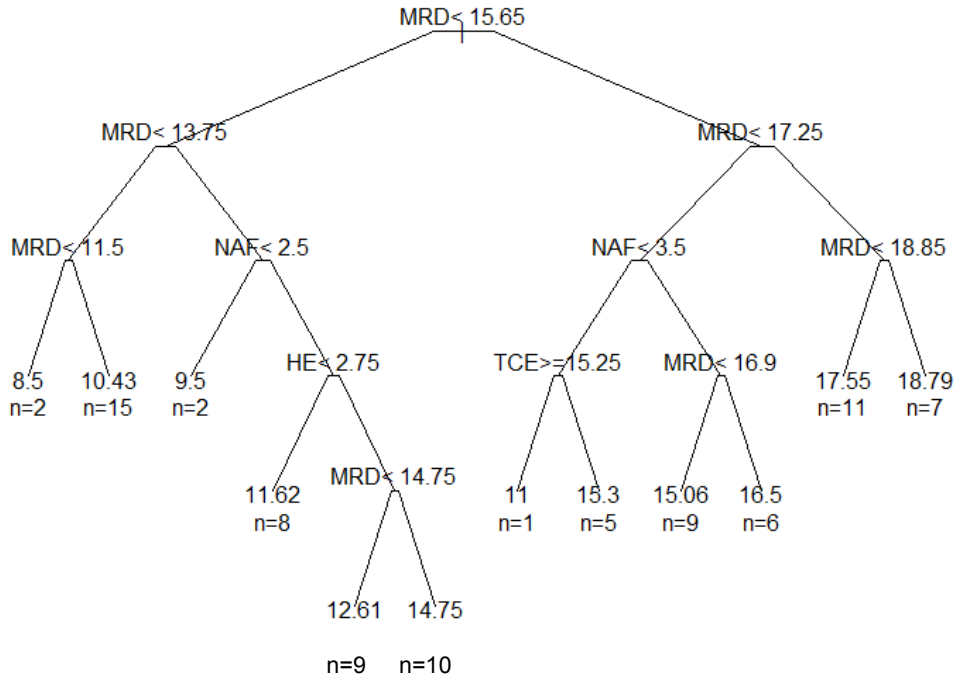



Gráfico 33 - Árvore final podada

>summary(arMp)

```
  n= 85

           CP nsplit rel error    xerror       xstd
1  0.565039227      0 1.0000000 1.0167687 0.10338771
2  0.091777210      1 0.4349608 0.5263764 0.09049713
3  0.085407155      2 0.3431836 0.4729713 0.08331713
4  0.029972945      3 0.2577764 0.3758443 0.07738939
5  0.026302314      5 0.1978305 0.3895626 0.07590496
6  0.012218399      6 0.1715282 0.3407537 0.07246474
7  0.009116598      8 0.1470914 0.3742397 0.06578644
8  0.008005978      9 0.1379748 0.4509090 0.07521259
9  0.007986787     10 0.1299688 0.4575382 0.07511801
10 0.007800000     11 0.1219820 0.4560186 0.07521499

Variable importance
 MRD   HE RMAF  NAF  TCE
  65   15   11    8    2
```

> tmp<-1-sum(pred!=dados$MDM)/nrow(dados) # taxa de má previsão

> tmp

[1] 0.01176471

Prever MDM para novos dados pela árvore ***arMp*** :

```
> prever<-predict(arMp,data.frame(NAF=4, TCE=2, RMAF=1000, MRD=14, HE=2))
> prever
      1
11.625
```

O que foi previsto pela função predict ( "prever " ) pode ser descrito pelo caminho que se indica na figura seguinte (figura 9):

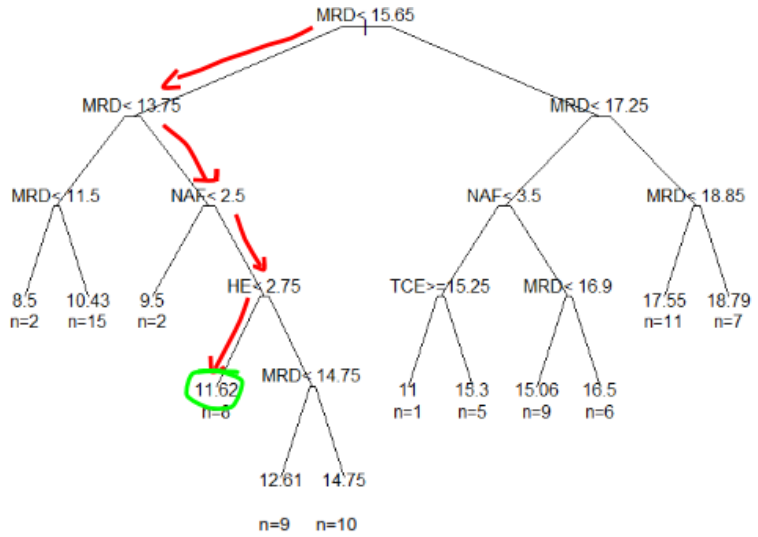



Figura 9 - Exemplo de caminho na árvore

Numa leitura mais detalhada – analisando e descrevendo a árvore de regressão (Gráfico 32 ou Figura10) – podemos fazer algumas considerações:

