UNIVERSIDADE DA BEIRA INTERIOR
Ciências

Modelo de Regressão Linear e suas Aplicações

Sandra Cristina Antunes Rodrigues

Relatório de Estágio para obtenção do Grau de Mestre em
Ensino de Matemática no 3º Ciclo do Ensino Básico e no Ensino Secundário
(2º ciclo de estudos)

Orientadora: Professora Doutora Célia Maria Pinto Nunes

Covilhã, Outubro de 2012

Dedicatória

Aos meus Filhos, João Pedro e Henrique Miguel.

Agradecimentos

Em primeiro lugar, quero agradecer a todos os que contribuíram para que fosse possível realizar este trabalho.

À minha orientadora, Professora Doutora Célia Maria Pinto Nunes, pela força e motivação, pelas suas orientações e apoio, e sobretudo pela sua amizade.

Um agradecimento especial à colega Sónia Ladeira pelas suas ideias e sugestões e à querida amiga Sílvia Melchior pela sua ajuda no Inglês.

À minha cunhada, Dra. Catarina Reis, que gentilmente me cedeu uma base de dados para uso nas aplicações.

À minha querida irmã, pela sua paciência infindável e carinho imensurável.

À minha família que sempre me apoiou e incentivou, em especial aos meus filhos, pelo tempo que não lhes dediquei, aos meus pais, pilares da minha vida, pelo permanente incentivo e pela formação que me permitiram adquirir, aos meus sogros pela ajuda incondicional.

Ao meu marido, quem mais sofreu com as minhas indisponibilidades e impaciências, pelo seu carinho, companheirismo e compreensão…

A todos, o meu Bem-Haja!

Resumo

O presente trabalho teve como principal objectivo apresentar os resultados mais importantes sobre os modelos de regressão linear, ilustrando a sua aplicabilidade através de estudos que foram elaborados com base em dados reais.

Como tal abordámos a análise de regressão linear simples e descrevemos sumariamente a regressão linear múltipla, que se distingue da anterior quando incorporadas mais do que uma variável independente no modelo de regressão.

Enquadrado na temática anterior, dedicámos um capítulo a Análise de resíduos.

Por último, e como complemento da investigação realizada ao longo deste trabalho, realizámos alguns estudos aplicados a dados reais, que dizem respeito à Variabilidade da Frequência Cardíaca.

Palavras-chave

Regressão Linear Simples, Regressão Linear Múltipla, Estimação dos Parâmetros, Aplicações a Dados Reais.

Abstract

This study's main objective was to present the most important results about the linear regression models, illustrating its applicability through studies that were prepared based on real data.

Therefore we touched the simple linear regression analysis and briefly describe the multiple linear regression, distinct from the previous embedded when more than one independent variable in the regress.

Framed in the previous issue, we devoted a chapter to Waste Analysis ion model.

Finally, and as a complement of research carried throughout this work we held some studies applied to real data, which concern the Heart Rate Variability.

Keywords

Simple Linear Regression, Multiple Linear Regression, Parameter Estimation, Applications to Real Data.

Índice

1. Introdução 1

2. Análise de Regressão simples 5

2.1. Modelo teórico 5

2.2. Pressupostos do modelo 6

2.3. Estimação dos parâmetros do Modelo 7

2.3.1. Método dos mínimos quadrados 7

2.3.2. Propriedades dos Estimadores 11

2.4. Estimador de σ^2 17

2.5. Testes e intervalos de confiança para os parâmetros do modelo 19

2.5.1. Testes e intervalos de confiança para β_1 20

2.5.2. Testes e intervalos de confiança para β_0 21

3. Breve abordagem à regressão linear múltipla 23

3.1. Modelo teórico e seus pressupostos 23

3.1.1. Interações 24

3.1.2. Pressupostos do modelo 25

3.1.3. Representação matricial do método de regressão linear múltipla 25

3.2. Estimação do parâmetro do modelo 26

3.2.1. Propriedades dos estimadores 27

3.3. Estimador de σ^2 28

3.4. Análise da Variância 30

4. Análise de Resíduos 33

4.1. Diagnóstico de normalidade 33

4.2. Diagnóstico de Homoscedasticidade (variância constante) 35

4.3. Diagnóstico de Independência 36

4.4. Diagnóstico de *Outliers* e observações influentes 37

4.4.1. Observações Influentes 38

4.5. Colinearidade e Multicolinearidade 39

5. Aplicações 43

5.1. Estudo 1 – Modelo de regressão Linear Simples 43

5.1.1. Verificação dos pressupostos do modelo 46

5.2. Estudo 2 – Modelo de Regressão linear Múltipla 49

5.2.1. Verificação dos pressupostos do modelo 51

6. Conclusões 57

Bibliografia 59

Anexos 61

Lista de Figuras

Figura 1.1- Classificação da correlação através do diagrama de dispersão, disponível em Santos (2007). 3

Figura 2.1- Interpretação geométrica dos parâmetros $\beta_0 \, e \, \beta_1$ 6

Figura 2.2- Representação gráfica dos resíduos 8

Figura 3.1- Hiperplano p-dimensional referente às variáveis explicativas. 24

Figura 4.1– Normal p-p plot de resíduos 34

Figura 4.2- Confirmação da homoscedasticidade dos resíduos (disponível em PortalAction). . 36

Figura 5.1– Diagrama de dispersão 44

Figura 5.2– Normal p-p plot 46

Figura 5.3– Gráfico dos resíduos estandardizados 47

Figura 5.4 – Gráfico resíduos *press* 47

Figura 5.5– Gráfico dos *Standardized* DFFIT 48

Figura 5.6– Normal p-p plot da regressão dos resíduos estandardizados 52

Figura 5.7- Gráfico dos resíduos estandardizados 53

Figura 5.8– Gráfico resíduos *press* 54

Figura 5.9– Gráfico dos *Standardized* DFFIT 55

Lista de Tabelas

Tabela 1.1- Interpretação do coeficiente de correlação de Pearson. 2

Tabela 3.1- Tabela da análise de variância (ANOVA) 31

Tabela 4.1- Tabela de decisão em função de d_U e d_L 37

Tabela 5.1- Estatística descritiva 43

Tabela 5.2– Sumário do Modelo 44

Tabela 5.3– Tabela da ANOVA 45

Tabela 5.4- Coeficientes 45

Tabela 5.5- Teste K-S 46

Tabela 5.6- Estatística dos resíduos 48

Tabela 5.7- Estatística descritiva 49

Tabela 5.8– Tabela de variáveis inseridas/removidas 49

Tabela 5.9- Sumário do Modelo 50

Tabela 5.10– Tabela da ANOVA 50

Tabela 5.11- Coeficientes 50

Tabela 5.12– Variáveis excluídas 51

Tabela 5.13– Teste K-S 52

Tabela 5.14- Diagnóstico da colinearidade 53

Tabela 5.15- Estatística dos resíduos 54

1. INTRODUÇÃO

"O termo 'regressão' foi proposto pela primeira vez por Sir Francis Galton em 1885 num estudo onde demonstrou que a altura dos filhos não tende a reflectir a altura dos pais, mas tende sim a regredir para a média da população. Actualmente, o termo "Análise de Regressão" define um conjunto vasto de técnicas estatísticas usadas para modelar relações entre variáveis e predizer o valor de uma ou mais variáveis dependentes (ou de resposta) a partir de um conjunto de variáveis independentes (ou predictoras)." (Maroco, 2003)

A temática deste trabalho será a análise de regressão linear, no entanto, faremos de seguida uma pequena abordagem ao coeficiente de correlação e consequentemente ao coeficiente de determinação.

A análise de correlação tem como objectivo a avaliação do grau de associação entre duas variáveis, X e Y, ou seja, mede a "força" de relacionamento linear entre as variáveis X e Y.

Para quantificar a relação entre duas variáveis quantitativas utiliza-se o coeficiente de correlação linear de Pearson.

O coeficiente de correlação linear de Pearson entre duas variáveis quantitativas, X e Y, é dado por:

$$R_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

onde

$$\bar{x} = \sum_{i=1}^{n}\frac{x_i}{n} \; e \; \bar{y} = \sum_{i=1}^{n}\frac{y_i}{n},$$

ou seja, é o quociente entre a covariância entre X e Y e o produto de desvios padrão de X e Y.

A partir de R_{xy} podemos tirar conclusões sobre a direcção e intensidade da relação existente entre as variáveis X e Y. Não existe uma "classificação" unânime da correlação. Nós optámos por seguir a considerada por Santos (2007) que é a apresentada na Tabela 1.1.

Tabela 1.1- Interpretação do coeficiente de correlação de Pearson.

Coeficiente de correlação	Correlação
$R_{xy} = 1$	Perfeita positiva
$0{,}8 \leq R_{xy} < 1$	Forte positiva
$0{,}5 \leq R_{xy} < 0{,}8$	Moderada positiva
$0{,}1 \leq R_{xy} < 0{,}5$	Fraca positiva
$0 \leq R_{xy} < 0{,}1$	Ínfima positiva
0	Nula
$-0{,}1 \leq R_{xy} < 0$	Ínfima negativa
$-0{,}5 \leq R_{xy} < -0{,}1$	Fraca negativa
$-0{,}8 \leq R_{xy} < -0{,}5$	Moderada negativa
$-1 \leq R_{xy} < -0{,}8$	Forte negativa
$R_{xy} = -1$	Perfeita negativa

Para investigar a relação entre duas variáveis, X e Y, podemos representar os valores das variáveis num gráfico de dispersão. Afirma-se que existe uma relação linear entre as variáveis se os dados se aproximarem de uma linha recta.

A partir da observação do diagrama de dispersão verificamos se a correlação entre as duas variáveis é mais ou menos forte, de acordo com a proximidade dos pontos em relação a uma recta. Na Figura 1.1, podemos observar alguns exemplos de gráficos de dispersão e a respectiva "classificação" da correlação.

Dependendo da relação entre as variáveis e da intensidade com que se relacionam, a recta obtida será um melhor ou pior modelo para traduzir a relação entre elas.

De seguida iremos definir o coeficiente de determinação, que é igual ao quadrado do coeficiente de correlação de Pearson.

Como vimos, o coeficiente de correlação linear de Pearson entre duas variáveis serve para medir a intensidade da relação linear entre elas. O coeficiente de determinação é mais indicado para medir a explicação da recta de regressão. Assim, quanto mais próximo de 1 estiver o valor do coeficiente de determinação, maior a percentagem da variação de Y explicada pela recta estimada, e por conseguinte, maior a qualidade do ajustamento.

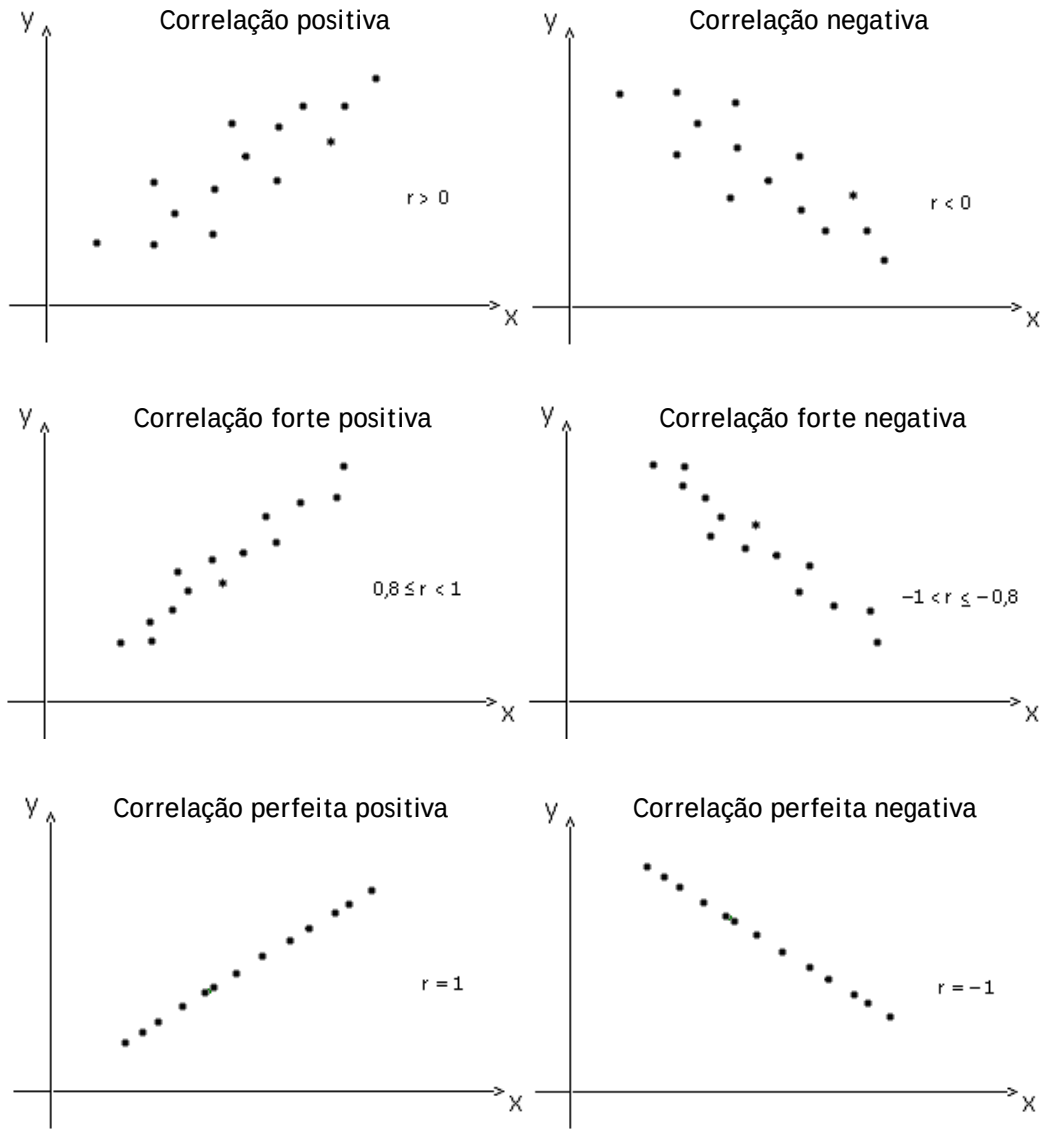

Figura 1.1- Classificação da correlação através do diagrama de dispersão, disponível em Santos (2007).

O coeficiente de determinação é dado por

$$R_{xy}^2 = \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

O R_{xy}^2 toma valores entre zero e um. A qualidade do ajuste será tanto maior quanto mais R_{xy}^2 se aproximar de 1.

Em resumo, a presença ou ausência de relação linear pode ser averiguada a partir de dois pontos distintos:

a) quantificando a força dessa relação, e para isso usamos a análise de correlação;
b) ou explicitando a forma dessa relação, fazendo uso da análise de regressão.

Ambas as técnicas, apesar de intimamente ligadas, diferem, pois na correlação todas as variáveis são aleatórias e desempenham o mesmo papel, não havendo nenhuma dependência, enquanto na regressão isso não acontece.

Assim, a análise de regressão estuda o relacionamento entre uma variável denominada de dependente, Y , e uma ou várias variáveis independentes, $X, (X_1, X_2, \dots, X_p)$. Caso se considere apenas uma variável independente apelidamos de análise de regressão simples, caso usemos duas ou mais variáveis, de análise de regressão múltipla.

A importância do estudo da análise de regressão advém da necessidade do estudo de determinados fenómenos nas Ciências da Natureza (Física, Biologia, Química, …), nas Ciências Sociais, nas Ciências da Saúde, …

Ainda que operacionalmente simples, existem certos aspectos do uso da regressão linear que merecem uma discussão adicional e sobre os quais nos debruçaremos neste trabalho.

Assim, no capítulo 2 debruçamo-nos sobre a regressão linear simples e apresentamos o modelo teórico. São discutidos temas como os parâmetros do modelo, as propriedades dos estimadores e inferência dos parâmetros.

É feita uma breve abordagem sobre a análise de regressão linear múltipla no capítulo 3. Neste capítulo, é apresentado o modelo teórico e os seus pressupostos. É ainda feita referência à análise de variância, de extrema importância para a regressão linear múltipla.

No capítulo 4 é feita a análise de resíduos, onde são apresentados alguns dos métodos existentes para verificação dos pressupostos.

No capítulo 5 são apresentadas algumas aplicações a dados reais recorrendo ao uso do SPSS (*Statistical Package for the Social Sciences*, versão 19), exemplificando algumas técnicas descritas no trabalho.

Por último, no capítulo 6, são apresentadas algumas conclusões.

