UNIVERSIDADE DA BEIRA INTERIOR
Ciências

Modelo de Regressão Linear:
"Aplicação ao Estudo sobre os Fatores que Influenciam o Rendimento Académico dos Alunos em Angola"

Ngaiele Muecheno Fundão

Dissertação para obtenção do Grau de Mestre em
Matemática para Professores
(2º ciclo de estudos)

Orientadora: Prof. Doutora Célia Maria Pinto Nunes

Covilhã, maio de 2018

Dedicatória

Dedico este trabalho ao meu querido pai Fundão Ngayele (In memoriam)

Agradecimentos

Começo por agradecer a Deus. Um agradecimento especial vai à minha orientadora Professora Doutora Célia Maria Pinto Nunes pela amabilidade, as sábias lições de vida e os conhecimentos que me transmitiu durante o período de formação, bem como, pelas contribuições que permitiram a materialização deste trabalho.

Agradeço ao Professor Doutor Jorge Gama, Professora Doutora Ana Paula e aos demais professores do departamento de Matemática da UBI pelos conhecimentos e lições de vida que me transmitiram. Agradeço também às Senhoras Filipa Raposo, Ana Silva e Ermelinda Calmão pela delicadeza e ajudas prestadas.

Agradeço à minha Esposa pelo encorajamento e apoio prestado durante a formação e aos meus pais, filhos, irmãos, tal como, aos demais elementos da família por vários motivos.

Agradeço ao Hermenegildo Simão, Orlando Cawende e Jacinto Comolehã pela amizade e camaradagem e aos colegas de curso e amigos pelo encorajamento e companheirismo.

Agradeço ainda ao Conselho de Administração da TPA, ao Senhor Victor Silva, Decano da Escola Superior Politécnica do Moxico e ao Ministério do Ensino Superior de Angola pela confiança depositada em mim e nos meus colegas e à Direção Provincial de Educação Ciência e Tecnologia do Moxico, aos diretores, subdiretores, professores de matemática e alunos do ensino secundário da província do Moxico, pela ajuda e atenção prestadas na recolha de dados.

Para finalizar agradeço às demais pessoas singulares e coletivas que direta ou indiretamente contribuíram para que este trabalho fosse concretizado.

O meu muito obrigado!

Resumo

As dificuldades no entendimento de conteúdos relacionados com a regressão linear e sua aplicação na resolução de problemas do dia a dia em Angola motivaram a realização do presente trabalho. Este, apresenta uma abordagem sobre os diversos aspetos da teoria de regressão linear e sua aplicação na área do ensino. O trabalho começa com uma abordagem introdutória sobre esta metodologia estatística e a problemática do rendimento académico a matemática dos alunos do ensino secundário em Angola. Nos capítulos 2, 3 e 4 apresentam-se os aspetos teóricos essenciais dos modelos de regressão linear simples e múltipla. No capítulo 5 apresenta-se um estudo prático que poderá auxiliar as autoridades educativas da província do Moxico em Angola, e não só, a justificar as notas dos alunos do ensino secundário à disciplina de matemática. O modelo obtido através deste estudo, distingue as notas a língua portuguesa e a física, a idade e a renda familiar média, como alguns dos fatores que influenciam as notas dos alunos a esta disciplina. Finalmente, apresenta-se uma síntese dos principais resultados obtidos no trabalho e aspetos a ter em conta na realização de futuros trabalhos ligados ao tema.

Palavras-chave

Modelo de regressão linear, Rendimento escolar a matemática, Província do Moxico em Angola.

Abstract

The present work was motivated by the difficulties in understanding linear regression analysis and its applications in solving daily life problems in Angola. The work includes a discussion of several aspects of linear regression theory and its applications on the teaching area. The work starts with an introduction to this statistical methodology. The chapters 2, 3 and 4 present the essential theoretical aspects of simple and multiple linear regression models. Chapter 5 presents a real study about the academic performance of students from Moxico province of Angola. It was obtained a multiple linear regression model that will help the education authorities to explain student performance in mathematics. This model distinguishes the classification in Portuguese language and physics, the age and the family budget as some factors that take significant influence in the students' performance in math. Finally, we present a summary of the main results and recommended the aspects to be taken into account in future works related to this subject.