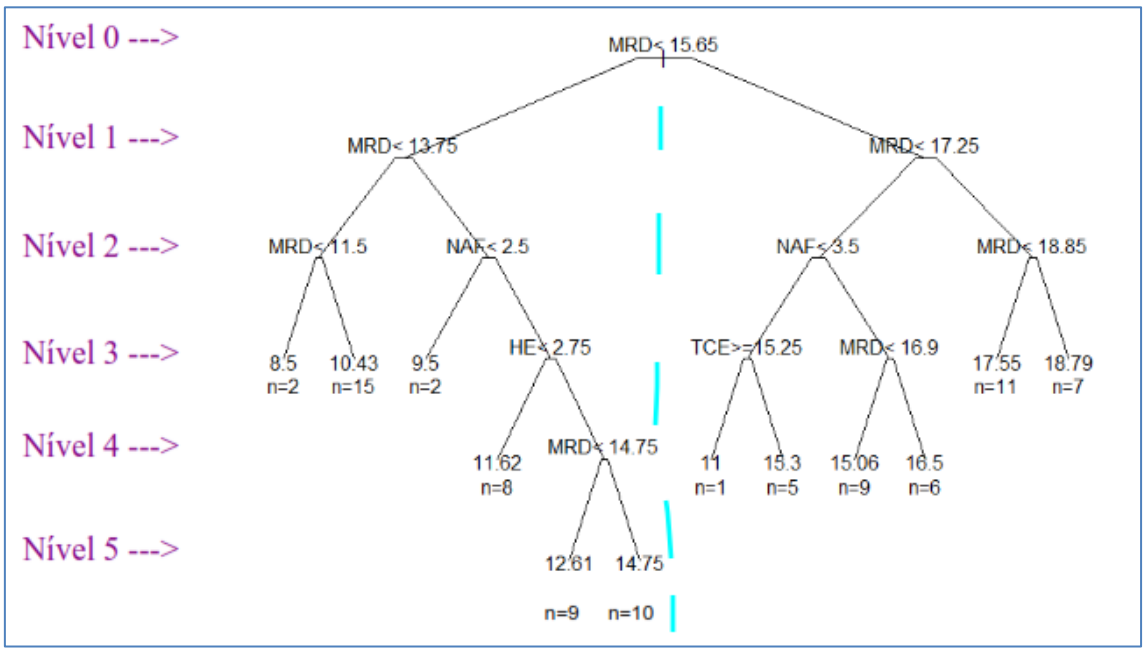



Figura 10 - Árvore final podada com níveis

A árvore é de complexidade diminuta – vantagem já referida - uma vez que apresenta 6 níveis e 12 nós terminais. A leitura far-se-á começando no "nível 0", onde se encontra a raiz da árvore (MRD < 15.65), e "sobe-se" ( a árvore está invertida) pelos ramos até ao próximo nó e assim – "recursivamente" – passando por nós intermédios chega-se aos nós terminais onde se encontram os valores da variável resposta. A saída de um nó para outro é através dos ramos: será pelo ramo da esquerda se se verificar a condição que está no nó ou pelo da direita no caso contrário. Ilustrando com o exemplo da previsão: - NAF=4, TCE=2, RMAF=1000, MRD=14, HE=2 – na raiz da árvore temos a condição MRD < 15.65 e nós queremos para MRD =14, então seguimos pelo ramo da esquerda uma vez que 14 é menor que 15.65 e chegamos ao nó com a condição MRD < 13.75, agora seguimos pelo ramo da direita pois 14 não é menor que 13.75 – resposta "NÃO" à condição MRD < 13.75 – e estamos no nó com a condição NAF < 2.5; se olharmos para a variável NAF – no exemplo - ela apresenta o valor 4 (NAF = 4), então o ramo a seguir será o da direita, tal como no anterior, porque a resposta foi "NÃO" á condição nesse nó e chegamos ao nó " HE < 2.75 " e daqui para o nó terminal onde MDM = 11.62. As RMAF e TCE ( com as condições RMAF = 1000 e TCE = 2) não apareceram neste caminho da raiz até ao nó terminal.