2. ANÁLISE DE REGRESSÃO SIMPLES

A análise de regressão linear estuda a relação entre a variável dependente ou variável resposta (Y) e uma ou várias variáveis independentes ou regressoras $(X_1, \dots, X_p)$.

Esta relação representa-se por meio de um modelo matemático, ou seja, por uma equação que associa a variável dependente (Y) com as variáveis independentes $(X_1, \dots, X_p)$.

O Modelo de Regressão Linear Simples define-se como a relação linear entre a variável dependente (Y) e uma variável independente (X).

Enquanto que o Modelo de Regressão Linear Múltiplo define-se como a relação linear entre a variável dependente (Y) e várias variáveis independentes $(X_1, \dots, X_p)$.

Neste capítulo vamos apenas debruçar-nos sobre o modelo de regressão linear simples. Será apresentado o modelo teórico e os seus pressupostos, assim como a estimação dos parâmetros do modelo pelo método dos mínimos quadrados. Serão ainda construídos testes e intervalos de confiança para os parâmetros do modelo.

2.1. MODELO TEÓRICO

A equação representativa do modelo de regressão linear simples é dado por:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \qquad i = 1, \dots, n \qquad (2.1)$$

onde:

. y_i representa o valor da variável resposta ou dependente, Y, na observação i , $i = 1, \dots, n$ (aleatória);

. x_i representa o valor da variável independente, X, na observação i, $i = 1, \dots, n$ (não aleatória);

. ε_i , $i = 1, \dots, n$ são variáveis aleatórias que correspondem ao erro (variável que permite explicar a variabilidade existente em Y e que não é explicada por X);

. β_0 e β_1 correspondem aos parâmetros do modelo.

O parâmetro β_0 representa o ponto em que a recta regressora corta o eixo dos yy quando $X = 0$ e é chamado de intercepto ou coeficiente linear.

O parâmetro β_1 representa a inclinação da recta regressora, expressando a taxa de mudança em Y, ou seja, indica a mudança na média da distribuição de probabilidade de Y para um aumento de uma unidade na variável X.

Na Figura 2.1 podemos observar a interpretação geométrica dos parâmetros β_0 e β_1.

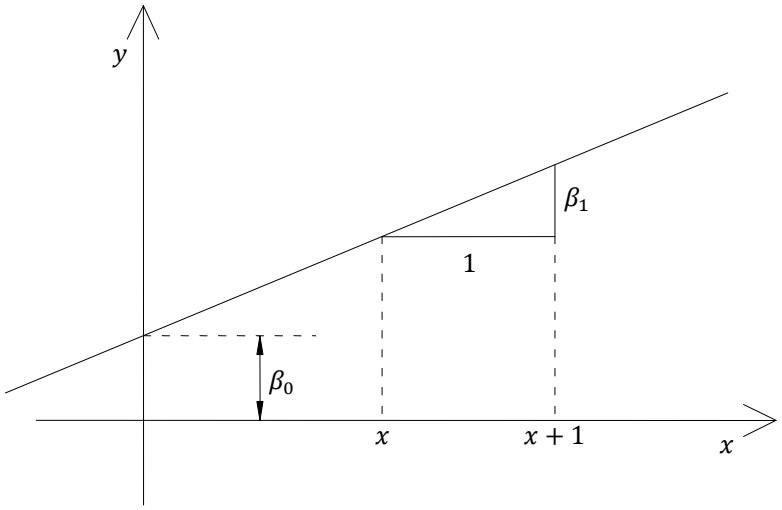

Figura 2.1- Interpretação geométrica dos parâmetros $\boldsymbol{\beta_0}$ *e* $\boldsymbol{\beta_1}$

2.2. PRESSUPOSTOS DO MODELO

Ao definir o modelo (2.1) estamos a pressupor que:

a) A relação existente entre Y e X é linear.
b) Os erros são independentes com média nula.

Pressupondo então que $E(\varepsilon_i) = 0$, tem-se:

$$E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$= \beta_0 + \beta_1 x_i + E(\varepsilon_i)$$

$$= \beta_0 + \beta_1 x_i. \qquad (2.2)$$

Por outro lado, podemos afirmar que o erro de uma observação é independente do erro de outra observação, o que significa que:

$$cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) - E(\varepsilon_i)E(\varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0, \quad \text{para } i \neq j, \quad i, j = 1, \dots, n\,.$$

c) A variância do erro é constante, isto é $var(\varepsilon_i) = \sigma^2, \ \ i = 1, \dots, n.$

Tem-se então

$$var(\varepsilon_i) = E(\varepsilon_i^2) - \underbrace{[E(\varepsilon_i)]^2}_{=0} = E(\varepsilon_i^2) = \sigma^2,$$

e consequentemente

$$var(y_i) = var(\beta_0 + \beta_1 x_i + \varepsilon_i) = \overbrace{var(\underbrace{\beta_0 + \beta_1 x_i}_{\text{termo constante}})}^{=0} + \underbrace{var(\varepsilon_i)}_{=\sigma^2} = \sigma^2.$$

d) Os erros, $\varepsilon_i,\ i = 1, \dots, n$, são normalmente distribuídos.

Concluímos portanto, de b) e c), que

$$\varepsilon_i \sim N(0, \sigma^2), \qquad i = 1, \dots n\,,$$

e portanto que

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \qquad i = 1, \dots, n\,.$$

2.3. ESTIMAÇÃO DOS PARÂMETROS DO MODELO

Supondo que existe efectivamente uma relação linear entre X e Y, coloca-se a questão de como estimar os parâmetros β_0 e β_1.

Karl Gauss entre 1777 e 1855 propôs estimar os parâmetros β_0 e β_1 visando minimizar a soma dos quadrados dos desvios, $e_i,\ i = 1, \dots, n$, chamando este processo de método dos mínimos quadrados. Este método será descrito de seguida. (Maroco, 2003)

2.3.1. Método dos mínimos quadrados

O método dos mínimos quadrados consiste na obtenção dos estimadores dos coeficientes de regressão $\hat{\beta}_0$ e $\hat{\beta}_1$, minimizando os resíduos do modelo de regressão linear, calculados como a diferença entre os valores observados, y_i, e os valores estimados, $\hat{y}_i$, isto é

$$e_i = y_i - \hat{y}_i, \qquad i = 1, \dots, n\,.$$

Em termos gráficos, os resíduos são representados pelas distâncias verticais entre os valores observados e os valores ajustados, como mostra a Figura **2.2**.

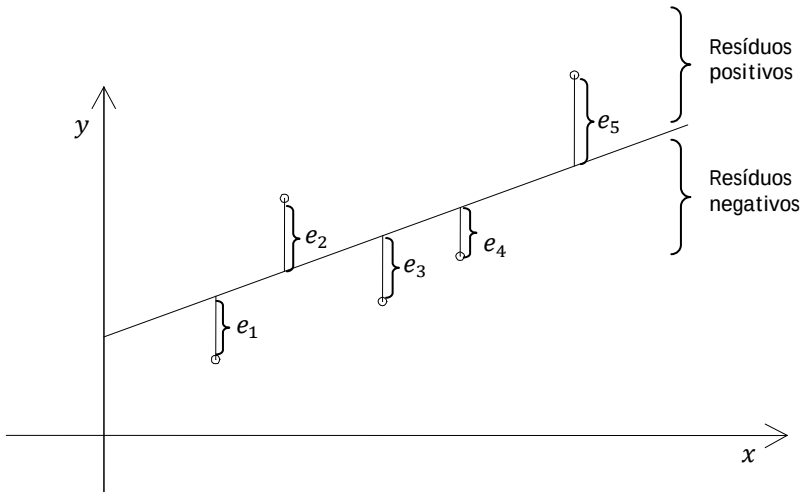

Figura 2.2- Representação gráfica dos resíduos

O método dos mínimos quadrados propõe então encontrar os valores de β_0 e β_1 para os quais a soma dos quadrados dos resíduos (SQE) é mínima. Tem-se então:

$$SQE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 \quad (2.3)$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 ,$$

com $\sum_{i=1}^{n} e_i = 0$ (daí o facto de ser considerado o quadrado de $e_i,\ i = 1, \dots, n$).

Precisamos agora de calcular as derivadas parciais de SQE em ordem a β_0 e β_1, obtendo-se:

$$\begin{cases} \dfrac{\partial SQE}{\partial \beta_0} = -2 \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \dfrac{\partial SQE}{\partial \beta_1} = -2 \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i \end{cases} .$$

Igualando estas derivadas a zero e substituindo β_0 e β_1 por $\hat{\beta}_0$ e $\hat{\beta}_1$, por forma a indicar valores concretos destes parâmetros, tem-se

$$\begin{cases} -2 \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ -2 \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \end{cases} \qquad (2.4)$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_0 = \dfrac{\sum_{i=1}^{n} y_i}{n} - \dfrac{\hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \\ \text{——} \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \text{——} \end{cases}, \qquad (2.5)$$

em que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, representam as médias de X e Y, respectivamente.

Vamos agora pegar na 2ª equação de (2.4) e tentar chegar à expressão de $\hat{\beta}_1$. Ora

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i$$

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

Como

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + n\hat{\beta}_1 \bar{x}^2 \, ,$$

vem

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + n\hat{\beta}_1 \bar{x}^2$$

$$\Leftrightarrow \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \Leftrightarrow$$

$$\Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \, . \qquad (2.6)$$

$\hat{\beta}_0$ e $\hat{\beta}_1$, anteriormente determinados em (2.5) e (2.6), são designados como os Estimadores de Mínimos Quadrados de β_0 e β_1.

De seguida serão apresentadas algumas propriedades do ajuste dos mínimos quadrados.

- Como vimos, os resíduos correspondem à diferença entre os valores observados, $y_i,\ i = 1, \dots, n$, e os correspondentes valores ajustados, $\hat{y}_i,\ i = 1, \dots, n$, isto é:

$$e_i = y_i - \hat{y}_i$$

$$= y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right),$$

com

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, \dots, n\,.$$

a) $\sum_{i=1}^{n} e_i = 0$, o que significa que a soma dos resíduos é sempre nula;
b) $\sum_{i=1}^{n} e_i^2$ é mínima;
c) $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$, o que significa que a soma dos valores observados y_i é igual à soma dos valores ajustados $\hat{y}_i$;
d) A recta obtida pelo método dos mínimos quadrados passa sempre pelo ponto $(\bar{x}, \bar{y})$.

Demonstração: Como

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$= \beta_0 + \beta_1 x_i + \varepsilon_i + \beta_1 \bar{x} - \beta_1 \bar{x}$$

$$= \beta_0 + \beta_1 (x_i - \bar{x}) + \beta_1 \bar{x} + \varepsilon_i$$

$$= (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x_i - \bar{x}) + \varepsilon_i$$

$$= {\beta_0}^0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i\,,$$

com

$${\beta_0}^0 = \beta_0 + \beta_1 \bar{x}\,.$$

Logo, os valores ajustados serão dados por:

$$\hat{y}_i = \hat{\beta}_0^0 + \hat{\beta}_1 (x_i - \bar{x})$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} + \hat{\beta}_1 (x_i - \bar{x})$$

$$= \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 \bar{x} + \hat{\beta}_1 (x_i - \bar{x})$$

$$= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \ ,$$

visto que

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \ .$$

Assim, no ponto de abcissa $\bar{x}$, vem

$$\hat{y}_i = \bar{y} + \hat{\beta}_1 (\bar{x} - \bar{x}) = \bar{y} \ .$$

2.3.2. Propriedades dos Estimadores

a) Valor esperado e variância de $\hat{\beta}_1$

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

Como vimos, de (2.6), tem-se

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

$$= \sum_{i=1}^{n} Q_i y_i , \qquad (2.7)$$

com

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

Desta forma, de (2.2), vem

$$E\left(\hat{\beta}_1\right) = E\left(\sum_{i=1}^{n} Q_i y_i\right) = \sum_{i=1}^{n} Q_i E(y_i)$$

$$= \sum_{i=1}^{n} Q_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum_{i=1}^{n} Q_i + \beta_1 \sum_{i=1}^{n} Q_i x_i \ . \qquad (2.8)$$

Visto que

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

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

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

e que

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

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

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

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

pegando em (2.8) concluímos que

$$E(\hat{\beta}_1) = \beta_0 \times 0 + \beta_1 \times 1 = \beta_1\,, \qquad (2.9)$$

o que significa que $\hat{\beta}_1$ é um estimador centrado de β_1.

Variância de $\hat{\beta}_1$

De (2.7) temos que

$$var(\hat{\beta}_1) = var\left(\sum_{i=1}^{n} Q_i y_i\right).$$

Como $y_i, i = 1, \dots, n$ são variáveis independentes, temos que

$$var(\hat{\beta}_1) = \sum_{i=1}^{n} var(Q_i y_i) = \sum_{i=1}^{n} {Q_i}^2 var(Y_i) = \sum_{i=1}^{n} {Q_i}^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,,$$

visto que

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

b) Valor esperado e variância de $\hat{\beta}_0$

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

Da 1ª equação de (2.5) tem-se $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, e visto que, de (2.9) se tem $E(\hat{\beta}_1) = \beta_1$, obtemos:

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

$$= E(\bar{y}) - \bar{x}E(\hat{\beta}_1)$$

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

$$= \sum_{i=1}^{n}\frac{E(y_i)}{n} - \bar{x}\beta_1$$

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

$$= \frac{n}{n}\beta_0 + \beta_1\frac{\sum_{i=1}^{n}x_i}{n} - \bar{x}\beta_1 = \beta_0\,. \qquad (2.10)$$

Logo $\hat{\beta}_0$ é um estimador centrado de β_0.

Variância de $\hat{\beta}_0$

Tem-se

$$var(\hat{\beta}_0) = var(\bar{y} - \hat{\beta}_1\bar{x}) = var(\bar{y}) + var(\hat{\beta}_1\bar{x}) - 2cov(\bar{y}, \hat{\beta}_1\bar{x})\,. \qquad (2.11)$$

Ora

$$cov(\bar{y}, \hat{\beta}_1\bar{x}) = E(\bar{y}\hat{\beta}_1\bar{x}) - E(\bar{y})E(\hat{\beta}_1\bar{x}) =$$

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

$$= E\left(\bar{x}\frac{\sum_{i=1}^{n}(\beta_0+\beta_1 x_i+\varepsilon_i)}{n}\hat{\beta}_1\right) - \frac{\bar{x}\beta_1}{n}\sum_{i=1}^{n}E(y_i)$$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$= \frac{\bar{x}}{n}\sum_{i=1}^{n}\left[\beta_0\beta_1 + x_i{\beta_1}^2 + E\left(\varepsilon_i\hat{\beta}_1\right)\right] - \frac{\bar{x}\beta_1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i)$$

$E(\hat{\beta}_1) = \beta_1$

$$= \frac{\bar{x}\beta_1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i) + \frac{\bar{x}}{n}\sum_{i=1}^{n}E\left(\varepsilon_i\hat{\beta}_1\right) - \frac{\bar{x}\beta_1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i)$$

$$= \frac{\bar{x}}{n}\sum_{i=1}^{n}E\left(\varepsilon_i\hat{\beta}_1\right).$$

Como

$$E\left(\varepsilon_i\hat{\beta}_1\right) = E\left[\varepsilon_i\frac{\sum_{j=1}^{n}(x_j-\bar{x})y_j}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] =$$

$$= \frac{\sum_{j=1}^{n}(x_j-\bar{x})E\left[\varepsilon_i y_j\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{j=1}^{n}(x_j-\bar{x})E\left[\varepsilon_i(\beta_0+\beta_1 x_j+\varepsilon_j)\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{j=1}^{n}(x_j-\bar{x})\left[\beta_0 E(\varepsilon_i)+\beta_1 x_j E(\varepsilon_i)+E\left(\varepsilon_j\varepsilon_i\right)\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{j=1}^{n}(x_j-\bar{x})E\left(\varepsilon_j\varepsilon_i\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

Uma vez que, quando $i \neq j$, $E\left(\varepsilon_j\varepsilon_i\right) = 0$, vem

$$E\left(\varepsilon_i\hat{\beta}_1\right) = 0\,.$$

Por outro lado, quando $i = j$,

$$E(\varepsilon_i\varepsilon_i) = E(\varepsilon_i^2).$$

Como

$$var(\varepsilon_i) = E(\varepsilon_i^2) - \big(\underbrace{E(\varepsilon_i)}_{=0}\big)^2 = E(\varepsilon_i^2) = \sigma^2$$

vem

$$E\left(\varepsilon_j\varepsilon_i\right) = \sigma^2$$

e consequentemente

$$E(\varepsilon_i\hat{\beta}_1) = \frac{\sum_{j=1}^{n}(x_j - \bar{x})\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sigma^2 \frac{\sum_{j=1}^{n}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0,$$

visto que

$$\sum_{j=1}^{n}(x_j - \bar{x}) = \sum_{j=1}^{n} x_j - n\bar{x} = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} x_j = 0\,.$$