Keywords

Linear regression model, School performance in mathematics, Moxico province of Angola.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

ANOVA	Análise da variância (Analysis of variance)
DGEEC	Direção Geral de Estatísticas de Educação e Ciências
DFBETA	Diferença em BETA (Difference in Beta)
DFFIT	Diferença de ajuste (Difference in Fit)
D-W	Durbin-Watson
K-S	Kolmogorov-Smirnov
MMQ	Método dos mínimos quadrados
MQE	Média quadrática dos resíduos
MQR	Média quadrática dos desvios explicados pela regressão
MRLM	Modelo de regressão linear múltipla
MRLS	Modelo de regressão linear simples
OCDE	Organização para a Cooperação e Desenvolvimento Económico
PISA	Programme for International Student Assessment
RLM	Regressão linear múltipla
RLS	Regressão linear simples
SDFBETA	Diferença em BETA estandardizada (Difference in Standardized Beta)
SDFFIT	Diferença de ajuste estandardizada (Difference in Standardized Fit)
SPSS	Statistical Package for the Social Sciences
SQE	Soma de quadrados dos resíduos
SQR	Soma de quadrados dos desvios explicados pela regressão
SQT	Soma de quadrados dos desvios totais
S-W	Shapiro-Wilk
TIMSS	Trends in International Mathematics and Science Study
UNESCO	Organização das Nações Unidas para a Educação, Ciência e Cultura
VIF	Fator de inflação da variância (Variance Inflation Factor)

Capítulo 1

Introdução

A estatística é um dos ramos da matemática que tem como finalidade observar os fenómenos, recolher informações, analisá-las e apresentar os resultados que possam ajudar na tomada de decisões sobre os mesmos. Atualmente a estatística é uma das ciências mais importante na sociedade, por esta razão é amplamente utilizada em diversas áreas da ciência.

As contribuições que fizeram com que a estatística fosse uma das ciências mais importantes no mundo atual, deveram-se a estudos realizados por investigadores de vários pontos do planeta, que viveram em distintas épocas da história da humanidade. Estes investigadores procuraram sempre dar explicações e prever resultados de diversos fenómenos que estavam associados a uma variedade de problemas que a humanidade vivia e vive até aos nossos dias.

Uma das metodologias estatística amplamente usada com o intuito de explicar ou prever soluções de diversos problemas ligados a várias áreas da vida social, económica, política e não só, é a regressão linear, cuja origem está associada ao antropólogo e matemático Inglês Francis Galton. Este introduziu pela primeira vez o termo "regressão" num estudo intitulado "Regression Towards Mediocrity in Hereditary Stature", que foi publicado no "Journal of the Anthropological Institute of Great Britain and Ireland" em 1886. Neste estudo, Galton mostrou que havia uma tendência nas populações humanas, de os filhos de pais de estatura mais elevada serem em média mais baixos do que os pais, enquanto que os filhos de pais de menor estatura tendiam a ser em média mais altos do que os seus pais; tendo denominado esta tendência de "regressão para a média" (Oliveira, Santos e Fortuna, 2011, p. 19).

Para se efetuarem os cálculos da reta de regressão, acorreu-se ao método dos mínimos quadrados criado pelo matemático Alemão Carl Friedrich Gauss em 1795 e publicado pela primeira vez pelo francês Adrien Marie Legendre em 1806, no seu livro "*Nouvelles méthodes pour la détermination des orbites des comètes*", onde figurava como uma ferramenta matemática que tinha como finalidade reduzir os erros provocados pelas medições astronómicas (Schivani e Sousa, 2015, p. 5).

Hoje este método é denominado "regressão linear" devido ao trabalho de Karl Pearson, um dos seguidores de Francis Galton, que deu contribuições extremamente importantes no desenvolvimento de muitos elementos da teoria de regressão, usados até aos nossos dias (Ferreira e Tavares, 2002, p. 24). Desde os tempos de Galton e Pearson até aos nossos dias, a regressão linear tem vindo a receber inúmeras contribuições, com a criação e incorporação de diversas técnicas estatísticas, que têm vindo a contribuir para a melhoria gradual do método, fazendo com que os resultados das análises realizadas com esta metodologia sejam considerados viáveis em muitas áreas do saber.