Na raiz da árvore encontra-se a variável **MRD** – Média obtida, nas restantes disciplinas, no 10º e 11º ano - que faz a partição entre os que obtiveram MRD **<** 15.65 ( lado esquerdo do tracejado vertical - partição da esquerda) e MRD **> =** 15.65 ( lado direito do tracejado vertical – partição da direita), sendo de prever para a variável dependente - **MDM**: Média obtida, na disciplina de Matemática, no 10º e 11º ano – tome valores entre 8.5 e 14.75 ( valores dos nós terminais da partição da esquerda) e valores entre 11 e 18.79 (valores dos nós terminais da partição da direita). Desta primeira partição que fizemos podemos concluir que há valores da variável dependente que podem ser obtidos por qualquer uma das duas partições - os valores resultantes da sua interseção: valores entre 11 e 14.75 – mas os valores entre 8.5 e 11 (excluído) só resultam da partição da esquerda ( MRD < 15.65), assim como os valores entre 14.75(excluído) e 18.79 só procedem da partição da direita (MRD > = 15.65). Nestas conclusões só teve influência a variável MRD – a mais influente nesta análise – mas as outras variáveis também influenciam a variável MDM.

Os extremos para MDM são 8.5(mínimo) de alunos que tiveram MRD < 11.5 e 18.79 (máximo) de MRD > = 18.85 ( Figura 11).

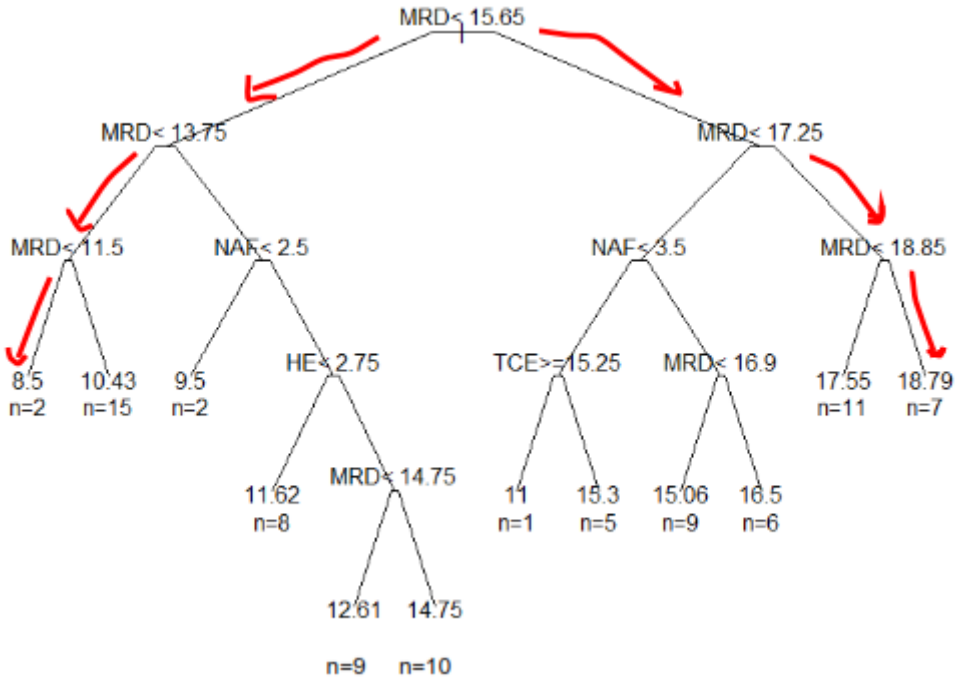



Figura 11 - Caminho dos extremos

As notas negativas – valores de MDM menores que 10 – são dadas por 2 caminhos: um – o vermelho ( da esquerda) - onde só a variável MRD tem influência e toma valores menores que 11.5 (ou seja, MRD < 11.5 que resulta de MRD < 15.65 e MRD < 13.75 e MRD < 11.5 ) e o outro – o azul ( da direita) - com MRD < 15.65 e MRD > = 13.75 e NAF < 2.5 ( ou seja 13.75 < = MRD < 15.65 e NAF < 2.5 ) ( Figura 12).