Podemos então concluir que

$$cov(\bar{y}, \hat{\beta}_1\bar{x}) = 0\,.$$

Assim, voltando a (2.11),

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

$$= var\left(\sum_{i=1}^{n}\frac{y_i}{n}\right) + \bar{x}^2 var(\hat{\beta}_1)\,.$$

Como $y_i,\ \ i = 1, \ldots, n$, são independentes, temos que:

$$var(\hat{\beta}_0) = \frac{1}{n^2}\sum_{i=1}^{n} var(y_i) + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

c) Covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = E(\hat{\beta}_0\hat{\beta}_1) - E(\hat{\beta}_0)E(\hat{\beta}_1)$$

$$= E\left[(\bar{y} - \hat{\beta}_1\bar{x})\hat{\beta}_1\right] - \beta_0\beta_1$$

$$= E\left[\bar{y}\hat{\beta}_1 - \bar{x}\hat{\beta}_1^{\,2}\right] - \beta_0\beta_1$$

$$= E(\bar{y}\hat{\beta}_1) - \bar{x}E\left(\hat{\beta}_1{}^2\right) - \beta_0\beta_1 .$$

Como $var(\hat{\beta}_1) = E(\hat{\beta}_1^2) - \left(E(\hat{\beta}_1)\right)^2$, vem

$$cov(\hat{\beta}_0, \hat{\beta}_1) = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\hat{\beta}_1\right] - \bar{x}\left[var(\hat{\beta}_1) + \left(E(\hat{\beta}_1)\right)^2\right] - \beta_0\beta_1$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[(\beta_0 + \beta_1 x_i + \varepsilon_i)\hat{\beta}_1\right] - \bar{x}\left[\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \beta_1{}^2\right] - \beta_0\beta_1$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[\beta_0\hat{\beta}_1 + \beta_1\hat{\beta}_1 x_i + \varepsilon_i\hat{\beta}_1\right] - \frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \bar{x}\beta_1{}^2 - \beta_0\beta_1$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\beta_0\beta_1 + \beta_1{}^2 x_i + E(\varepsilon_i\hat{\beta}_1)\right] - \frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \bar{x}\beta_1{}^2 - \beta_0\beta_1$$

$$= \beta_0\beta_1 + \beta_1{}^2\bar{x} + \frac{1}{n}\sum_{i=1}^{n} E(\varepsilon_i\hat{\beta}_1) - \frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \bar{x}\beta_1{}^2 - \beta_0\beta_1$$

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

uma vez que, como provado anteriormente, $E(\varepsilon_i\hat{\beta}_1) = 0$.

d) Distribuição amostral de $\hat{\beta}_0$ e de $\hat{\beta}_1$

Como vimos anteriormente, de (2.7) temos que:

$$\hat{\beta}_1 = \sum_{i=1}^{n} Q_i y_i ,$$

com $Q_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_1 - \bar{x})^2}$.

Logo $\hat{\beta}_1$ é uma combinação linear dos y_i, $i = 1, \ldots, n$. Assim como o $\hat{\beta}_0$, definido em (2.5). Concluímos portanto que, uma vez que y_i são normalmente distribuídos, com

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \qquad i = 1, \ldots, n .$$

quer $\hat{\beta}_0$ quer $\hat{\beta}_1$ são normalmente distribuídos.

Assim, considerando o valor esperado e a variância de $\hat{\beta}_0$ e $\hat{\beta}_1$ que obtivemos em a) e b) temos que:

. A distribuição amostral para $\hat{\beta}_0$ será:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right);$$

. A distribuição amostral para $\hat{\beta}_1$ será:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

2.4. ESTIMADOR DE σ^2

Tal como os parâmetros do modelo β_0 e β_1, também é necessário obter um estimador da variância dos erros, isto é, um estimador de σ^2.

Como vimos anteriormente

$$SQE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

$$= \sum_{i=1}^{n}\left(y_i^2 - 2y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2\right) =$$

$$= \sum_{i=1}^{n}\left[y_i^2 - 2y_i(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i) + \underbrace{(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i)}_{(\bar{y} - \hat{\beta}_1(\bar{x} - x_i))}^2\right]$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

$$= \sum_{i=1}^{n}\left[y_i^2 - 2y_i\bar{y} + 2y_i\hat{\beta}_1\bar{x} - 2y_i\hat{\beta}_1 x_i + \bar{y}^2 - 2\bar{y}\hat{\beta}_1(\bar{x} - x_i) + \hat{\beta}_1^2(\bar{x} - x_i)^2\right]$$

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

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

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

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

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

Vamos agora calcular o valor esperado de SQE, isto é,

$$E(SQE) = E\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})y_i\right)$$

$$= E\left[\sum_{i=1}^{n}(y_i^2) - n\,\bar{y}^2\right] - E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})y_i\right)$$

$= \sigma^2 + (\beta_0 + \beta_1 x_i)^2$

$= var(y_i) + \left(E(y_i)\right)^2$

$$\sum_{i=1}^{n}\overbrace{E(y_i^2)} - n\underbrace{E(\bar{y}^2)} - \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}E\left[\left(\sum_{i=1}^{n}(x_i - \bar{x})y_i\right)^2\right]$$

$= var(\bar{y}) + \left(E(\bar{y})\right)^2$

$= \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2$

$$= \sum_{i=1}^{n}(\sigma^2 + (\beta_0 + \beta_1 x_i)^2) - n\left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2\right) - \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \left[var\left(\sum_{i=1}^{n}(x_i - \bar{x})y_i\right) + \left(E\left[\sum_{i=1}^{n}(x_i - \bar{x})y_i\right]\right)^2\right]. \quad (2.12)$$

Calculamos

$$var\left(\sum_{i=1}^{n}(x_i - \bar{x})y_i\right) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \times var(y_i) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sigma^2$$

e

$$E\left(\sum_{i=1}^{n}(x_i - \bar{x})y_i\right) = \sum_{i=1}^{n}(x_i - \bar{x})\,E(y_i) = \sum_{i=1}^{n}(x_i - \bar{x})\,(\beta_0 + \beta_1 x_i)\,.$$

Pegando novamente em (2.12), obtemos

$$E(SQE) =$$

$$\underbrace{\sum_{i=1}^{n}\sigma^2}_{=n\sigma^2} + \sum_{i=1}^{n}(\beta_0+\beta_1 x_i)^2 - \sigma^2 - n(\beta_0+\beta_1\bar{x})^2 - \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\times\sigma^2\right) - \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})\,(\beta_0+\beta_1 x_i)\right)^2$$

$$= n\sigma^2 + \sum_{i=1}^{n}(\beta_0+\beta_1 x_i)^2 - \sigma^2 - n(\beta_0+\beta_1\bar{x}_i)^2 - \sigma^2 - \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left[\beta_0\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})}_{=0} + \beta_1\sum_{i=1}^{n}(x_i-\bar{x})\,x_i\right]^2$$

$$= \sigma^2(n-2) + \sum_{i=1}^{n}\overbrace{(\beta_0+\beta_1 x_i)^2}^{\beta_0^2+2\beta_0\beta_1 x_i+\beta_1^2 x_i^2} - n\underbrace{(\beta_0+\beta_1\bar{x}_i)^2}_{\beta_0+2\beta_0\beta_1\bar{x}+\beta_1^2\bar{x}^2} - \beta_1^2\frac{\overbrace{\left(\sum_{i=1}^{n}(x_i-\bar{x})\,x_i\right)^2}^{=\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \sigma^2(n-2) + n\beta_0^2 \underbrace{+ 2\beta_0\beta_1\sum_{i=1}^{n}x_i}_{=2\beta_0\beta_1 n\bar{x}} + \beta_1^2\sum_{i=1}^{n}x_i^2 - n\beta_0 - 2n\beta_0\beta_1\bar{x} - n\beta_1^2\bar{x}^2 - \beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \sigma^2(n-2) + \beta_1^2\left(\overbrace{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}^{=\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) - \beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \sigma^2(n-2) \qquad (2.13)$$

Concluímos portanto que $E(SQE) = \sigma^2(n-2)$, o que implica que o estimador centrado de σ^2 será

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

em que QME representa o quadrado médio dos erros.

2.5. TESTES E INTERVALOS DE CONFIANÇA PARA OS PARÂMETROS DO MODELO

Nesta secção construiremos testes de hipóteses e intervalos de confiança para β_1 e β_0, considerando os pressupostos anteriormente referidos. Estes pressupostos levaram-nos a concluir que as observações

$$y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2), \qquad i = 1, \ldots, n\,.$$

2.5.1. Testes e intervalos de confiança para β_1

Como vimos atrás, o estimador pontual de β_1 é dado por

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

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Vimos também que a distribuição amostral de β_1 para o modelo de regressão normal também é normal, uma vez que β_1 é uma combinação linear dos y_i, com:

$$E(\hat{\beta}_1) = \beta_1 ;$$

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Daí

$$\hat{\beta}_1 \sim N\left(\beta_1; \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Suponhamos que pretendemos testar as hipóteses

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

o que significa que pretendemos testar se β_1 é igual a um determinado valor β_1'.

Assim, a estatística de teste será dada por

$$T = \frac{\hat{\beta}_1 - \beta_1'}{S_{\hat{\beta}_1}} \sim t_{(n-2)},$$

com

$$S_{\hat{\beta}_1} = \sqrt{\frac{QME}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

T tem distribuição t-student com $n-2$ graus de liberdade, $t_{(n-2)}$ (ver por exemplo Maroco, 2003).

Suponhamos agora que pretendemos testar

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}, \qquad (2.14)$$

que são as hipóteses que queremos testar no modelo em questão. Neste caso a estatística de teste poderá ser reescrita da seguinte forma

$$T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t_{(n-2)}. \qquad (2.15)$$

Logo, rejeita-se H_0, para um nível de significância α, se $|T_{obs}| > t_{(1-\alpha/2, n-2)}$, onde T_{obs} representa o valor observado da estatística T e $t_{(1-\alpha/2, n-2)}$ o quantil de ordem $1 - \alpha/2$ da distribuição t com $n - 2$ graus de liberdade.

No que diz respeito ao intervalo de confiança, a $(1-\alpha) \times 100\%$, para β_1 esse será dado por

$$\left[\hat{\beta}_1 - t_{(1-\alpha/2; n-2)} S_{\hat{\beta}_1}; \hat{\beta}_1 + t_{(1-\alpha/2; n-2)} S_{\hat{\beta}_1}\right].$$

2.5.2. Testes e intervalos de confiança para β_0

Como vimos, o estimador pontual de β_0 é dado por:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}\,.$$

Assumindo a normalidade das observações e visto que:

$$E(\hat{\beta}_0) = \beta_0$$

e

$$var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

tem-se

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right).$$

Consideremos as hipóteses:

$$\begin{cases} H_0: \beta_0 = \beta_0' \\ H_1: \beta_0 \neq \beta_0' \end{cases},$$

a estatística de teste será dada por

$$T^0 = \frac{\hat{\beta}_0 - \beta_0'}{S_{\hat{\beta}_0}} \sim t_{(n-2)}\,,$$

com

$$S_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$= \sqrt{QME \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}.$$

T^0 também segue uma distribuição t com $n-2$ graus de liberdade, $t_{(n-2)}$.

Se por outro lado pretendermos testar as hipóteses:

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases},$$

a estatística de teste poderá ser reescrita de seguinte forma

$$T^0 = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} \sim t_{(n-2)}\,.$$

Assim, rejeita-se H_0, para um nível de significância de α, se $|T^0_{obs}| > t_{(1-\alpha/2,n-2)}$, onde T^0_{obs} representa o valor observado de estatística T^0.

Quanto ao intervalo de confiança para β_0, com $(1-\alpha)\times 100\%$ de confiança, será dado por

$$\left[\hat{\beta}_0 - t_{(1-\alpha/2,n-2)} S_{\hat{\beta}_0}; \hat{\beta}_0 + t_{(1-\alpha/2,n-2)} S_{\hat{\beta}_0}\right].$$

3. BREVE ABORDAGEM À REGRESSÃO LINEAR MÚLTIPLA

Neste capítulo faremos uma breve abordagem à regressão linear múltipla.

Como referido anteriormente, a diferença entre a regressão linear múltipla e a regressão linear simples é que na múltipla são consideradas duas ou mais variáveis explicativas (independentes). As variáveis independentes são as ditas variáveis explicativas, uma vez que explicam a variação de y.

Na regressão linear múltipla assumimos que existe uma relação linear entre uma variável y (variável dependente) e p variáveis independentes (preditoras), $x_1, x_2, \dots, x_p$.

3.1. MODELO TEÓRICO E SEUS PRESSUPOSTOS

O modelo de regressão linear múltipla com p variáveis explicativas é definido da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, \dots, n\ , \qquad (3.1)$$

em que

. y_i representa o valor de vaiável resposta na observação $i,\ i = 1, \dots, n$;

. $x_{i1}, x_{i2} \dots x_{ip}, i = 1, \dots, n$ são os valores da i-ésima observação das p variáveis explicativas, (constantes conhecidas);

. $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ são os parâmetros ou coeficientes de regressão;

. $\varepsilon_i,\ i = 1, \dots, n$ correspondem aos erros aleatórios.

Este modelo descreve um hiperplano p-dimensional referente às variáveis explicativas como mostra a Figura **3.1**.

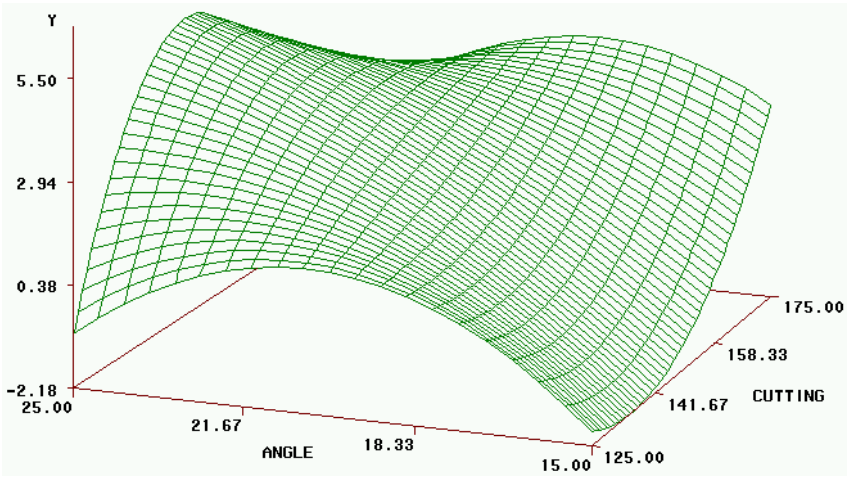

Figura 3.1- Hiperplano p-dimensional referente às variáveis explicativas.

Os parâmetros $\beta_j, j = 1, \dots, p$, representam a média esperada na variável resposta, Y, quando a variável X_j, $i = 1, \dots, p$ sofre um acréscimo unitário, enquanto todas as outras variáveis $X_k, k \neq j$ são mantidas constantes.

Por esse motivo os β_j, $j = 1, \dots, p$ são chamados de coeficientes parciais.

O parâmetro β_0 corresponde ao intercepto do plano de regressão. Se a abrangência do modelo incluir $X_j = 0$, $j = 1, \dots, p$, então β_0 será a média de Y nesse ponto. Caso contrário não existe interpretação prática para β_0.

3.1.1. Interações

Vamos considerar o caso particular do modelo de regressão linear múltipla com duas variáveis explicativas X_1 e X_2. Assim, o modelo será definido por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon\,. \qquad (3.2)$$

Se considerarmos um modelo mais complexo, em que existe interacção entre as variáveis explicativas, obtemos

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \underbrace{X_1 X_2}_{\text{Interacção}} + \varepsilon\,. \qquad (3.3)$$

Neste caso, $X_1 X_2$ representa a interacção existente entre as variáveis X_1 e X_2. Se a interação existir e for significativa, o efeito de X_1 na resposta média depende do nível X_2 e vice-versa.