Na prática, a aplicação da regressão linear é feita com base nos chamados "modelos de regressão linear". Estes podem ser classificados segundo o número de variáveis independentes que possuem. Assim, quando o modelo possui apenas uma variável independente denomina-se "Modelo de Regressão Linear Simples" e se tiver mais de uma variável independente denomina-se "Modelo de Regressão Linear Múltipla".

A regressão linear, embora seja considerada atualmente em alguns círculos como uma metodologia de análise estatística antiga, fruto dos seus longos anos de utilização na modelação de problemas estatísticos, ainda continua a ter bastante utilidade e uma ampla utilização por parte dos especialistas ligados à estatística em todo o mundo. Por isso mesmo, o seu domínio afigura-se importante por parte dos estudantes, investigadores e profissionais que têm a necessidade de usar a estatística como uma ferramenta fundamental nos seus estudos e investigações.

1.1 Justificação do estudo

Apesar dos seus longos anos de utilização pelo mundo fora, em Angola a regressão linear ainda é muito pouco difundida no meio académico e não só. Como resultado, são muitos os estudantes que chegam ou até mesmo terminam os estudos secundários e universitários, sem terem abordado em algum ano escolar, conteúdos ligados a esta metodologia estatística.

Este trabalho propõe efetuar uma abordagem sobre os aspetos teóricos essenciais que conformam a teoria da regressão linear e mostrar como esta pode ser aplicada à situações reais, através de um estudo sobre o rendimento académico dos alunos do ensino secundário da província do Moxico[1] em Angola, à disciplina de Matemática.

Para além da abordagem dos aspetos mais relevantes ligados à teoria da regressão linear e suas aplicações, este estudo visa despertar a população estudantil, professores, dirigentes escolares e outras entidades angolanas, a necessidade de se efetuarem estudos que ajudem a resolver a problemática do rendimento escolar dos alunos nos mais variados níveis de ensino.

A razão que justifica a realização do estudo sobre o rendimento académico dos alunos a matemática, prende-se com as constantes preocupações demonstradas por parte de muitos dos alunos do ensino secundário e pelas autoridades educativas angolanas, sobre o que devem fazer para que se registem melhorias no rendimento académico a esta disciplina.

[1] Moxico é uma das 18 províncias de Angola. É a maior do país em extensão territorial, ocupando uma área de 223 023 quilómetros quadrados. Em termos populacionais, os resultados do censo realizado em 2014 no país indicam que a província possuía na altura cerca de 758 500 habitantes. Segundo as autoridades locais do estado, a província tinha cerca de 285 600 alunos matriculados no ensino básico e secundário em 2017.

1.2 Problema

Apoiando-se na argumentação exposta anteriormente levanta-se a seguinte questão: Como identificar os fatores que influenciam o rendimento académico dos alunos do ensino secundário da província do Moxico em Angola, à disciplina de matemática?

1.3 Objetivos do trabalho

1.3.1 Objetivo geral

- Obter um modelo de regressão linear que possa identificar os fatores que influenciam as notas dos alunos do ensino secundário da província do Moxico em Angola, à disciplina de matemática e proporcionar às autoridades educativas de Angola, uma ferramenta que as ajude a explicar o rendimento académico dos alunos do ensino secundário a matemática.

1.3.2 Objetivos específicos

- Realizar uma abordagem sobre os aspetos teóricos essenciais dos modelos de regressão linear simples e múltipla;
- Obter informações sociodemográficas e socioeconómicas que ajudem a caracterizar os alunos do ensino secundário da província do Moxico em Angola.
- Recorrer ao modelo de regressão linear, por forma a identificar os fatores que influenciam o rendimento académico dos alunos do ensino secundário da província do Moxico em Angola, à disciplina de matemática.

1.4 Resultados esperados

Espera-se que este trabalho possa ajudar os alunos do ensino secundário e universitário e as demais pessoas que estejam interessadas em aprender esta metodologia estatística, a interpretar e compreender os aspetos teóricos essenciais dos modelos de regressão linear simples e múltipla. Por outro lado, através do estudo prático espera-se também obter resultados que ajudem a caracterizar os alunos do ensino secundário da província do Moxico em Angola e a identificar os fatores que influenciam as notas a matemática destes alunos. Pretende-se ainda proporcionar às autoridades educativas de Angola uma ferramenta, que as auxilie a explicar o rendimento académico dos alunos a esta disciplina.