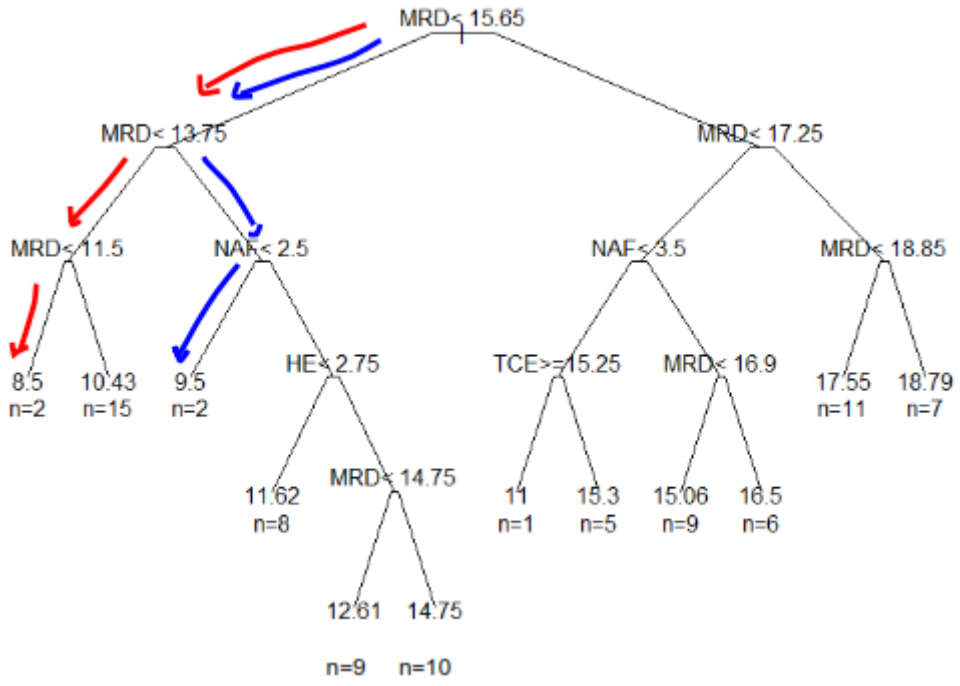



Figura 12 - Caminhos preditores de MDM < 10

Outras análises e descrições da árvore de regressão (Gráfico 32 ) poderiam ser feitas, mas a relação entre as variáveis preditoras e a variável resposta pode ser feita (Matematicamente) através de sentenças lógicas que fazem de cada um dos caminhos, que se considerem, regras ( ditas de predição ) associadas ao modelo do Gráfico 4.32. Neste caso concreto, 12 regras podem ser enumeradas(começando por predizer o nó terminal da esquerda até ao da direita):

**Regra 1**: **SE** ( MRD<15.65 e MRD<13.75 e MRD<11.5) **ENTÃO** a melhor predição para MDM é 8.5

ou seja

**SE** MRD <11.5 **ENTÃO** a melhor predição para MDM é 8.5

**Regra 2**: **SE** (( MRD<15.65 e MRD<13.75 e MRD > = 11.5) e MRD<11.5) **ENTÃO** a melhor predição para MDM é 10.43

Equivalente a

**SE** 11.5 < = MRD <13.75 **ENTÃO** a melhor predição para MDM é 10.43

**Regra 3**: **SE** (( 13.75 <= MRD<15.65) e NAF < 2.5) **ENTÃO** a melhor predição para MDM é 9.5

**Regra 4**: **SE** ( 13.75 <= MRD<15.65 e NAF > = 2.5 e HE < 2.75 ) **ENTÃO** a melhor predição para MDM é 11.62

**Regra 5**: **SE** ( 13.75 <= MRD<14.75 e NAF > = 2.5 e HE > = 2.75 ) **ENTÃO** a melhor predição para MDM é 12.61

**Regra 6**: **SE** ( 14.75 <= MRD<15.65 e NAF > = 2.5 e HE < 2.75 ) **ENTÃO** a melhor predição para MDM é 14.75

**Regra 7**: **SE** (15.65<= MRD<17.25 e NAF < 3.5 e TCE > = 15.25 ) **ENTÃO** a melhor predição para MDM é 11

**Regra 8**: **SE** (15.65<= MRD<17.25 e NAF < 3.5 e TCE < 15.25 ) **ENTÃO** a melhor predição para MDM é 15.3

**Regra 9**: **SE** (15.65<= MRD< 16.9 e NAF > = 3.5) **ENTÃO** a melhor predição para MDM é 15.06

**Regra 10**: **SE** (16.9 < = MRD < 17.25 e NAF > = 3.5) **ENTÃO** a melhor predição para MRD é 16.5

**Regra 11**: **SE** 17.25< = MRD < 18.85 **ENTÃO** a melhor predição para MDM é 17.55

**Regra 12**: **SE** MRD > = 18.85 **ENTÃO** a melhor predição para MDM é 18.79

( As regras 3 à 12 foram obtidas através de operações lógicas – conjunção ou disjunção de condições - e não por leitura simples da árvore – o que já tinha acontecido com as regras finais 1 e 2)

## 5. Conclusões e trabalho futuro

Este estudo teve a apresentação de dois métodos – duas técnicas – para modelar e analisar dados - recolhidos por questionário – fazendo uso da linguagem R, a qual se revelou bastante versátil e de fácil aplicabilidade, como era propósito demonstrar.

Com o primeiro método – Regressão Linear Múltipla – o processo não foi terminado uma vez que não houve lugar à validação do modelo encontrado. Tal não aconteceu porque não se verificaram os pressupostos da análise dos resíduos: linearidade da função de regressão; homocedasticidade (homogeneidade de variância dos erros);independência dos erros e normalidade dos erros. Não verificados, haveria importantes implicações relativamente à inferência proporcionada pelas estimativas do modelo. Algo mais se poderia fazer para encontrar um modelo válido mas seria à custa da "desvirtualização" dos dados. Poder-se-ia não contar com *outliers,* mas eles eram tantos que, a relação, variáveis preditoras e observações, seria muito pequena e assim teríamos poucas variáveis explicativas. Como tínhamos outro método para aplicar, resolvemos parar a análise uma vez que das tentativas feitas – e de estudos sobre a caracterização dos dados – só validaríamos uma Regressão Linear Simples com as vaiáveis MDM e MRD (Média obtida, na disciplina de Matemática, no 10º e 11º ano (variável dependente) e Média obtida, nas restantes disciplinas, no 10º e 11º ano), o que seria muito pouco para o que nos propusemos.