3.1.2. Pressupostos do modelo

Os pressupostos para o modelo de regressão linear múltipla são análogos ao do modelo de regressão linear simples. Assim tem-se:

a) $E[\varepsilon_i] = 0\ , i = 1, \dots, n$;
b) Os erros são independentes;
c) $V[\varepsilon_i] = \sigma^2,\ i = 1, \dots, n$ (variâncias constantes);
d) Os erros têm distribuição normal.

Destes pressupostos, concluímos que $\varepsilon_i \sim N(0, \sigma^2),\ i = 1, \dots, n$ e consequentemente que $\underline{y}$ tem distribuição normal com varância σ^2 e, para o caso de modelo definido em (3.1),

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\ .$$

3.1.3. Representação matricial do método de regressão linear múltipla

Como vimos em (3.1), a expressão geral de i-ésima observação no modelo de regressão linear (sem interacção) é dada por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \dots, n\ .$$

Este modelo pode ser reescrito em notação matricial da seguinte forma:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}\ , \qquad (3.4)$$

onde

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \underline{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1,p} \\ 1 & x_{21} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{n,p} \end{bmatrix} = [1 \quad \underline{x_1} \quad \cdots \quad \underline{x_p}]$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}\ .$$

Concluímos então que:

. $\underline{\varepsilon}$ é um vector de dimensão $n \times 1$ cujas componentes são os erros aleatórios, $\varepsilon_i,\ i = 1, \dots, n$;

. $\underline{Y}$ é um vector $n \times 1$ cujas componentes correspondem às n respostas, $y_1, \dots, y_n$, constituído pelas observações da variável resposta;

. $\underline{X}$ é uma matriz de dimensão $n \times (p+1)$ denominada matriz do modelo, cujas colunas são constituídas pelos vectores $\underline{1} = (1, \dots, 1)'$ e $\underline{x_j} = \left(x_{1,j}, \dots, x_{n,j}\right)'$, $j = 1, \dots, p$. A notação A' representa a transposta da matriz A.

. $\underline{\beta}$ é um vector coluna $(p+1) \times 1$ cujos elementos são os coeficientes de regressão, $\beta_0, \beta_1, \dots, \beta_p$.

Uma vez que $\underline{\varepsilon}$ é normalmente distribuído, tendo-se $\varepsilon \sim N(\underline{0}, \sigma^2 I_n)$, com $\underline{0}$ o vector nulo e I_n a matriz identidade de ordem n, $\underline{Y}$ será normalmente distribuído com $E(\underline{Y}) = \underline{X}\underline{\beta}$ e matriz de variâncias-covariâncias $cov(\underline{Y}) = \sigma^2 I_n$, isto é

$$\underline{Y} \sim N\left(\underline{X}\underline{\beta}, \sigma^2 I_n\right).$$

3.2. ESTIMAÇÃO DO PARÂMETRO DO MODELO

De modo análogo à regressão simples, usando o método dos mínimos quadrados, pretendemos encontrar o vector de estimadores $\underline{\hat{\beta}}$, com componentes $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$, que minimiza

$$SQE = \sum_{i=1}^{n} {e_i}^2 = \underline{e}'\underline{e} = \left(\underline{Y} - \underline{X}\underline{\beta}\right)' \left(\underline{Y} - \underline{X}\underline{\beta}\right)$$

$$= \underline{Y}'\underline{Y} - \underline{Y}'\underline{X}\underline{\beta} - \underline{\beta}'\underline{X}'\underline{Y} + \underline{\beta}'\underline{X}'\underline{X}\underline{\beta}$$

$$= \underline{Y}'\underline{Y} - 2\underline{\beta}'\underline{X}'\underline{Y} + \underline{\beta}'\underline{X}'\underline{X}\underline{\beta},$$

uma vez que se tem $\underline{Y}'\underline{X}\underline{\beta} = \underline{\beta}'\underline{X}'\underline{Y}$, pois este produto é igual a um escalar.

Derivando $\underline{\hat{\beta}}$ obtemos

$$\frac{\partial SQE}{\partial \underline{\beta}} = -2\underline{X}'\underline{Y} + 2\underline{X}'\underline{X}\underline{\beta}.$$

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

$$-2\underline{X}'\underline{Y} + 2\underline{X}'\underline{X}\underline{\hat{\beta}} = 0$$

$$\Leftrightarrow \left(\underline{X}'\underline{X}\right)\underline{\hat{\beta}} = \underline{X}'\underline{Y}$$

$$\Leftrightarrow \underline{\hat{\beta}} = \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y}, \qquad (3.5)$$

onde, $\underline{X}^{-1}$ representa a matriz inversa de $\underline{X}$. De (3.4), concluimos que o modelo de regressão linear ajustado é

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

e o vector dos resíduos

$$\underline{e} = \underline{Y} - \underline{\hat{Y}}\,.$$

3.2.1. Propriedades dos estimadores

a) Valor esperado de $\underline{\hat{\beta}}$:

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

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

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

$$= E\left[I_{p+1}\underline{\beta}\right] + E\left[\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\varepsilon\right]$$

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

visto que $E\left[\underline{\varepsilon}\right] = 0$ e $\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{X} = I_{p+1}$.

b) Matriz de covariâncias de $\underline{\hat{\beta}}$:

Sendo $\underline{W}$ um vector das variáveis aleatórias $W_1, \ldots, W_n$, então a matriz de covariâncias de $\underline{W}$ é dada por

$$cov\left(\underline{W}\right) = E(WW') - E(W)E(W)'\,,$$

que na forma matricial é escrita como

$$cov\left(\underline{W}\right) = \begin{bmatrix} var(W_1) & cov(W_1,W_2) & cov(W_1,W_3) & \ldots & cov(W_1,W_n) \\ cov(W_2,W_1) & var(W_2) & cov(W_2,W_3) & \ldots & cov(W_2,W_n) \\ cov(W_3,W_1) & cov(W_3,W_2) & var(W_3) & \ldots & cov(W_3,W_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(W_n,W_1) & cov(W_n,W_2) & cov(W_n,W_3) & \ldots & var(W_n) \end{bmatrix}$$

Assim, a matriz de covariâncias de $\underline{\hat{\beta}}$ será definida por:

$$cov\left(\underline{\hat{\beta}}\right) = E\left(\underline{\hat{\beta}}\underline{\hat{\beta}}'\right) - E\left(\underline{\hat{\beta}}\right)E\left(\underline{\hat{\beta}}\right)'$$

$$= E\left(\left(\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y}\right)\left(\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y}\right)'\right) - \underline{\beta\beta}'$$

$$= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'E\left(\underline{Y}\underline{Y}'\right)\underline{X}\left(\underline{X}'\underline{X}\right)^{-1} - \underline{\beta\beta}'$$

$$= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\left[cov\left(\underline{Y}\right) + E\left(\underline{Y}\right)E\left(\underline{Y}\right)'\right]\underline{X}\left(\underline{X}'X\right)^{-1} - \underline{\beta\beta}'$$

$$= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'cov\left(\underline{Y}\right)\underline{X}\left(\underline{X}'\underline{X}\right)^{-1} + \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'E\left(\underline{Y}\right)E\left(\underline{Y}\right)'\underline{X}\left(\underline{X}'\underline{X}\right)^{-1} - \underline{\beta\beta}'.$$

Como vimos anteriormente $cov\left(\underline{Y}\right) = \sigma^2 I_n$ e $E\left(\underline{Y}\right) = \underline{X\beta}$, logo

$$cov\left(\underline{\hat{\beta}}\right) = \sigma^2\underbrace{\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'I_n\underline{X}\left(\underline{X}'\underline{X}\right)^{-1}}_{=I_{p+1}} + \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\left(\underline{X\beta}\right)\left(\underline{X\beta}\right)'\underline{X}\left(\underline{X}'\underline{X}\right)^{-1} - \underline{\beta\beta}'$$

$$= \sigma^2\left(\underline{X}'\underline{X}\right)^{-1} + \underbrace{\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{X}}_{=I_{p+1}}\underline{\beta\beta}'\underbrace{\underline{X}'\underline{X}\left(\underline{X}'\underline{X}\right)^{-1}}_{=I_{p+1}} - \underline{\beta\beta}'$$

$$= \sigma^2\left(\underline{X}'\underline{X}\right)^{-1} + \underline{\beta\beta}' - \underline{\beta\beta}'$$

$$= \sigma^2\left(\underline{X}'\underline{X}\right)^{-1},$$

visto que $\underline{X}'\underline{X}\left(\underline{X}'\underline{X}\right)^{-1} = I_{p+1}$.

3.3. ESTIMADOR DE σ^2

Consideremos a soma do quadrado dos resíduos, que como vimos anteriormente, é definido por:

$$SQE = \sum_{i=1}^{n} {e_i}^2 = \left(\underline{Y} - \underline{\hat{Y}}\right)'\left(\underline{Y} - \underline{\hat{Y}}\right)$$

$$= \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}'\underline{X}'\underline{Y} + \underline{\hat{\beta}}'\underline{X}'\underbrace{\underline{X}\underline{\hat{\beta}}}_{=\underline{\hat{Y}}}.$$

Uma vez que $\underline{X}'\underline{X}\underline{\hat{\beta}} = \underline{X}'\underline{Y}$ e de (3.5) se tem $\underline{\hat{\beta}} = \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y}$ obtemos:

$$SQE = \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}'\underline{X}'\underline{Y}' + \underline{\hat{\beta}}'\underline{X}'\underline{Y} = \underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\underline{X}'\underline{Y}$$

$$= \underline{Y}'\underline{Y} - \underbrace{\underline{Y}'\underline{X}\left(\underline{X}'\underline{X}\right)^{-1}}_{=\underline{\hat{\beta}}'}\underline{X}'\underline{Y} = \underline{Y}'\left(I_n - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\right)\underline{Y}.$$

Pelo que

$$SQE = \underline{Y}'\left(I_n - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\right)\underline{Y}.$$

Se $\underline{Y} \sim N\left(\underline{\mu}; \Sigma\right)$ então, $\underline{Y}'\underline{AY}$ segue uma distribuição qui-quadrado não central com g graus de liberdade e parâmetro de não centralidade de $\delta = \frac{1}{2}\underline{\mu}'\underline{A}\underline{\mu}$, $\underline{Y}'\underline{AY} \sim \chi^2_{g,\delta}$, (ver, por exemplo, Mexia, 1995) e g corresponde à característica de matriz A, $g = car(A)$.

Como assumimos que o vector dos erros $\underline{\varepsilon} \sim N(\underline{0}; \sigma^2 I_n)$, segue que $\underline{Y} \sim N\left(\underline{X\beta}; \sigma^2 I_n\right)$.

Desta forma, obtemos que

$$\frac{SQE}{\sigma^2} = \frac{\underline{Y}'}{\sigma^2}\left[I_n - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\right]\underline{Y} \sim \chi^2_{r;\delta}$$

com $r = car\left(I_n - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\right)$. Neste caso

$$\delta = \frac{1}{2}\frac{\underline{\beta}'\underline{X}'\left(I_n - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\right)\underline{X\beta}}{\sigma^2} = 0$$

e

$$r = n - (p+1),$$

então $\frac{SQE}{\sigma^2}$ segue uma distribuição qui-quadrado central com $n-(p+1)$ graus de liberdade,

$$\frac{SQE}{\sigma^2} \sim \chi^2_{n-(p+1)}.$$

Portanto, um estimador não viciado para σ^2 é dado por:

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

3.4. ANÁLISE DA VARIÂNCIA

A análise de variância é importante para a análise de regressão linear múltipla. Este tema não foi abordado na análise de regressão linear simples, uma vez que não traz novidades em termos de aplicação dos testes, já que o teste t e o teste F darão os mesmos resultados. Basta-nos observar que o teste F é o quadrado do teste t.

Na análise de regressão múltipla, o teste F produz um teste mais geral. Através da sua utilização determina-se se qualquer das variáveis independentes no modelo possui poder de explicação. Cada variável pode então ser testada individualmente com o teste t para determinar se é uma das variáveis significativas.

A análise de variância, baseia-se na decomposição da soma dos quadrados total, SQT, (que corresponde à variação da variável resposta), na soma dos quadrados explicada, SQR, (que corresponde à variação da variável resposta que é explicada pelo modelo) e na soma dos quadrados dos resíduos, SQE, (que corresponde à variação da variável resposta que não é explicada pelo modelo).

Desta forma, podemos escrever,

$$SQT = SQR + SQE \Leftrightarrow$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Assim, no conceito de regressão linear múltipla, as hipóteses a testar serão

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_1: \exists_j: \beta_j \neq 0, \qquad j = 1, \ldots, p \end{cases}.$$

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

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

com $\frac{SQR}{\sigma^2} \sim \chi^2_p$, $\frac{SQE}{\sigma^2} \sim \chi^2_{n-p-1}$ e SQR e SQE independentes. Assim, sob H_0, a estatística de F segue uma distribuição F central com p e $n-(p+1)$ graus de liberdade, $F_{p,n-p-1}$.

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

Estas somas de quadrados podem ser apresentadas numa tabela como a que apresentamos de seguida.

Tabela 3.1- Tabela da análise de variância (ANOVA)

Causas de Variação	Soma Quadrados	Graus Liberdade	Quadrados Médios	F
Regressão	SQR	p	$QMR = \frac{SQR}{p}$	$F = \frac{QMR}{QME}$
Erro (resíduo)	SQE	$n-p-1$	$QEM = \frac{SQE}{n-p-1}$	
Total	SQT	$n-1$		

Como vimos anteriormente o coeficiente de determinação é igual ao quadrado do coeficiente de correlação de Pearson, que agora poderá ser reescrito da seguinte forma

$$R^2 = \frac{variação\ explicada}{variação\ total}$$

$$= \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}.$$

Este coeficiente é usado para quantificar a capacidade explicativa do modelo, ou seja, segundo Esteves and Sousa (2007), é uma medida da proporção da variação da variável resposta Y que é explicada pela equação de regressão quando estão envolvidas as variáveis independentes $X_1, X_2, \dots, X_p$.

Como já foi referido anteriormente,

$$0 \leq R^2 \leq 1\,.$$

Temos no entanto de ter atenção ao facto de que $R^2 \simeq 1$ não significa que o modelo de regressão providencia um bom ajustamento aos dados, dado que a adição de uma variável aumenta sempre o valor deste coeficiente (mesmo que tenha muito pouco poder explicativo sobre a variável resposta).

Desta forma, quando R^2 é elevado em determinados modelos, leva-nos a interpretações erradas de novas observações ou estimativas pouco fiáveis do valor esperado de Y. Por isso, concluímos que R^2 poderá não ser um bom indicador do grau de ajustamento do modelo.

Assim sendo, é preferível utilizar o coeficiente de determinação ajustado, que é uma medida ajustada do coeficiente de determinação e que é "penalizada" quando são adicionadas variáveis pouco explicativas.

O coeficiente de determinação ajustado é definido por:

$$R_a^2 = 1 - \left(\frac{n-1}{n-(p+1)}\right)(1-R^2)\,.$$

Note-se que a inclusão de mais variáveis diminui o valor de R_a^2, pois aumenta p, e não traz muito "incremento" a R^2.

Ou seja, ao contrário do coeficiente de determinação R^2, o coeficiente de determinação ajustado, R_a^2, não aumenta sempre quando adicionamos uma nova variável. Aliás, se adicionarmos variáveis com pouco poder explicativo este tende a decrescer. Pelo que, quando existe uma diferença significativa entre R^2 e R_a^2, estamos perante uma situação em que provavelmente tenham sido incluidas no modelo variáveis estatisticamente não significativas.

4. ANÁLISE DE RESÍDUOS

Como vimos nos capítulos anteriores os resíduos são dados pela diferença entre os valores da variável resposta observada e a variável resposta estimada, isto é,

$$e_i = y_i - \hat{y}_i, \qquad i = 1, \dots, n\,.$$

Ao realizarmos uma análise de resíduos pretendemos verificar se o modelo de regressão que está a ser utilizado é adequado. Para tal os resíduos devem verificar os pressupostos anteriormente impostos ao erro do modelo. Tais pressupostos são, considerando o modelo

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}\,,$$

com

$$\underline{Y} = (y_1, \dots, y_n)', \underline{X} \text{ a matriz do modelo}, \underline{\beta} = \left(\beta_0, \dots, \beta_p\right)' \text{ e } \underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'\,,$$

a) $\varepsilon_i,\ i = 1, \dots, n$ são normalmente distribuídos;
b) $var(\varepsilon_i) = \sigma^2,\ i = 1, \dots, n$, têm variância constante (homoscedasticidade);
c) ε_i e $\varepsilon_j, i \neq j$, são independentes;
d) não existem *Outliers* influentes.