1.5 Estrutura do trabalho

O presente trabalho possui 6 capítulos. O Capítulo 1 corresponde à introdução onde são apresentados os objetivos do presente trabalho e outros aspetos associados ao tema. No Capítulo 2 faz-se uma abordagem sobre o modelo de regressão linear simples (MRLS), descrevendo os aspetos essenciais ligados a este tipo de modelo, tais como, a equação e pressupostos do modelo, a estimação dos parâmetros, a obtenção dos intervalos de confiança, a realização dos testes de hipóteses para os parâmetros, entre outros. No Capítulo 3 apresenta-se uma abordagem sobre o modelo de regressão linear múltipla (MRLM), focando os aspetos já mencionados para o MRLS e outros que apenas se consideram neste modelo, como é o caso dos métodos de seleção das variáveis independentes e a inserção de variáveis qualitativas no modelo. Já no Capítulo 4 efetua-se uma abordagem sobre a validação dos pressupostos impostos ao modelo de regressão. O Capítulo 5 traz a abordagem prática relacionada com o estudo sobre os fatores que influenciam o rendimento académico dos alunos do ensino secundário da província do Moxico em Angola, à disciplina de Matemática. Nas considerações finais do trabalho apresentam-se os principais resultados obtidos com a realização do mesmo e algumas recomendações a ter em conta na eventualidade de no futuro haver a possibilidade de se realizarem mais trabalhos relacionados com este tema.

Capítulo 2

Modelo de Regressão Linear Simples - MRLS

O Modelo de Regressão Linear Simples (MRLS) é uma metodologia estatística que permite estabelecer uma relação linear entre duas variáveis, usualmente representas por X e Y, ambas de natureza quantitativa, onde Y representa a variável dependente (ou variável resposta), cujo valor se pretende explicar e X representa a variável independente (ou variável explicativa), cujo valor é conhecido.

Este tipo de modelo é aplicado aos estudos que visam explicar ou efetuar previsões de vários fenómenos ligados aos diversos ramos da vida social, económica e não só, onde a explicação ou previsão dos mesmos depende apenas de um único fator, que desde já deve ser observável.

2.1 Equação do modelo de regressão linear simples

O MRLS é dado através da seguinte equação:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \cdots, n, \tag{2.1}$$

onde
y_i - representa a i-ésima observação da variável dependente;
x_i - representa a i-ésima observação da variável independente;
β_0 e β_1 - representam os parâmetros do modelo, também designados como coeficientes de regressão; sendo que β_0 representa a interceção da reta de regressão com o eixo vertical, quando $x = 0,$ e β_1 representa o declive (coeficiente angular) da reta e mede o efeito de x_i em y_i;
ε_i - variável aleatória residual (erro) que descreve os efeitos em y_i, não explicados por x_i;
n - é o número de observações.

2.2 Pressupostos do modelo

Como se pode notar na expressão (2.1), para além de y_i e x_i e dos parâmetros β_0 e β_1 aparece também a variável aleatória residual (erro) ε que representa outros fatores que influênciam a variável dependente y_i, não observáveis em x_i. A existência desta variável na equação do

modelo leva a teoria de regressão a considerar um conjunto de pressupostos (hipóteses), cujo cumprimento deve ser obrigatório, para que o modelo de regressão obtido seja considerado válido.

Assim, um MRLS para que seja considerado válido deve observar os seguintes pressupostos:

1. Linearidade do fenómeno em análise, isto é, deve existir uma relação linear entre as variáveis em análise.

2. Os resíduos das observações devem ser mutuamente independentes, isto é, não devem estar autocorrelacionados.

3. A variável residual deve ter um valor esperado nulo, $E(\varepsilon_i) = 0$, logo tem-se:

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

   Por outro lado, uma vez que os resíduos são independentes,

$$cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) - \underbrace{E(\varepsilon_i)}_{=0}\underbrace{E(\varepsilon_j)}_{=0} = E(\varepsilon_i, \varepsilon_j) = 0, \text{ para } i \neq j \text{ e } i, j = 1, 2, \cdots, n.$$

4. Os resíduos devem ser homoscedásticos, o que significa que a variância dos resíduos deve ser constante, $var(\varepsilon_i) = \sigma^2,\ i = 1, 2, \cdots, n$. Logo

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

   assim,

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

5. Os resíduos devem ter distribuição normal, com média zero e variância constante σ^2,

$$\varepsilon_i \sim N\left(0, \sigma^2\right),\ i = 1, \cdots, n.$$

   Por outro lado, para cada valor fixo da variável independente, a variável dependente deve apresentar uma distribuição normal com média $\beta_0 + \beta_1 x_i$ e variância constante σ^2, ou seja, $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right),\ i = 1, \cdots, n.$

Os valores dos resíduos para além de serem utilizados na verificação dos pressupostos impostos ao erro do modelo, também são usados na estimação dos parâmetros.