O segundo método - Árvores de Regressão Binárias – mostrou ter uma boa capacidade descritiva, preditiva e de fácil interpretação – ver árvore, Gráfico 4.32 – e assim atingir os objetivos deste estudo. A variável dependente - **MDM**: Média obtida, na disciplina de Matemática, no 10º e 11º ano – é explicada pelas variáveis independentes mencionadas, destacando ( da observação da árvore ) com as seguintes conclusões:

1. A variável RMAF - rendimento mensal líquido do agregado familiar – não se encontra como variável de partição na árvore, o que leva a concluir que a situação económica do aluno não tem " influência " no seu desempenho.

2. Quando olhamos para as partições geradas pela variável NAF - número de pessoas do agregado familiar – verificamos que a um maior valor da variável, leva a um

desempenho melhor e vice-versa; (alunos provenientes de famílias com maior número de pessoas têm "mais sucesso").

3. As variáveis MRD - média obtida, nas restantes disciplinas, no 10º e 11º ano – e HE – horas de estudo – fazem com que um crescimento dos seus valores conduz a valores maiores de MDM; já a variável TCE - tempo de casa à escola – influencia MDM em sentido oposto ao que foi dito; ou seja, alunos com boas médias às disciplinas que não Matemática, estudiosos e que gastam pouco tempo em deslocações, também tenham boas médias a Matemática.

Do que foi dito acerca da Regressão Linear Múltipla, não é de excluí-lo como método. No presente estudo não se "adaptou" aos dados, pelas razões já descritas, mas a amostra recolhida não era possível alterá-la, uma vez que coincidia com a população.

Futuramente, há que repensar o método a aplicar em função da amostra – população – e também o tipo de variáveis que elegemos para explicativas. No caso presente só nos cingimos às quantitativas. Outros aspetos, haverá, que podiam ter sido tratados e explorados, mas serão alvo de trabalho futuro.

**Referências Bibliográficas**


[1] Atkinson, E.J., Therneau, T.M. (1997):An introduction to Recursive Partitioning : Using the RPART Routines (long version) http://eric.univ-lyon2.fr/~ricco/cours/didacticiels/r/long_doc_rpart.pdf.

[2] Austin P.(2007): A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med. , 26 (15): 2937-2957.

[3] Barreto, Alexandre Serra (2011): Modelos de Regressão: Teoria e Aplicações com o Programa Estatístico R, Edição do Autor, Brasília.

[4] Breiman,L., Friedman,J.H., Olshen,R.A. & Stone,C.J. (1984): Classification and Regression Trees, Wadsworth Int. Group, Belmont, California, USA.

[5] Celeux G, Nakache J.P. (1994): Analyse discriminante sur variables qualitatives, Polytechnica,Paris.

[6] Chaudhuri et al.(1995): Generalized regression trees. Statist. Sinica. v5. 641-666

[7] Chen and Ahn, (1996) : Fitting mixed Poisson regression models using quasi-likelihood methods. Biometrical J. v38. 81-96.

[8] Choi, Y.,A. Hongshik, Chen, J.J.(2005): Regression trees for analysis of count data with extra Poisson variation. Computational Statistics & Data Analysis , 893-915 Elsevier Science Publishers

[9] Cornillon, Pierre-André (2010): Statistique avec R, PUR

[10] Cornillon, Pierre-André, Matzner-Lober, E. (2011): Régression avec R, Springer.

[11] Crawley, Michael J. (2007): The R book, 950p. Wiley.

[12] Draper,N.R., Smith,H. (1981) : Applied Regression Analysis, 2nd edition, John Wiley.

[13] Everitt BS and Hothorn T (2006): A Handbook of Statistical Analyses using R. Boca Raton, FL: Chapman and Hall/CRC.

[14] Ferreira, M. F. M. (2009): Árvores de regressão e generalizações: Aplicações. Universidade do Porto. Reitoria. Dissertação de Mestrado.

[15] Fonseca, Jaime (2001): Estatística Matemática, Vol. I e II, Edições Sílabo, Lisboa.

[16] Hill, Manuela Magalhães; Hill, Andrew (2009): Investigação por questionário,

Edições Sílabo, Lisboa.

[17] Husson, F., Lê, S., Pagès, J. (2009): Analyse de données avec R, PUR

[18] Kim, H. and Loh, W.-Y.: Classification trees with bivariate linear discriminant node models. Journal of Computational and Graphical Statistics, 12:512–530, 2003.