No caso da regressão linear múltipla, para além destes pressupostos, é preciso ainda verificar se existe colinearidade ou multicolinearidade entre as variáveis explicativas.

De seguida apresentamos algumas "técnicas" por forma a verificar estes pressupostos.

4.1. DIAGNÓSTICO DE NORMALIDADE

A normalidade dos resíduos pode ser analisada quer através de gráficos, quer usando alguns testes, nomeadamente através do

i. gráfico P-P plot dos resíduos;
ii. histograma dos resíduos estandardizados;
iii. teste de Kolmogorov-Smirnov;
iv. teste de Shapiro-Wilk.

Vejamos:

i. Gráfico P-P plot dos resíduos;

Neste gráfico, vamos visualizar a distribuição de probabilidades dos valores observados com os valores esperados, representada por uma diagonal, segundo uma distribuição normal.

Caso a normalidade se verifique, as observações registadas aproximam-se dessa diagonal, sem nenhum afastamento significativo.

A Figura 4.1 mostra o gráfico p-p plot de resíduos. Nesta situação a normalidade é verificada já que os pontos se aproximam da recta.

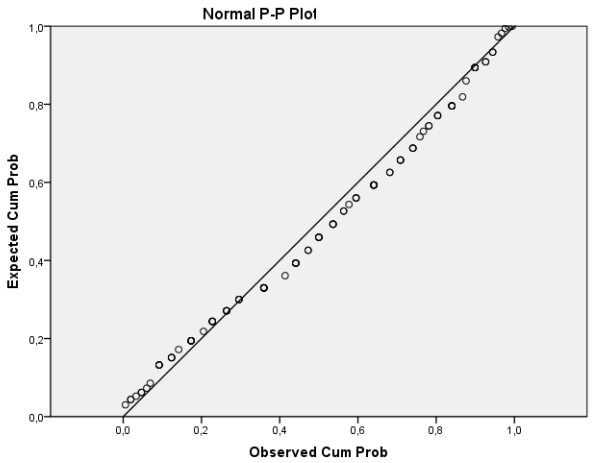

Figura 4.1– Normal p-p plot de resíduos

ii. Histograma dos resíduos estandardizados

Também se pode fazer um histograma dos resíduos no qual se procuram afastamentos evidentes em relação à forma simétrica e unimodal da distribuição normal. Este gráfico apenas deverá ser utilizado em amostras de dimensão elevada, já que quando se trabalha com amostras de dimensão reduzida o histograma não é muito conclusivo.

iii. Teste de Kolmogorov-Smirnov (K-S)

Neste caso o teste de K-S é utilizado para testar as hipóteses:

$$\begin{cases} H_0\text{: A distribuição é normal} \\ H_1\text{:A distribuição não é normal} \end{cases}.$$

A estatística de teste, é dada por, ver Maroco (2003),

$$D = \max\{\max(|F(x_i) - F_0(x_i)|)\,;\max(|F(x_i - 1) - F_0(x_i)|)\}$$

em que $F(x_i) - F_0(x_i)$ representa a diferença entre a frequência acumulada de cada uma das observações e a frequência acumulada que essa observação teria, sendo a sua distribuição normal.

Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada assumida pelos dados, neste caso da distribuição normal, e a função de distribuição empírica dos dados.

iv. Teste de Shapiro-Wilk (S-W)

Este teste sugere-nos preferência em relação ao teste de K-S para amostras de pequenas dimensões $(n < 30)$. Neste caso, as hipóteses a serem testadas são as definidas anteriormente para o teste de K-S.

A estatística de teste é definida da seguinte forma:

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

onde:

. a_i são constantes geradas a partir da média, variância e covariância de n ordens, ver Maroco (2003).

4.2. DIAGNÓSTICO DE HOMOSCEDASTICIDADE (VARIÂNCIA CONSTANTE)

Um dos pressupostos do modelo de regressão linear é a de que os erros devem ter variância constante. Esta condição é designada por homoscedasticidade.

A variância ser constante equivale a supor que não existem observações incluídas na variável residual cuja influência seja mais intensa na variável dependente.

Uma das técnicas usadas para verificar a suposição de que os resíduos são homoscedásticos, é a análise do gráfico dos resíduos versus valores ajustados. Este gráfico deve apresentar pontos dispostos aleatoriamente sem nenhum padrão definido, como se pode ver, por exemplo na Figura 4.2.

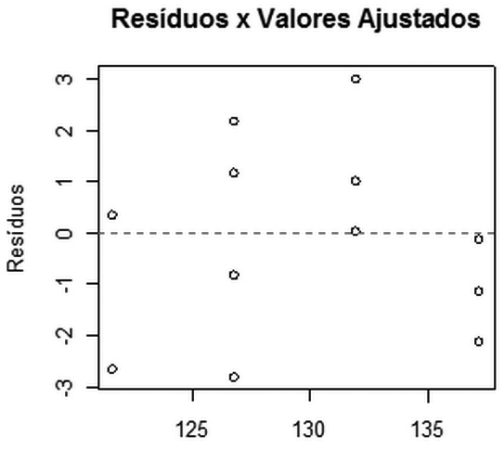

Figura 4.2- Confirmação da homoscedasticidade dos resíduos (disponível em PortalAction).

Por isso, se os pontos estão aleatoriamente distribuídos em torno da recta $y = 0$, sem nenhum comportamento ou tendência, temos indícios de que a variância dos resíduos é constante. Já a presença, por exemplo, de "funil" é um indicativo da presença de heteroscedasticidade.

4.3. DIAGNÓSTICO DE INDEPENDÊNCIA

Para testar o pressuposto da independência dos resíduos, ou a presença de autocorrelação entre eles, pode utilizar-se o teste de Durbin-Watson (DW).

O teste de Durbin-Watson testa as hipóteses:

$$\begin{cases} H_0: Não\ existe\ autocorrelação\ dos\ resíduos \\ H_1: Existe\ autocorrelação\ positiva\ dos\ resíduos \end{cases}.$$

A estatística de teste é dada por:

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

e toma valores entre zero e quatro, $0 \leq dw \leq 4$.

Esta estatística mede a correlação entre cada resíduo e o resíduo correspondente à observação imediatamente anterior.

Podemos tomar a decisão comparando o valor de dw com os valores críticos d_L e d_U da tabela de Durbin-Watson disponível no Anexo 1.

A tabela seguinte dá-nos as decisões a tomar em função dos valores críticos, d_U e d_L.

Tabela 4.1- Tabela de decisão em função de $\boldsymbol{d_U}$ e $\boldsymbol{d_L}$

	Zona de Rejeição e de não-rejeição de H_0				
dw	$[0; d_L[$	$[d_L; d_U[$	$[d_U; 4-d_U[$	$[4-d_U; 4-d_L[$	$[4-d_L; 4]$
Decisão	Rejeitar H_0	Nada se pode concluir	Não rejeitar H_0	Nada se pode concluir	Rejeitar H_0
	Auto-correlação positiva		Os resíduos são independentes		Auto-correlação negativa

4.4. DIAGNÓSTICO DE *OUTLIER*S E OBSERVAÇÕES INFLUENTES

De acordo com Pires e Branco (2007), *Outlier*s são observações extremas que se encontram de tal forma afastadas da maioria dos dados que surgem dúvidas sobre se elas poderão ou não ter sido geradas pelo modelo proposto para explicar essa maioria dos dados.

Os *Outlier*s podem ser classificados em severos ou moderados consoante o seu afastamento em relação às restantes observações. Os *Outliers* moderados encontram-se fora do intervalo $[Q_1 - 1{,}5Q; Q_1 + 1{,}5Q]$ e os *Outliers* severos encontram-se fora do intervalo $[Q_1 - 3Q; Q_1 + 3Q]$, em que Q_1 representa o 1º quartil dos dados e Q a amplitude interquartil, isto é, é a diferença entre o 3º e o 1º quartil, $Q = Q_3 - Q_1$.

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

Uma medida que serve para diagnosticar *Outliers* é *Leverage* (LEV). Para uma dada observação, um *Leverage* elevado indica que essa observação se distancia do centro das observações exercendo influência sobre o valor previsto. O *Leverage* varia entre 0 e 1.

Acontece que um elevado *Leverage* indica apenas que a observação poderá ser influente.

Considera-se um *Leverage* elevado quando, ver Pestana e Gageiro, 2005a,

$$LEV > \frac{3(p+1)}{n}, \qquad para\ amostras\ de\ dimensão\ reduzida$$

$$LEV > \frac{2(p+1)}{n}, \qquad para\ amostras\ grandes$$

onde n é a dimensão da mostra e p o número de variáveis independentes.

Segundo Pires e Branco, (2007), para se perceber os problemas que a presença de *Outliers* podem causar à estimação dos mínimos quadrados é conveniente distinguir vários tipos de *Outliers*:

a) *Outlier* de regressão:

Trata-se de um ponto que se afasta significativamente da estrutura linear descrita pelos dados e que influencia a estimação, conduzindo a modelos ajustados impróprios.

b) *Outlier* em x (ponto de *Leverage* ou alavanca)

É um ponto que é um *Outlier* em relação à coordenada x, isto é, a coordenada x está demasiado afastada das restantes. É um potencial *Outlier* de regressão.

c) *Outlier* em y

É um ponto que é *Outlier* em relação à coordenada y. Pode ou não ser um *Outlier* de regressão.

d) *Outlier* em (x, y)

Um ponto que é *Outlier* nas duas coordenadas. Este pode ou não ser um *Outlier* de regressão."

Uma vez que uma observação pode ser considerada um *Outlier* e pode ou não ser uma observação influente é importante identificar quais as observações influentes. De seguida serão apresentadas algumas "técnicas" que permitem essa identificação.

4.4.1. Observações Influentes

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

Nesta subsecção, apresentamos várias medidas que são utilizadas para identificar as observações influentes.

1) SDFFIT

   É uma das medidas de utilização mais frequente para medir a influência de cada observação. SDFFIT trata-se de uma medida estandardizada que mede a influência que a observação i tem sobre o seu valor ajustado.

   Considera-se que uma observação é influente se, ver Pestana e Gageiro, 2005a,

$$|SDFFIT| > 2\sqrt{\frac{p+1}{n-p-1}}\ .$$

2) SDFBETA

A influência que uma observação tem sobre a estimação de cada um dos coeficientes de regressão pode ser calculada pelo SDFBETA. Trata-se de uma medida estandardizada que corresponde à alteração nos coeficientes estimados, $\hat{\beta}_j,\ j = 0, \dots, p$, quando se exclui essa observação.

Neste caso, a observação é influente quando,

$$|SDFBETA| > 1{,}96, \qquad para\ amostras\ de\ dimensão\ reduzida$$

$$|SDFBETA| > \frac{2}{\sqrt{n}}, \qquad para\ amostras\ grandes$$

3) Para verificar se uma observação é influente também podemos usar a distância de Cook que mede a influência da i-ésima observação sobre todos os n valores ajustados $\hat{y}_i,\ i = 1, \dots, n$.

Uma distância de Cook elevada significa que o resíduo e_i é elevado, ou a *Leverage* para essa observação é elevada, ou ambas as situações.

De tal forma que, uma observação é influente quando, ver Pestana e Gageiro, 2005a,

$$COOK > \frac{4}{n - p - 1},$$

em que n é a dimensão da amostra e o p o número de variáveis independentes. Considera-se que observações com Distância de Cook superior a 1 são excessivamente influentes.

4.5. COLINEARIDADE E MULTICOLINEARIDADE

Como foi referido atrás, na regressão linear múltipla é importante efectuar uma análise de colinearidade e multicolinearidade.

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

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

Quando uma das variáveis independentes já se encontra no modelo de regressão, a inclusão de outra variável independente, não implica, uma explicação adicional significativa da variação total da variável dependente.

A colinearidade poderá ser diagnosticada:

. verificando se a matriz de correlações das variáveis independentes demonstra correlações elevadas. Caso a correlação de duas variáveis seja muito próxima de 1, indica de facto um problema;

. Verificando se, ao se realizar a regressão de X_i em função das outras variáveis independentes, o valor de $R^2 \cong 1$.

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

A variância de cada um dos coeficientes de regressão associados às variáveis independentes é dada por, ver Maroco (2003):

$$var(\hat{\beta}_i) = \sigma^2 \left(\frac{1}{1-R_i^2}\right) \times \frac{1}{\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2},$$

em que R_i^2 é o R^2 de regressão de X_i sobre as restantes variáveis explicativas.

Esta variância é tanto maior quanto maior for a correlação múltipla entre X_i e as variáveis independentes.

O termo $\frac{1}{1-R_i^2}$ designa-se, em concreto, por VIF para o coeficiente de regressão β_i associado à variável X_i.

Segundo Maroco (2003), caso se obtenham valores de $VIF > 5$ conclui-se que estamos perante problemas com a estimação de β_i devido à presença de multicolinearidade nas variáveis independentes.

Suponhamos que temos a equação de regressão

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon\,,$$

em que X_1 e X_2 são altamente correlacionadas.

Numa situação deste género devíamos eliminar uma das variáveis e reestimar o modelo.

Existem vários métodos que permitem, na egressão linear múltipla, fazer uma selecção das variáveis independentes que melhor explicam a variável resposta, nomeadamente:

. *FORWARD* – o método começa apenas com a constante e adiciona uma variável independente de cada vez. A primeira variável selecionada é a que apresenta maior correlação com a variável resposta (maior *score statistic*)

. *BACKWARD* – o método faz o "contrário" do método *Forward*. Neste caso todas as variáveis independentes são incorporadas no modelo. Depois, por etapas, cada uma pode ser ou não eliminada.

. *STEPWISE* - o método *Stepwise* é uma "modificação" do método *Forward* que permite resolver problemas de multicolinearidade. Consiste no seguinte: fazemos entrar no modelo a variável explicativa que apresenta maior coeficiente de correlação com a variável dependente.

Em seguida, calculam-se os coeficientes de correlação parcial para todas as variáveis que não fazem parte da primeira equação de regressão, para que, a próxima variável a entrar, seja a que apresenta maior coeficiente de correlação parcial.

Estima-se a nova equação de regressão e analisa-se se uma das duas variáveis independentes deve ser excluída do modelo.

No final, se ambas as variáveis apresentarem valores *t* significativos, novos coeficientes de correlação parcial são calculados para as variáveis que não entraram.

Este processo finda, assim que se chegue à situação em que nenhuma variável deva ser acrescentada à equação.

5. APLICAÇÕES

Os dados que usamos neste capítulo foram cedidos pela Doutora Catarina Reis Santos, Nefrologista na Unidade Local de Saúde de Castelo Branco. A amostra é constituída por 35 utentes da Consulta de Hipertensão e Dislipidémia, no Hospital de Sta. Marta, Lisboa, em 2010, e os dados foram recolhidos através de documento próprio, Avaliação da Variabilidade da Frequência Cardíaca, disponível no anexo 2. Estes dados foram recolhidos com o intuito de avaliar a existência de diferenças da variabilidade da frequência cardíaca, entre utentes diabéticos e não diabéticos.

O nosso estudo vai-se concentrar nas seguintes variáveis: peso, colesterol total (CT), triglicéridos e *high density lipoprotein* (HDL).

Estes dados foram tratados recorrendo ao *software* SPSS, versão 19, de onde provêm as tabelas e figuras que apresentamos neste capítulo.

Serão apresentados dois estudos. No primeiro estudo foi utilizado o Modelo de Regressão Linear Simples, enquanto que no segundo considerámos o Modelo de Regressão Linear Múltipla, com três variáveis explicativas.

5.1. ESTUDO 1 – MODELO DE REGRESSÃO LINEAR SIMPLES

Com vista a perceber se o nível de HDL no sangue, (mg/dl), influencia o nível de CT no sangue (mg/dl), foi realizada uma análise de regressão linear simples.

O modelo de regressão linear simples que representa a relação entre a variável dependente, CT, e a variável independente, HDL, é dado pela seguinte equação:

$$CT = \beta_0 + \beta_1 HDL + \varepsilon \qquad (5.1)$$

Tabela 5.1- Estatística descritiva

	Mean	Std. Deviation	N
CT	154,8857	103,45467	35
HDL	33,6000	19,72040	35

A Tabela 5.1- Estatística descritiva mostra o valor médio e o desvio padrão de CT e HDL. Concluimos que, nesta amostra, o nível de concentração de CT no sangue é em média 154,8857 (mg/dl), enquanto que de HDL é 33,6.

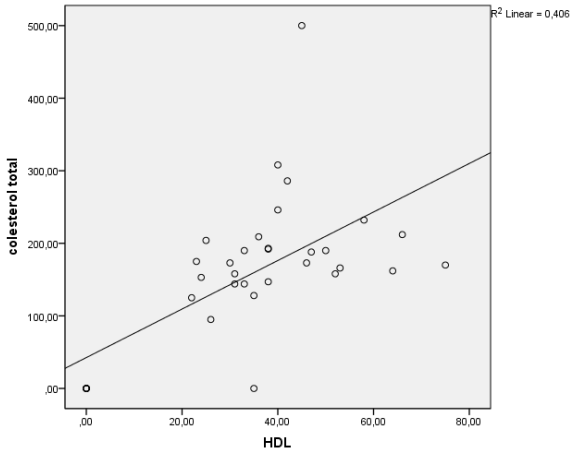

Figura 5.1– Diagrama de dispersão

Elaborámos um diagrama de dispersão com o intuito de perceber se a relação existente entre as duas variáveis é de facto linear.

De acordo com a observação do diagrama de dispersão (Figura 5.1) somos tentados a concluir que existe uma relação linear entre o CT e o HDL e que as duas variáveis tendem a variar no mesmo sentido, o que significa que o aumento da variável independente, HDL, provoca um aumento da variável dependente, CT.

Analisando a Tabela 5.2 podemos afirmar que a correlação existente entre as variáveis é positiva moderada ($R = 0{,}637$).

Tabela 5.2– Sumário do Modelo

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,637[a]	,406	,388	80,91691	1,739

a. Predictors: (Constant), HDL
b. Dependent Variable: colesterol total

Continuando a análise à Tabela 5.2 concluímos que o valor de $R^2(0{,}406)$ e de $R_a^2\ (0{,}388)$ não são muito diferentes.

Como foi dito no capítulo 3, a nossa preferência recai sobre o valor do coeficiente de correlação ajustado que, neste caso, nos leva a afirmar que 38,8% da variabilidade da variável dependente CT é explicada pela variável independente HDL, sendo a restante variabilidade explicada por factores não incluídos no modelo.

Tabela 5.3– Tabela da ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	147828,525	1	147828,525	22,578	,000[a]
	Residual	216069,018	33	6547,546		
	Total	363897,543	34			

a. Predictors: (Constant), HDL
b. Dependent Variable: colesterol total

Para efectuar a análise de variância do modelo recorreu-se ao teste de F que tem associado o seguinte $p-value(sig)$de $0{,}000$. De acordo com o seu valor, rejeitamos $H_0: \beta_1 = 0$, pelo que podemos dizer que o modelo é significativo.

Tabela 5.4**-** Coeficientes

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42,538	27,315		1,557	,129
	HDL	3,344	,704	,637	4,752	,000

a. Dependent Variable: colesterol total

A equação do modelo ajustado de regressão será, segundo a Tabela 5.4

$$CT = 42{,}538 + 3{,}344HDL\,. \qquad (5.2)$$

O teste ao coeficiente de regressão β_0 é dado pelo teste t-student ao qual está associado um valor de significância de 0,129 (>0,05). Concluímos, portanto, que não se deve rejeitar a hipótese $H_0: \beta_0 = 0$, o que significa que a recta ajustada passa pela origem.

Quanto ao teste para β_1 é dado pelo teste t-student ao qual está associado um valor de significância de 0,000 (<0,05). Logo rejeita-se a hipótese $H_0: \beta_1 = 0$, o que significa que a variável HDL influencia significativamente o CT.

Fazendo a comparação dos resultados do teste t com o teste F, verificamos que foram obtidos os mesmos resultados como podemos confirmar pelas tabelas Tabela 5.3 e 5.4, tal como era esperado pelo que foi justificado no capítulo 2.

5.1.1. Verificação dos pressupostos do modelo

O modelo definido em (5.2) só será adequado se validados todos os pressupostos. Vamos nesta subsecção fazer uma análise desses pressupostos.

- NORMALIDADE DOS RESÍDUOS

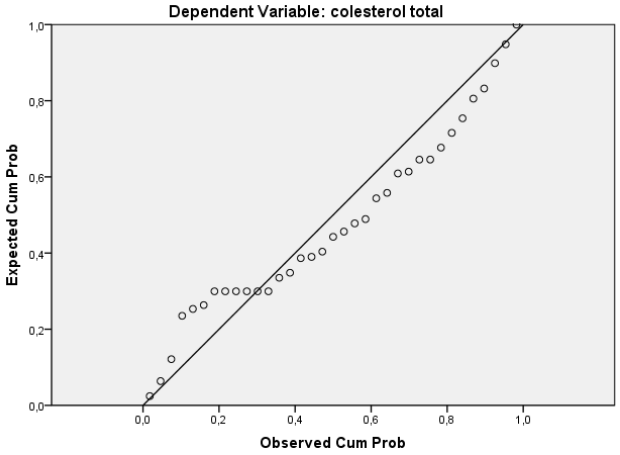

Figura 5.2– Normal p-p plot

A partir da análise da Figura 5.2, podemos concluir que as observações se aproximam da recta sem nenhum afastamento sistemático, pelo que somos levados a concluir que os resíduos são normalmente distribuídos.

Tabela 5.5- Teste K-S

		Unstandardized Residual
N		35
Normal Parameters[a,b]	Mean	,0000000
	Std. Deviation	75,68119963
Most Extreme Differences	Absolute	,157
	Positive	,139
	Negative	-,157
Kolmogorov-Smirnov Z		,930
Asymp. Sig. (2-tailed)		,353

a. Test distribution is Normal.
b. Calculated from data.

Com o intuito de confirmar a normalidade dos resíduos realizámos o teste K-S apresentado na Tabela 5.5. Pelo valor obtido de significância (0,353) concluímos que não se rejeita H_0, pelo que os resíduos são normalmente distribuídos.

- AUTOCORRELAÇÃO DOS RESÍDUOS

  O valor do teste de Durbin-Watson foi de 1,739, como se pode ver na Tabela 5.2.
  Uma vez que este valor pertence ao intervalo $[d_U; 4 - d_U[$ (ver Tabela **4.1** e Tabela do Anexo 1) somos levados a concluir que os resíduos são independentes.

- HOMOSCEDASTICIDADE DOS RESÍDUOS

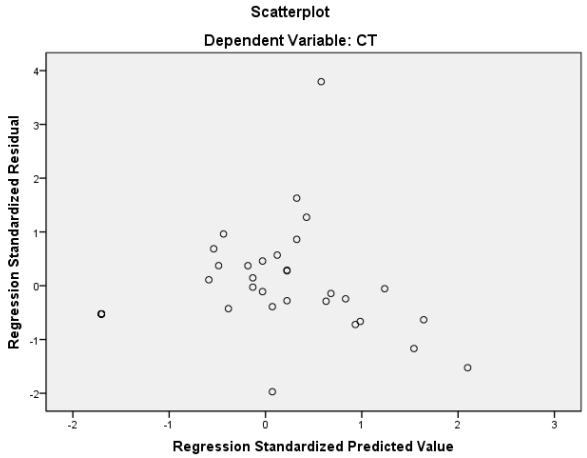

Figura 5.3– Gráfico dos resíduos estandardizados

A partir da análise gráfica da Figura 5.3 concluímos que os resíduos são homoscedásticos uma vez que estes se distribuem de forma aleatória em torno zero (0). (ver Maroco (2003) e Pestana e Gageiro (2005b))

- *OUTLIERS* E OBSERVAÇÕES INFLUENTES

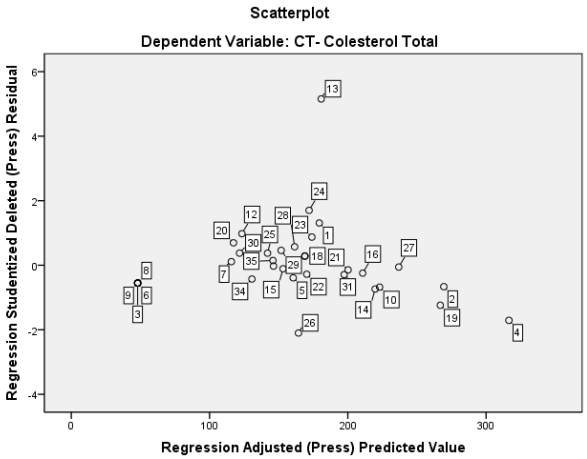

Figura 5.4 – Gráfico resíduos *press*

Pela análise gráfica dos resíduos estandardizados (Figura 5.3) e dos resíduos *press* (Figura 5.4) podemos concluir que existem *Outliers*, dado que há resíduos que apresentam valores absolutos superiores a 1,96, sendo eles os correspondentes às observações 4, 13 e 26, como mostra a Figura 5.4. (ver Maroco, 2003)

Tabela 5.6- Estatística dos resíduos

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	42,5383	293,3138	154,8857	65,93859	35
Std. Predicted Value	-1,704	2,099	,000	1,000	35
Standard Error of Predicted Value	13,684	32,184	18,486	5,776	35
Adjusted Predicted Value	48,0091	316,4877	156,4813	66,73882	35
Residual	-159,56686	306,99640	,00000	79,71807	35
Std. Residual(ZRE_1)	-1,972	3,794	,000	,985	35
Stud. Residual(SRE_1)	-2,001	3,869	-,009	1,013	35
Deleted Residual	-164,28506	319,25589	-1,59559	84,28194	35
Stud. Deleted Residual(SDRE_1)	-2,102	5,154	,026	1,177	35
Mahal. Distance	,001	4,407	,971	1,275	35
Cook's Distance	,000	,299	,029	,065	35
Centered Leverage Value	,000	,130	,029	,038	35

a. Dependent Variable: colesterol total

A confirmação da existência de *Outlier*s pode ser feita, por exemplo, através do valor máximo do *student deleted residual* que neste caso corresponde ao valor 5,154 > 1,96. E também analizando o valor da *Leverage* centrada máxima, que é igual a $0{,}130 > \frac{2(p+1)}{n} = 0{,}12,\ \ n = 35,\ \ p = 1$. Interessa verificar se *Outliers* são ou não observações influentes. Olhando ainda para a Tabela 5.6 tudo leva a crer que sim, já que temos como valor máximo da distância de COOK $0{,}299 > \frac{4}{n-p-1} = 0{,}121$.

Vamos usar mais uma técnica para averiguar a existência de Observações Influentes, recorrendo à análise dos SDFFIT. Para isso apresentamos o gráfico dos SDFFIT (Figura 5.5).

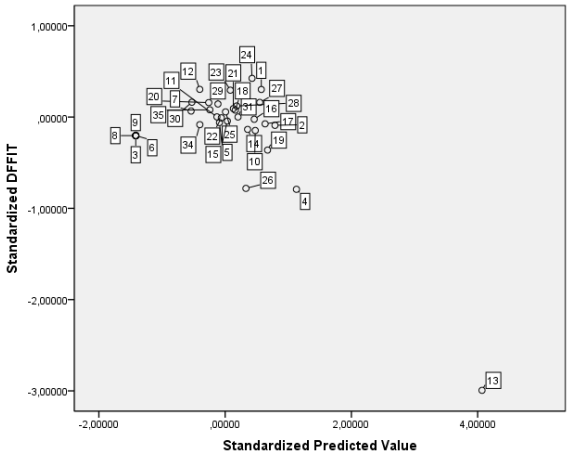

Figura 5.5– Gráfico dos *Standardized* DFFIT

Visto que as observações 4, 13 e 26 da Figura 5.5 têm $|SDFFIT| > 2\sqrt{\frac{p+1}{n-p-1}} \simeq 0{,}49$, concluímos que se tratam de observações influentes, devendo ser mantidas no estudo como é sugerido, por exemplo, por Pestana e Gageiro (2005a).

Conclusão

Uma vez que todos os pressupostos da regressão linear simples foram validados, podemos concluir que o modelo (5.2) é adequado justificando correctamente os dados.

5.2. ESTUDO 2 – MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Neste estudo passamos a considerar o modelo de regressão linear múltipla, que será "estimado" através do método *Stepwise*. O que pretendemos averiguar é de que forma o peso, o nível de HDL, e o nível de triglicéridos (mg/dl) influenciam o nível de CT no sangue.

O modelo de regressão linear múltipla que representa a relação entre a variável dependente, CT, e as variáveis independentes, peso, HDL, triglicéridos, é dado pela seguinte equação:

$$CT = \beta_0 + \beta_1 peso + \beta_2 HDL + \beta_3 triglicéridos + \varepsilon \qquad (5.3)$$

Tabela 5.7- Estatística descritiva

	Mean	Std. Deviation	N
CT- Colesterol Total	154,8857	103,45467	35
HDL	33,6000	19,72040	35
Peso	73,8429	21,21624	35
Triglicéridos	160,7429	298,36134	35

O valor médio de CT é aproximadamente 154,8857 mg/dl enquanto que o valor médio de HDL é de 33,6 mg/dl aproximadamente e dos triglicéridos é 160,7429 mg/dl. O peso médio dos indivíduos da amostra é de 73,8 Kg, aproximadamente.

Tabela 5.8– Tabela de variáveis inseridas/removidas

Model	Variables Entered	Variables Removed	Method
1	Triglicéridos	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	HDL	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: CT- Colesterol Total

A Tabela 5.8 confirma a utilização do método *Stepwise*. Verificamos que a primeira variável a entrar é triglicéridos, seguindo-se a variável HDL.

Tabela 5.9- Sumário do Modelo

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,682[a]	,465	,449	76,81933	
2	,856[b]	,733	,716	55,10980	1,716

a. Predictors: (Constant), Triglicéridos
b. Predictors: (Constant), Triglicéridos, HDL
c. Dependent Variable: CT- Colesterol Total

Observando a Tabela 5.9, concluímos que R^2 e R_a^2 tomam valores aproximados, sendo que o maior valor de R_a^2 corresponde ao modelo em que são consideradas as duas variáveis explicativas, HDL e triglicéridos. Concluímos que este modelo será provavelmente o que melhor explica os valores de CT. Este valor permite-nos afirmar que 71,6% ($R_a^2 = 0.716$) da variabilidade de CT é explicada por este modelo.

Tabela 5.10– Tabela da ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	169157,648	1	169157,648	28,665	,000[a]
	Residual	194739,895	33	5901,209		
	Total	363897,543	34			
2	Regression	266710,661	2	133355,331	43,909	,000[b]
	Residual	97186,882	32	3037,090		
	Total	363897,543	34			

a. Predictors: (Constant), Triglicéridos
b. Predictors: (Constant), Triglicéridos, HDL
c. Dependent Variable: CT- Colesterol Total

Pela análise do valor de significância do teste F (0,000) concluímos que o modelo é altamente significativo. Constata-se que o CT é explicado pelas duas variáveis independentes (triglicéridos e HDL). Esta conclusão pode ser confirmada observando a significância do teste t da Tabela 5.11 ($sig = 0{,}000$).

Tabela 5.11- Coeficientes

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	116,885	14,798		7,899	,000	86,778	146,992		
	Triglicéridos	,236	,044	,682	5,354	,000	,147	,326	1,000	1,000
2	(Constant)	29,500	18,720		1,576	,125	-8,630	67,631		
	Triglicéridos	,202	,032	,582	6,256	,000	,136	,268	,964	1,037
	HDL	2,766	,488	,527	5,667	,000	1,772	3,760	,964	1,037

a. Dependent Variable: CT- Colesterol Total

Ainda da análise da Tabela 5.11 concluímos que o modelo ajustado tem como equação

$$CT = 29{,}5 + 0{,}202triglicéridos + 2{,}766HDL \qquad (5.4)$$

Ambas as variáveis explicativas apresentam um coeficiente positivo o que parece fazer sentido, uma vez que, o aumento do nível de triglicéridos e de HDL fazem aumentar o valor de CT.

Tabela 5.12– Variáveis excluídas

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	HDL	,527[a]	5,667	,000	,708	,964	1,037	,964
	Peso	-,075[a]	-,586	,562	-,103	1,000	1,000	1,000
2	Peso	-,070[b]	-,762	,452	-,136	,999	1,001	,964

a. Predictors in the Model: (Constant), Triglicéridos
b. Predictors in the Model: (Constant), Triglicéridos, HDL
c. Dependent Variable: CT- Colesterol Total

A Tabela 5.12 dá-nos a informação de quais as variáveis excluídas da análise. A variável excluída é o peso, dado que pelo valor da significância para o teste t (0.452) leva à não rejeição da hipótese nula. Assim, esta variável não influencia significativamente os níveis do CT.

O que vai de encontro à realidade, uma vez que o aumento de peso nos individuos não significa que estes venham a "sofrer" de níveis de CT no sangue superiores ao níveis normais. Pode inclusivamente surgir a situação de que individuos com peso abaixo do peso ideal "sofram" de níveis bastante elevados de CT.