[19] Loh, W.-Y.(2002): Regression trees with unbiased variable selection and interaction detection.Statistica Sinica, 12:361–386.

[20] Loh, W-Y.(2008): Classification and Regression Tree Methods. In Encyclopedia of Statistics in Quality and Reliability, Ruggeri, Kenett and Faltin (eds.), 315–323, Wiley.

[21] Murteira, B., Carlos, C. S., Silva, j. A., Pimenta, C.(2007): Introdução à Estatística, 2ª Ed..,McGraw-Hill.

[22] Nakache J.P., Confais J., (2003): Statistique Explicative Appliquée, Technip,Paris.

[23] Oliveira, Teresa P. C. A.(2004): Estatística Aplicada, Universidade Aberta, Lisboa

[24] Pestana, M. Helena; Gageiro, J.N. (2005): Descobrindo a Regressão com a Complementaridade do SPSS, Edições Sílabo, Lisboa.

[25] Pestana, M. Helena; Gageiro, J.N. (2009): Análise Categórica, Árvores de Decisão e Análise de Conteúdo, Edições LIDEL, Lisboa

[26] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.Rproject.org/.

[27] Rocha, M., Cortez, P., Neves, J. M. (2008): Análise Inteligente de Dados: Algoritmos e Implementação em Java, FCA Editora, Lisboa.

[28] Torgo, Luís (2006) - Introdução à Programação em R, 99p., Faculdade de Economia, Universidade do Porto, Porto.

[29] Torgo, Luís (2009) - A linguagem R, Programação para a análise de dados, Escolar Editora, Lisboa.

[30] Torgo,Luís.(1999): Inductive Learning of Tree-based Regression Models. Dissertação de doutoramento disponível on-line: http://www.liaad.up.pt/~ltorgo/Ph

[31] Venables, W. N.; Smith, D. M. et al. (2010) – An Introduction to R.Disponível em: http://cran.r-project.org/doc/manuals/R-intro.pdf. Acedido a 20 de Janeiro

de 2011.

**Anexos**

**Anexo1 - Questionário**

# QUESTIONÁRIO

1.Escola_____

2.Sexo: Masculino ☐ Feminino ☐

3. Habilitações literárias dos teus pais :

| | Pai | Mãe |
|---|---|---|
| Analfabeto | ☐ | ☐ |
| 4º ano (1º Ciclo) | ☐ | ☐ |
| 6º ano (2º Ciclo) | ☐ | ☐ |
| 9º ano (3º Ciclo) | ☐ | ☐ |
| 12º ano ( ou Curso Profissional) | ☐ | ☐ |
| Curso Superior | ☐ | ☐ |

4. Situação laboral dos teus pais:

| | Pai | Mãe |
|---|---|---|
| Trabalha por conta própria | ☐ | ☐ |
| Trabalha por conta de outrem | ☐ | ☐ |
| Temporariamente desempregado | ☐ | ☐ |
| Reformado/pensionista | ☐ | ☐ |

5. Profissão do pai:_____

6. Profissão da mãe:_____

7. Estado civil dos pais?

| | |
|---|---|
| Casados | ☐ |
| União de facto | ☐ |
| Divorciados/Separados | ☐ |
| Viúvo/a | ☐ |
| Outra | ☐ |

8. Quem é o teu encarregado de educação?

| | |
|---|---|
| Mãe/Madrasta | ☐ |
| Pai/Padrasto | ☐ |
| O próprio | ☐ |
| Irmão/Irmã | ☐ |
| Outro | ☐ |

9. Número de pessoas do agregado familiar:_____

10.Pessoas com quem vives:

| | |
|---|---|
| Pais | ☐ |
| Pais e irmão(s) | ☐ |
| Só com Pai | ☐ |
| Só com Mãe | ☐ |
| Outra situação | ☐ |

11. Local de residência : Meio Urbano ☐ Meio Rural ☐

12. Quanto tempo demoras de casa à escola?

| | |
|---|---|
| Menos de 15 min. | ☐ |
| 15 a 30 min. | ☐ |
| 30 min. a 1 hora | ☐ |
| Mais de 1 hora | ☐ |

13. Como classificas as relações familiares em casa?

| Escala | 1 | 2 | 3 | 4 | 5 | Escala |
|---|---|---|---|---|---|---|
| Conflituosas | | | | | | Excelentes |

14. Rendimento mensal líquido do agregado familiar:

| | |
|---|---|
| Até 500 € | ☐ |
| 500 a 1000 € | ☐ |
| 1000 a 1500 € | ☐ |
| 1500 a 2000 € | ☐ |
| 2000 a 3000 € | ☐ |
| 3000 a 5000 € | ☐ |
| Mais de 5000 € | ☐ |

15.Quantas vezes já reprovaste?

| 0 | 1 | 2 | 3 | 4 | Mais de 4 |
|---|---|---|---|---|---|
| | | | | | |

16. Média obtida, na disciplina de Matemática, no 10º e 11º ano: ____

17. Média obtida, nas restantes disciplinas, no 10º e 11º ano: ____

18.Quantas horas estudas (por semana)?

Menos de 2 horas ☐
2 a 5 horas ☐
5 a 10 horas ☐
Mais de 10 horas ☐

19.Tens Computador em casa? Sim ☐ Não ☐

20. Tens Internet em casa? Sim ☐ Não ☐

21. Onde estudas habitualmente? *( marcar só uma)*

Em casa ☐
Em casa de amigos ☐
Na escola ☐
Noutro local ☐

22. Fazes os TPC? *( marcar só uma)*

Sempre ☐
Muitas vezes ☐
Às vezes ☐
Nunca ☐

23. Tens ajuda nos TPC?

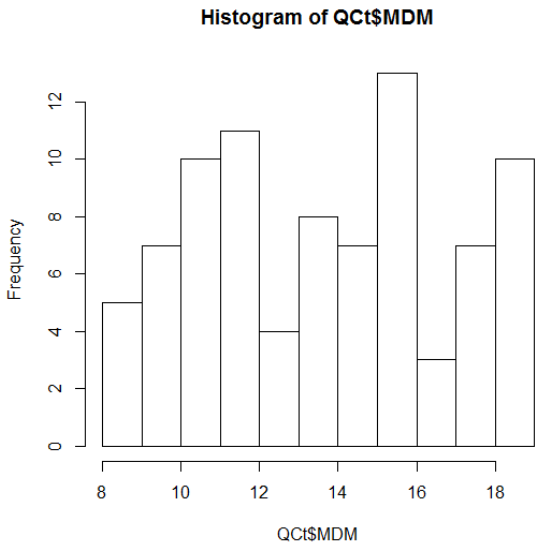
Sim ☐
Não ☐
Às vezes ☐

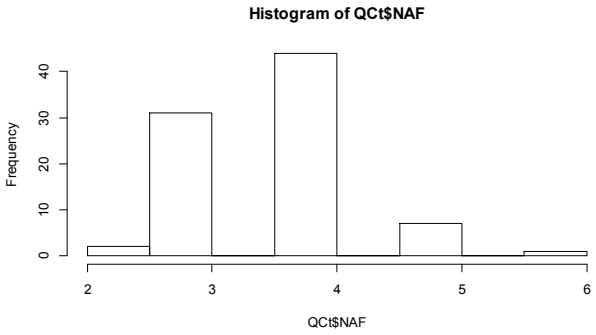
24. Usas recursos informáticos nas aulas de Matemática?

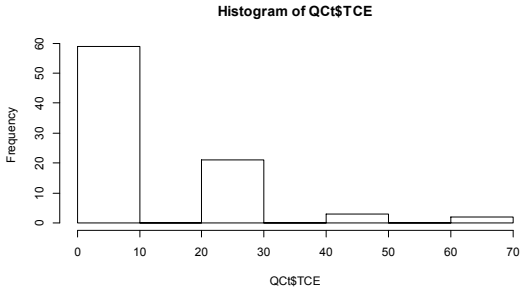
Nunca ☐
Às vezes ☐
Frequentemente ☐

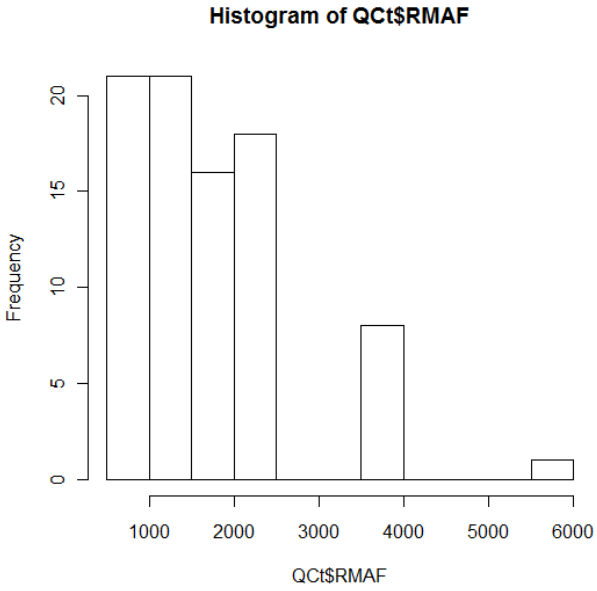
25. Desejas frequentar um curso superior? Sim ☐ Não ☐