5.2.1. Verificação dos pressupostos do modelo

Com o intuito de verificar se o modelo (5.4) é adequado, de seguida é feita uma análise dos pressupostos da regressão linear múltipla.

- NORMALIDADE DOS RESÍDUOS

  Observando a Figura 5.6 parecem existir alguns pontos que se afastam da diagonal principal, não sendo conclusivos quanto à normalidade dos resíduos.

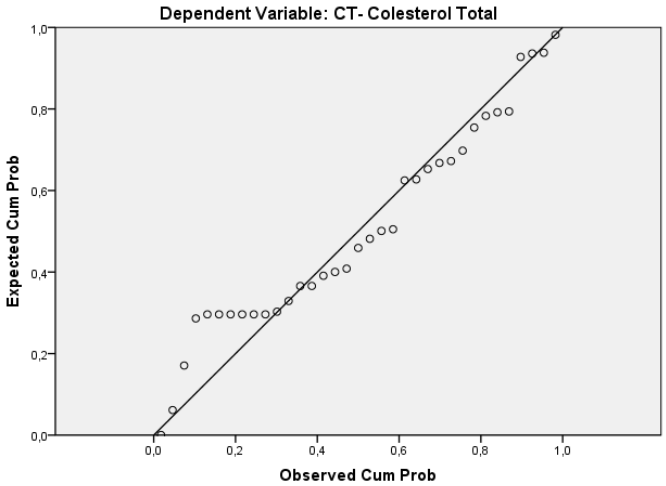

Figura 5.6– Normal p-p plot da regressão dos resíduos estandardizados

Para confirmar a normalidade realizamos o teste K-S apresentado na Tabela 5.13. Mediante o valor de significância obtido (0,123) confirmamos que os resíduos são normalmente distribuídos, devido a não se rejeitar a hipótese nula.

Tabela 5.13– Teste K-S

		Unstandardized Residual
N		35
Normal Parameters[a,b]	Mean	,0000000
	Std. Deviation	53,46435912
Most Extreme Differences	Absolute	,195
	Positive	,095
	Negative	-,195
Kolmogorov-Smirnov Z		1,152
Asymp. Sig. (2-tailed)		,141
Exact Sig. (2-tailed)		,123
Point Probability		,000

a. Test distribution is Normal.
b. Calculated from data.

- AUTOCORRELAÇÃO DOS RESÍDUOS

Considerando o resultado obtido para o teste de Durbin-Watson, apresentado na Tabela 5.9 (1,716) e uma vez que eese valor pertence ao intervalo $[d_U; 4 - d_U[$, concluímos que os resíduos são independentes (ver Tabela 4.1 e Tabela do Anexo 1).

- HOMOSCEDASTICIDADE DOS RESÍDUOS

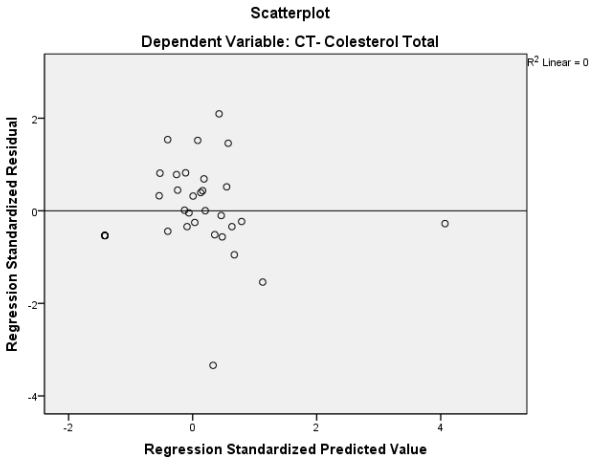

Figura 5.7- Gráfico dos resíduos estandardizados

A partir da análise gráfica dos resíduos estandardizados, Figura 5.7, como os resíduos se distribuem aleatoriamente em torno de zero, concluímos que os resíduos são homoscedásticos. (ver Maroco, (2003) e Pestana e Gageiro, (2005b))

- COLINEARIDADE

Como se trata de uma análise de regressão linear múltipla, um dos pressupostos que terá de ser verificado é se existe colinearidade entre as duas variáveis independentes.

Tabela 5.14- Diagnóstico da colinearidade

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Triglicéridos	HDL
1	1	1,480	1,000	,26	,26	
	2	,520	1,686	,74	,74	
2	1	2,249	1,000	,04	,08	,04
	2	,617	1,909	,05	,92	,04
	3	,134	4,095	,91	,00	,92

a. Dependent Variable: CT- Colesterol Total

Dado que para as duas variáveis independentes os valores de $\text{VIF} < 5$, como podemos confirmar pela Tabela 5.11, concluímos que não existem problemas de colinearidade. Esta conclusão pode ser confirmada pelos *Condition Index* na Tabela 5.14, já que estes valores são inferiores a 15 (ver Maroco, 2003).

- OUTLIERS E OBSERVAÇÕES INFLUENTES

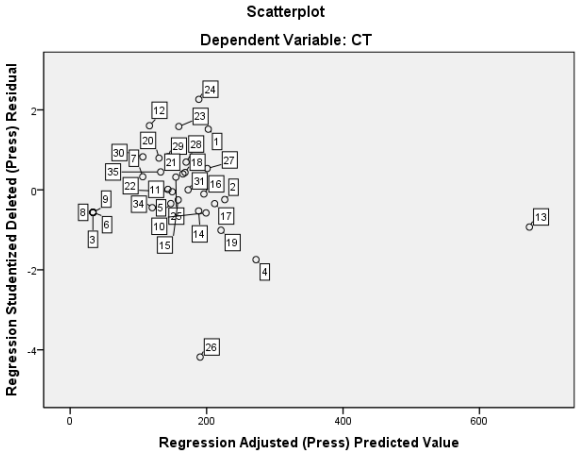

Figura 5.8– Gráfico resíduos *press*

Pela análise gráfica dos resíduos *press*, Figura 5.8, temos que, existem *Outlier*s, dado que apresenta resíduos com valores absolutos superiores a 1,96 (ver Pestana e Gageiro, 2005b). São *Outliers* as observações 4, 24 e 26.

Tabela 5.15- Estatística dos resíduos

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	29,5005	515,2530	154,8857	88,56879	35
Std. Predicted Value	-1,416	4,069	,000	1,000	35
Standard Error of Predicted Value	9,354	52,627	14,193	7,785	35
Adjusted Predicted Value	33,3483	673,1512	160,4467	109,00174	35
Residual	-184,03899	115,37788	,00000	53,46436	35
Std. Residual	-3,339	2,094	,000	,970	35
Stud. Residual	-3,397	2,130	-,028	1,009	35
Deleted Residual	-190,46869	119,42819	-5,56101	63,58802	35
Stud. Deleted Residual	-4,182	2,263	-,044	1,104	35
Mahal. Distance	,008	30,033	1,943	5,064	35
Cook's Distance	,000	3,001	,105	,505	35
Centered Leverage Value	,000	,883	,057	,149	35

a. Dependent Variable: CT- Colesterol Total

A confirmação de existência de *Outlier*s pode ser feita através do valor máximo de *Student Deleted Residual* (2,263>1,96) e do *Leverage* ($0{,}883 > \frac{(2(p+1))}{n} \simeq 0{,}17$). (ver Maroco, 2003)

Olhando ainda para a Tabela 5.15 concluímos que estes *Outlier*s poderão ser influentes uma vez que o valor máximo da distância de $COOK = 3{,}001 > \frac{4}{n-p-1} \simeq 0{,}125$.

Para averiguar se as observações são efectivamente influentes, devemos ainda recorrer à análise dos SDFFIT, pelo que apresentamos de seguida o gráfico destes.

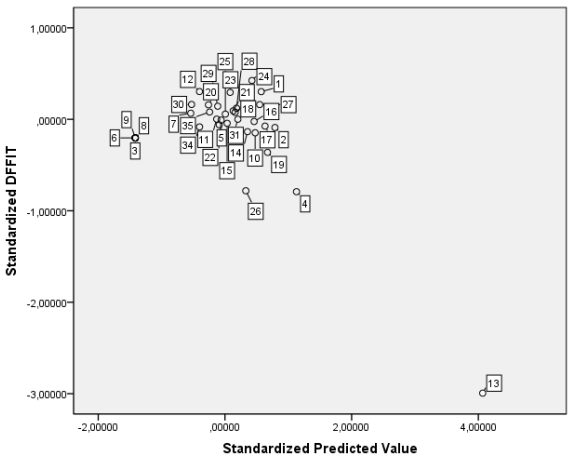

Figura 5.9– Gráfico dos *Standardized* DFFIT

Concluímos então que as observações 4 e 26 são observações influentes, uma vez que $|SDFFIT| > 2\sqrt{\frac{p+1}{n-p-1}} \simeq 0{,}61$. No entanto a observação 24 não satisfaz esta condição, o que significa que não é influente.

Conclusão

O que fazer em situações como esta é uma questão ainda nos tempos actuais colocada por muitos autores da área. Podemos tomar a decisão mais fácil, que passa por excluir esta observação da análise e reestimar de novo o modelo. No entanto há quem defenda que se deve manter este tipo de observações, ver por exemplo, Figueira (1995).

À excepção desta questão todos os restantes pressupostos foram validados. Optando pela não exclusão da observação 24, teríamos o modelo (5.4) como o modelo válido, adequado aos dados.

6. CONCLUSÕES

O presente trabalho teve como principal objectivo aprofundar o conhecimento dos modelos de regressão linear.

Centrámo-nos no entanto no estudo do modelo de regressão linear simples onde procuramos apresentar de forma detalhada os pressupostos do modelo e o método dos mínimos quadrados, que nos leva à obtenção de estimadores dos parâmetros do modelo.

Realizamos ainda inferência para os parâmetros, construindo testes e intervalos de confiança para os mesmos.

Este trabalho não abordou de forma exaustiva o modelo de regressão linear múltipla, pela extensão do tema e pela restrição do tempo, daí termos designado o capítulo referente a este tema como "Breve Abordagem".

Tendo em conta a importância de validação dos pressupostos impostos ao erro do modelo, por forma a concluir se o modelo é adequado, foi apresentado um capítulo sobre a análise de resíduos.

Desenvolvemos alguns estudos, no capítulo das Aplicações, com vista a mostrar a grande aplicabilidade do modelo de regressão linear em diversas áreas, nomeadamente, na área das ciências da saúde, área contemplada nos estudos que apresentámos. Dada a restrição de tempo, não foi considerado neste capítulo o modelo mais complexo da regressão linear múltipla, onde teríamos de inserir a interacção entre as variáveis. A partir da investigação realizada neste último capítulo consideramos importante a apresentação de algumas considerações finais.

Podemos afirmar que existe uma relação entre os níveis (mg/dl) no sangue de CT e HDL, triglicéridos. Constatámos que uma outra variável, peso, não influencia significativamente os níveis de CT. Este facto corrobora a existência de pessoas que, mesmo com excesso de peso, não têm níveis elevados de CT no sangue, e o contrário, pessoas magras, podem "sofrer" de colesterol total elevado.

Embora a variabilidade da variável CT não seja inteiramente explicada pelos níveis de HDL e triglicéridos, existindo outros factores que não foram identificados neste estudo, podemos afirmar que, em média, um aumento das variáveis HDL e triglicéridos provoca um aumento de CT.

BIBLIOGRAFIA

[1]. Afonso, A., Nunes, C.; Probabilidades e Estatística – Aplicações e soluções em SPSS; Escolar Editora; 2011.

[2]. Branco, J.A. e Pires, A.M.; Introdução aos Métodos Estatísticos Robustos; Edições SPE; 2007.

[3]. Cordeiro, N., Magalhães, A.; Introdução à Estatística; Lidel; 2004.

[4]. Cunha, G, Martins, M.R, Sousa, R., Oliveira, F.F.; Estatística Aplicada às Ciências e Tecnologias da Saúde; Lidel; 2007.

[5]. Draper, N.R., Smith, H.; Applied Regression Analisys, 3ª edição, John Wiley and Sons; 1998.

[6]. Edwards, A. L.; An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman, 1976.

[7]. Esteves, E.& Sousa, C.; Apontamentos de ADPE; UALG; 2007.

[8]. Figueira, M.M.C. (1995): "Identificação de outliers: uma aplicação ao conjunto das maiores empresas com actividade em Portugal" Tese de Mestrado – Instituto Superior de Economia e Gestão.

[9]. Franco, S. C. A.; Comportamento pedagógico dos instrutores de fitness em aulas de grupo localizada; Ph.D. Thesis, INEFC; 2009.

[10]. Guimarães, R. C., Cabral, J. A. S.; Estatística; McGraw-Hill; 1997.

[11]. Marinho, J. L. A.; Proposta de um modelo para avaliação de imóveis urbanos da Região de Cariri utilizando variáveis sócio-econômicas; Fortaleza 2007; Available from pt.scribd.com/doc/56277599/74/coeficiente-de-determinacao.

[12]. Mexia, J.T.; Introdução à Inferência Estatística Linear; Centro de Estudos de Matemática Aplicada; Edições Lusófonas; 1995

[13]. Matos, M.A.; Manual operacional para a regressão linear; FEUP; 1995.

[14]. Murteira, B., Ribeiro, C., Silva, J. A., Pimenta, C.; Introdução à estatística; Escolar editora; 2010.

[15]. Maroco, J.; Análise Estatística – Com utilização do SPSS, 2ª edição; Edições Sílabo; 2003.

[16]. Murteira, B.J.F.; Probabilidades e Estatística, volume I, 2ª edição; McGraw-Hill; 1998.

[17]. Pagano, M. & Gauvreau, K; Princípios de Bioestatística; Pioneira Thomson Learning; 2004.

[18]. Pestana, D.D., Velosa, S.F.; Introdução à Probabilidade Estatística – Volume I; Fundação Calouste Gulbenkian; 2002.

[19]. Pestana, M. H. e Gageiro, J.N.; Análise de Dados para Ciências Sociais: A Complementaridade do SPSS. 4ª ed., Lisboa: Sílabo; 2005a.

[20]. Pestana, M. H. & Gageiro, J. N.; Descobrindo a Regressão – Com a Complementaridade do SPSS. Edições Sílabo; 2005b.

[21]. Pereira, A.; Guia prático de utilização do SPSS – Análise de dados para ciências sociais e psicologia, 6ª edição; Edições Sílabo; 2006.

[22]. Portal Action; Copyright 1997-2011 Estatcamp; Análise de regressão; Available from: www.portalaction.com.br

[23]. Santos, C. M. A.; Estatística Descritiva – Manual de auto-aprendizagem; Edições Sílabo; 2007.

[24]. Spiegel, M.R.; Probabilidade Estatística; McGraw-Hill; 1977.

[25]. Weisberg, S.; Applied Linear Regression (Wiley Series in Probability and Statistics), John Wiley & Sons, Inc.; 3th Edition, 2005.

[26]. Werkena, C e Aguiar, S.; Análise de Regressão: como entender o relacionamento entre as variáveis de um processo; Werkena Editora; 1996.

[27]. Wikipedia, A enciclopédia livre; Regressão linear; Available from: pt.wikipedia.org/wiki/Regressão_linear.