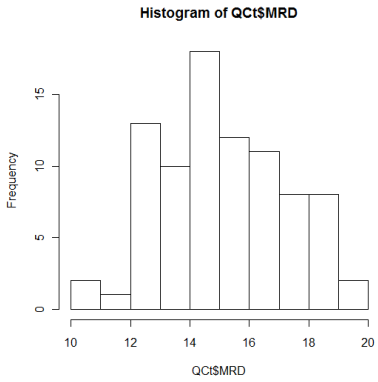
## Anexo2 – Histogramas

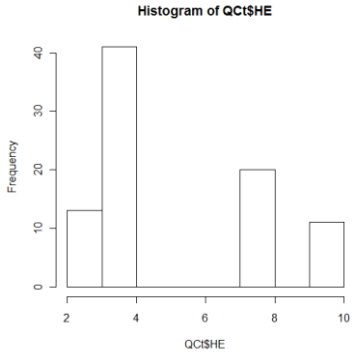


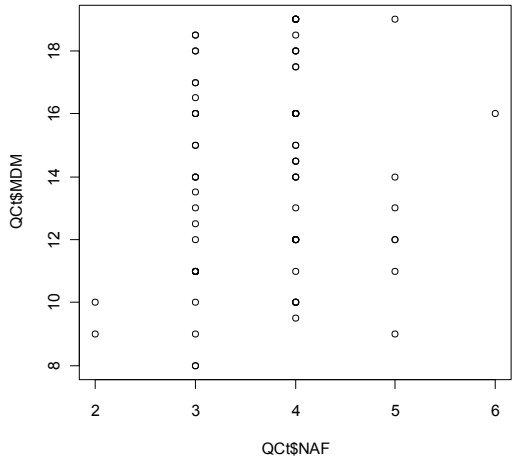



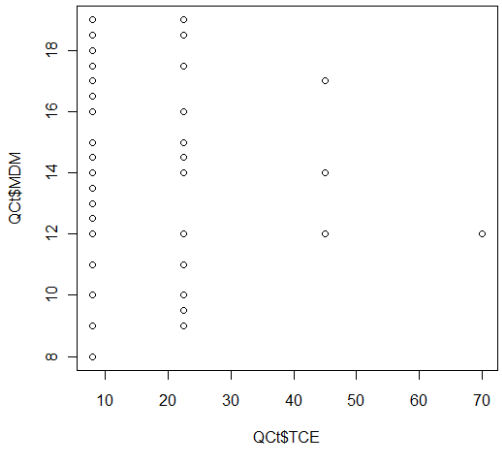



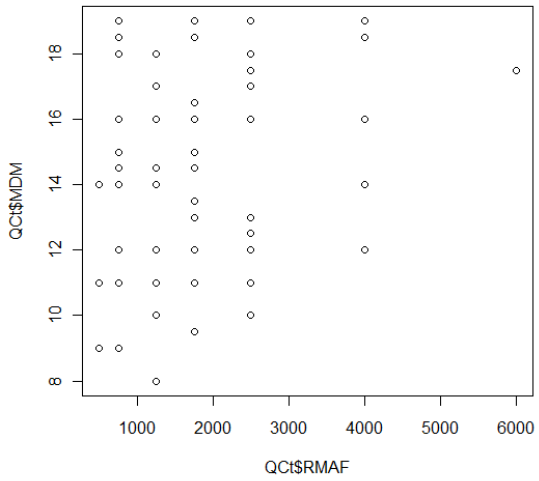



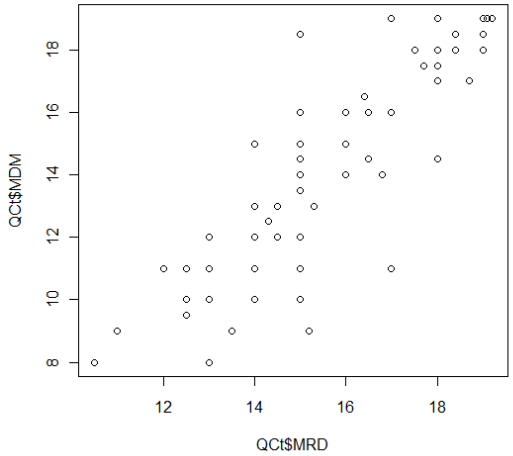



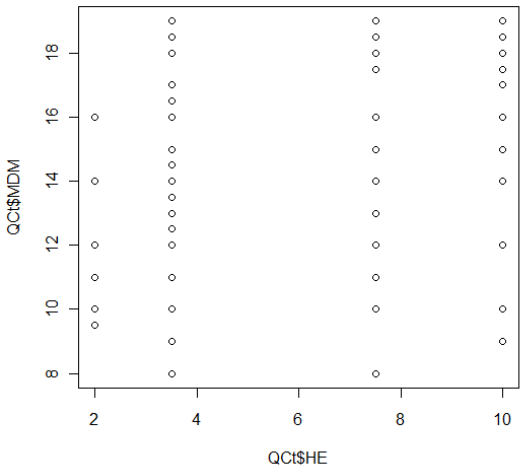
## Anexo3 - Diagrama de dispersão para MDM com as outras variáveis




QCt$MDM
8
10
12
14
16
18
2
4
6
8
10
QCt$HE