ANEXOS

Anexo 1- Tabela de valores críticos de d_L e d_U, (disponível em Maroco (2003))

Tabela de valores críticos de d_L e d_U, (disponível em Maroco (2003)).

	para $\alpha = 0.05$													
p	1		2		3		4		5		10		15	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.61	1.40												
10	0.88	1.32	0.70	1.64	0.53	2.02	0.38	2.41	0.24	2.82				
20	1.20	1.41	1.10	1.54	1.00	1.68	0.89	1.83	0.79	1.99	0.34	2.89	0.06	3.68
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83	0.71	2.36	0.39	2.94
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79	0.95	2.15	0.68	2.56
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77	1.11	2.05	0.88	2.35
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77	1.22	1.98	1.03	2.28
70	1.58	1.64	1.55	1.67	1.53	1.70	1.49	1.74	1.46	1.77	1.31	1.95	1.14	2.15
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77	1.37	1.93	1.22	2.09
90	1.64	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78	1.42	1.91	1.29	2.06
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78	1.46	1.90	1.35	2.03
200	1.76	1.78	1.75	1.79	1.74	1.8	1.73	1.81	1.72	1.82	1.67	1.87	1.61	1.93

Anexo 2 - Avaliação da variabilidade da frequência cardíaca

Avaliação da variabilidade da frequência cardíaca

Caracterização de variáveis demográficas: sexo, raça, data de nascimento

Caracterização de variáveis clínicas: peso, altura, TAS, TAD, DM, dislipidémia, doença coronária, doença cerebrovascular, tabagismo, actividade física, história familiar de qualquer uma das doenças referidas anteriormente

Terapêutica anti-hipertensora: ACC, beta-bloqueantes, IECA/ARA, diurético, associação

Avaliação analítica: colesterol total, HDL, triglicéridos, glicemia, creatinina

Avaliação de lesão de órgão-alvo: HVE, microalbuminúria

Parâmetros de variabilidade da frequência cardíaca: medidas de "time-domain" e frequência

Análise dos resultados

- caracterização da variabilidade da frequência cardíaca por géneros, classe etária e grupos patológicos

- avaliação a longo prazo de eventos cardiovasculares (EAM, AVC, mortalidade, …)

AVALIAÇÃO DA VARIABILIDADE DA FREQUÊNCIA CARDÍACA

Nome: nº processo:

Raça: data de nascimento:

Peso: Altura:

Variáveis clínicas:

Diabetes ☐ Dça cerebrovascular ☐ História familiar ☐

Terapêutica anti-hipertensora

☐ ☐ ☐ ☐ ☐

Avaliação analítica:

Lesão-orgão alvo:

Parâmetros de variabilidade da frequência cardíaca:

SDNN: _____ HF: _____

Anexo 3 – Variabilidade da frequência cardíaca (FC) – Base de dados

Variáveis demográficas			
nome	nº processo	raça	data de nascimento
	9043085	caucasiano	04-09-1968
	9047058	caucasiano	15-02-1944
	9050121	caucasiano	24-09-1948
	5144834	negra	11-09-1949
		caucasiano	
		caucasiano	
		caucasiano	
		caucasiano	17-01-1946
		negra	21-05-1989
		caucasiano	02-01-1932
		caucasiano	04-10-1959
	9007759	caucasiano	19-03-1971
	5099458	caucasiano	27-11-1940
	8027480	caucasiano	15-07-1962
	8027228	caucasiano	14-10-1961
	99018573	caucasiano	02-10-1948
	8070668	negra	10-03-1951
	98042393	caucasiano	16-09-1948
	97050698	caucasiano	08-06-1965
	7283618	caucasiano	15-11-1957
	7302865	caucasiano	18-04-1978
	5122080	caucasiano	29-09-1947
	8014858	caucasiano	15-12-1948
	9044483	caucasiano	04-09-1957
	4017870	caucasiano	20-01-1949
		caucasiano	24-10-1946
	8038698	caucasiano	31-12-1948
		caucasiano	01-03-1944
	8016803	caucasiano	26-02-1952
		caucasiano	29-09-1982
	8095648	caucasiano	06-04-1966
		caucasiano	
		caucasiano	23-04-1944
	8022849	caucasiano	12-12-1937
	3001906	caucasiano	26-01-1937

Variáveis clínicas										
peso	altura	HTA	DM 2	Dislipidémia	CAD	DVC	Tabagismo	Act. Física	Hx familiar	Apneia sono
69	162	0	1	1	0	0	0	0	1	0
54	150	0	0	1	0	0	0	0	1	0
67	155	1	0	1	0	0	0	0	1	0
73	165	0	1	1	0	0	0	0	1	0
97	171	1	1	1	0	0	1	0	1	1
80	162	0	1	0	0	0	1	1	1	0
66	154	0	1	0	1	0	0	1	1	0
89,5	177	1	1	1	0	0	1	0	1	0
52	161	0	1	1	0	0	0	0	0	0
85	168	1	1	1	0	1	1	0	1	0
100	179	1	1	1	0	0	1	0	1	0
105	160	1	0	0	0	0	1	0	1	0
65	156	1	1	1	1	0	0	0	1	0
58	162	1	0	0	0	0	0	1	1	0
60	158	0	0	1	0	0	1	0	0	0
78	170	1	1	1	0	0	0	0	0	0
97	174	1	1	1	0	0	1	0	1	0
82	160	1	0	0	0	0	1	1	0	0
52	155	1	0	1	0	0	0	0	1	0
91	165	0	1	1	0	0	0	0	1	0
90	184	0	0	1	0	0	0	0	1	0
67	161	1	1	1	0	0	0	0	1	0
59	150	1	1	1	0	0	0	1	1	0
75	166	1	1	1	0	0	1	0	1	0
		1	1	1	0	1	1	1	0	0
108	172	1	1	1	0	0	1	0	1	1
64	155	1	1	1	0	0	0	0	1	0
73	168	1	1	1	1	0	1	0	0	0
64	161	0	1	1	0	0	0	1	1	0
51	157	0	1	0	0	0	0	0	0	0
108	160	1	1	1	0	0	0	0	0	0
55	146	1	1	1	0	0	0	0	1	0
74	154	1	1	1	0	0	0	0	0	0
87	165	1	1	1	1	0	0	1	1	0
89	171	1	1	1	0	0	1	0	1	0

Terapêutica anti-hipertensora					Avaliação analítica						
ACC	BB	IECA/ARA	Diurético	Associação	glicemia	creatinina	CT	HDL	Triglic	HbA1c	ác.úrico
0	0	0	0	0	129	0,8	286	42	297		
0	0	0	0	0	95	0,8	212	66	63		
0	0	0	0	0							
1	0	1	1	0	162	0,7	170	75	89	8,8	
1	0	1	1	0	153	1,2	128	35	102	5,4	7,6
0	0	0	0	0							
0	1	0	0	0	162	1,1	125	22	83	8,1	3
0	0	0	0	1	140	1,4					
0	0	0	0	0							
1	0	1	1	0	114	1,4	166	53	104		5,4
1	0	1	1	0	146	0,9	144	31	139		4,9
1	0	0	1	0	100	0,7	204	25	102		
1	0	1	1	0	142	0,7	500	45	1790	7	4,5
0	0	1	0	0	75	0,9	158	52	65		4,2
0	0	0	0	0	100	0,9	144	33	184		4,8
0	0	0	0	1	122	1,1	190	50	138	5,9	7,1
0	0	0	1	0	243	1,7	192	38	378		
0	0	1	0	0	96	0,7	193	38	171		4,3
0	1	0	0	0	91	0,6	162	64	39		
0	0	0	0	0	108	0,8	175	23	192		4,7
0	0	0	0	0	100	1,1	188	47	34		5,2
1	0	1	1	0	209	0,9	147	38	74	10	4,9
1	0	0	0	1	114	0,9	246	40	109		
0	0	1	0	0	119	0,8	308	40	260		
0	0	0	0	1	131	1,3	173	30	213		
1	0	0	0	1	220	1		35	286	9,4	5,2
1	0	1	1	0	171	0,7	232	58	67	8,5	
0	1	1	1	0	122	1,9	209	36	208	9,4	
0	0	0	0	0	129	1,2	190	33	119	6,5	
0	0	0	0	0	157	0,7	153	24	61	8,4	
0	0	1	1	0	92	0,4	173	46	80	6,8	5,4
0	1	1	1	0		0,8					
0	1	0	1	1	140	1,1	95	26	89	10,9	6,2
0	0	0	0	1	222	1,3	158	31	90	7,9	5,8

Lesão orgão-alvo		Variabilidade da frequência cardíaca															
HVE	microalbuminúria	RRI	SDNN	RMSSD	NN50	pNN50	SD1	SD2	SD1/SD2	RRI	SDNN	RMSSD	NN50	pNN50	SD1	SD2	SD1/SD2
		788	43,36	14,43	0	0	10,2	60,46	0,17		27,54	14,44	0	0	10,21	37,58	0,27
1		917	33,37	18,81	3	1,03	13,3	45,27	0,29		34,39	18,84	3	1,03	13,31	46,78	0,28
		933,14	66,1	53,07	108	35,64	37,52	85,59	0,44		62,45	53,07	108	35,64	37,52	79,94	0,47
		718,6	22,81	16,82	3	0,72	11,89	29,99	0,4		22,64	16,82	3	0,72	11,89	29,72	0,4
		976	62,93	53,4	29	12,24	37,75	80,59	0,47		62,94	53,38	29	12,24	37,74	80,62	0,47
		930	23,25	17,99	5	1,56	12,72	30,32	0,42		22,25	17,99	5	1,56	12,72	28,78	0,44
0	1	754,6	12,52	11,34	0	0	8,01	15,78	0,51		12,23	11,35	0	0	8,02	15,32	0,52
0		1003	28,05	26,83	12	4,18	18,97	34,83	0,54		27,22	26,84	12	4,18	18,98	33,5	0,57
		736,14	14,62	15,91	2	0,5	11,25	17,34	0,65		13,6	15,92	2	0,5	11,25	15,6	0,72
		750,38	12,97	9,96	0	0	7,04	16,93	0,42		12,66	9,96	0	0	7,04	16,46	0,43
1	0	562,9	14,47	5,5	0	0	3,89	20,09	0,19		12,62	5,5	0	0	3,89	17,42	0,22
0	1	723,28	17,94	9,56	0	0	6,76	24,45	0,28		15,49	9,56	0	0	6,76	20,84	0,32
		823,31	40	17,35	2	0,56	12,27	55,23	0,22		21,33	17,35	5	0,56	12,27	27,56	0,45
		1012	48,45	51,25	102	34,69	36,24	58,14	0,62		43,73	51,3	102	34,69	36,27	50,09	0,72
		874	30,43	18,31	5	1,47	12,94	41,03	0,32		29,89	18,3	5	1,47	12,94	40,24	0,32
		679	10,76	4,96	0	0	3,51	14,8	0,24		8,52	4,94	0	0	3,49	11,53	0,3
		858	23,52	22,43	5	1,46	15,86	29,23	0,54		23,43	22,44	5	1,46	15,86	29,08	0,55
		1028	64,05	63,65	97	36,47	44,85	78,7	0,57		63,15	63,67	97	36,47	44,87	77,22	0,58
		744	24,32	17,03	0	0	12,04	32,22	0,37		23,88	17,04	0	0	12,05	31,55	0,38
		975,5	25,64	19,57	0	0	13,83	33,52	0,41		25,24	19,56	0	0	13,83	32,9	0,42
		774	25,03	15,37	1	0,26	10,86	33,69	0,32		24,46	15,38	1	0,26	10,88	32,83	0,33
1	1	819	28,98	14,4	1	0,27	10,18	39,7	0,26		26,42	14,38	1	0,27	10,17	35,95	0,28
		1017	47,65	35,73	45	15,36	25,26	62,47	0,4		45,83	35,75	45	15,36	25,28	59,68	0,42
		847	24,81	17,63	2	0,69	12,45	32,8	0,38		25,83	17,65	2	0,69	12,46	34,34	0,36
1		1165	46,39	27,85	21	8,24	19,69	62,58	0,31		42,02	27,86	21	8,24	19,7	56,06	0,35
		621	9,93	8,62	0	0	6,09	12,66	0,48		9,44	8,61	0	0	6,09	11,88	0,51
0	1	832	21,14	16,63	1	0,28	11,76	27,49	0,43		17,73	16,62	1	0,28	11,75	22,14	0,53
		1087	14,45	17,13	0	0	12,11	16,45	0,74		13,74	17,14	0	0	12,12	15,19	0,8
	1	799	35,22	21,61	7	1,88	15,28	47,4	0,32		29,32	21,58	7	1,88	15,26	38,55	0,4
		766	33,87	13,96	2	0,51	9,87	46,87	0,21		26,15	13,97	2	0,51	9,88	35,64	0,28
0	1	729,87	29,42	18,93	2	0,49	13,39	33,39	0,34		23,37	18,97	2	0,49	13,41	30,2	0,44
		788	108,23	82,44	54	39,13	58,24	141,54	0,41		107	82,36	54	39,13	58,19	139,69	0,42
		796	24,65	10,81	0	0	7,64	34,02	0,22		23,06	10,81	0	0	7,64	31,7	0,24
		577	7,35	9,2	1	0,19	6,51	8,1	0,8		7,19	9,2	1	0,19	6,5	7,81	0,83
		666,7	9,67	4,22	0	0	2,99	13,34	0,22		8,15	4,23	0	0	2,99	11,13	0,27

VLF NON-DETREND		VLF DETREND		LF				HF			
FFT	AR	FFT	AR	FFT	AR	FFT	AR	FFT	AR	FFT	AR
718	1145	174	187	171	162	144	142	45	43	38	38
382	467	351	502	138	137	165	153	58	61	66	69
593	654	372	375	950	961	949	965	485	494	486	493
85	94	80	89	91	86	92	86	70	74	70	74
441	552	309	370	386	371	396	373	652	636	641	635
80	91	52	61	120	119	123	121	54	56	54	56
21	23	17	19	32	31	31	31	13	13	13	13
160	159	134	136	98	111	100	110	101	106	103	108
39	42	24	25	20	21	21	20	15	14	14	13
44	42	40	37	19	23	19	22	25	25	25	25
81	98	54	66	16	15	19	15	6	6	6	6
89	105	50	53	34	36	32	33	32	33	31	32
634	1266	98	100	79	78	61	67	58	59	51	51
335	462	129	143	306	312	291	299	505	498	488	479
199	216	182	190	189	211	188	211	51	48	49	47
46	51	24	24	7	9	7	8	2	2	2	2
119	119	115	115	61	73	62	73	64	67	64	67
603	725	534	645	599	588	603	590	718	698	724	702
134	146	126	138	98	102	99	104	50	49	50	50
179	205	168	196	87	88	89	88	54	55	55	55
114	122	104	103	115	132	113	130	69	65	67	64
239	275	172	189	100	101	100	101	57	57	57	57
666	715	589	629	219	216	216	214	202	205	202	208
105	122	129	151	168	162	165	161	81	78	82	79
644	812	467	537	282	289	269	280	116	103	110	102
21	23	16	17	10	10	10	10	14	14	14	14
117	135	51	57	30	29	32	29	59	58	60	58
36	40	24	28	28	27	31	28	9	10	10	10
347	411	161	165	216	214	195	209	56	57	56	57
447	544	217	229	74	86	77	79	35	33	32	32
248	302	83	100	96	93	101	93	70	69	73	70
668	993	471	720	1630	1737	1648	1740	2919	2843	2918	2849
182	285	157	218	88	80	74	74	34	27	29	25
11	13	10	12	4	4	4	4	8	7	7	7
35	45	23	25	6	6	4	5	3	2	2	2

Total power				LF/HF			
FFT	AR	FFT	AR	FFT	AR	FFT	AR
935	1351	358	368	3,78	3,74	3,76	3,66
579	666	584	725	2,36	2,21	2,47	2,22
2029	2110	1809	1835	1,96	1,94	1,95	1,96
247	255	243	250	1,29	1,16	1,31	1,17
1479	1561	1347	1379	0,59	0,58	0,62	0,59
255	268	230	239	2,23	2,13	2,25	2,15
66	69	63	65	2,37	2,34	2,36	2,34
361	377	338	354	0,97	1,04	0,97	1,02
75	77	60	60	1,34	1,5	1,45	1,48
89	91	85	86	0,77	0,94	0,76	0,89
104	120	79	88	2,75	2,46	3,09	2,5
155	175	114	119	1,06	1,08	1,02	1,05
773	1404	211	219	1,36	1,31	1,2	1,3
1146	1273	909	922	0,61	0,63	0,6	0,63
440	476	421	449	3,7	4,38	3,82	4,42
56	63	34	35	3,08	3,43	3,07	3,3
246	261	242	256	0,95	1,09	0,96	1,09
1920	2012	1862	1938	0,83	0,84	0,83	0,84
283	298	276	292	1,96	2,07	1,96	2,08
322	349	313	340	1,6	1,61	1,6	1,61
299	319	285	298	1,66	2,03	1,67	2,03
396	434	330	348	1,75	1,76	1,76	1,76
1088	1137	1008	1052	1,09	1,05	1,07	1,03
355	364	377	393	2,06	2,06	2,02	2,02
1042	1205	847	920	2,43	2,8	2,43	2,74
46	47	41	42	0,76	0,74	0,77	0,7
207	223	143	145	0,5	0,51	0,53	0,5
75	78	65	67	2,85	2,75	3,09	2,83
619	683	413	432	3,82	3,73	3,49	3,62
558	663	327	341	2,08	2,58	2,41	2,48
414	466	258	264	1,36	1,34	1,39	1,33
5218	5574	5038	5310	0,56	0,61	0,56	0,61
304	393	261	318	2,57	2,97	2,52	2,89
23	25	22	24	0,53	0,55	0,53	0,55
44	54	31	33	1,98	2,1	1,75	1,99