2.3 Estimação dos parâmetros do modelo

Como em muitos estudos estatísticos a totalidade da população não está acessível, frequentemente recorre-se a uma amostra de n observações das variáveis x_i e y_i, que contém informações de uma parte da população, para estimar os parâmetros β_0 e β_1 do modelo.

Na regressão linear, não se pode afirmar que dados os valores de x_i, se obtêm os valores de y_i, porque como já vimos, a equação do modelo $(y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\, i = 1, 2, \cdots, n)$ possui a variável aleatória erro ε_i que representa quantidades aleatórias desconhecidas. O que se pode afirmar é que o valor esperado de y_i (que não pode ser confundido com a média de y_i) é dado por (Marôco, 2014),

$$E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i = \widehat{y}_i$$

o que significa que se tem,

$$\widehat{y}_i = \beta_0 + \beta_1 x_i, \tag{2.2}$$

devido às propriedades do valor esperado, já que os valores correspondentes a β_0, β_1 e x_i, $i = 1, 2, \cdots, n$ são constantes e $E(\varepsilon_i) = 0$.

Com este resultado, pode-se estimar o valor de ε_i através da expressão

$$\varepsilon_i = y_i - \widehat{y}_i,\ i = 1, 2, \cdots, n. \tag{2.3}$$

Os valores das estimativas dos resíduos para além de quantificarem a margem de erro do modelo, são usados para estimar os coeficientes de regressão e validar os pressupostos impostos ao erro do modelo.

Assim, os parâmetros β_0 e β_1 do modelo de regressão são estimados por meio dos pontos experimentais obtidos através da amostra, resultando a equação da reta estimada representada pela expressão:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \tag{2.4}$$

onde $\widehat{\beta}_0$ é a estimativa do parâmetro β_0, $\widehat{\beta}_1$ é a estimativa do parâmetro β_1 e $\widehat{y}_i$ representa os valores obtidos pela reta estimada, através do Método dos Mínimos Quadrados.

2.3.1 Método dos mínimos quadrados

O método dos mínimos quadrados (MMQ) permite obter os estimadores dos parâmetros β_0 e β_1 da função de regressão, através da soma dos quadrados dos resíduos $\left[SQE = \sum_{i=1}^{n} \varepsilon_i^2\right]$.

Para cada valor de x_i haverá uma diferença entre y_i e o valor determinado pela equação da reta estimada $\widehat{y}_i$. Esta diferença corresponde ao erro (ε_i), (ver expressão 2.3). A mesma diferença pode ser positiva, negativa ou ainda nula (caso a observação i esteja sobre a reta).

Em termos gráficos estes desvios correspondem à distância vertical entre a reta estimada e as observações y_i, como mostra a figura a seguir.

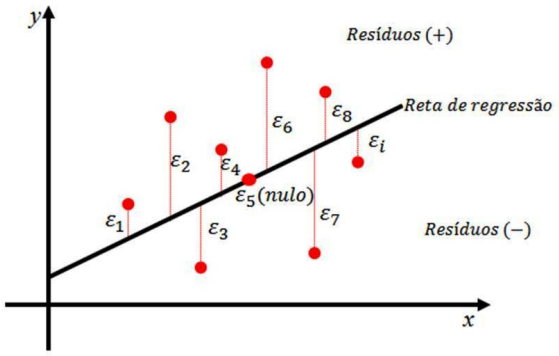

Figura 2.1: Distribuição dos resíduos em relação à reta de regressão

De todas as retas que se ajustam a um conjunto de pontos, a que tem a propriedade de apresentar o desvio mínimo dos pontos em relação à reta estimada, é a reta dos mínimos quadrados, também conhecida como a melhor reta de ajustamento.

Deste modo, o problema de estimação de β_0 e β_1, passa a resumir-se a um problema de determinação do mínimo da função $SQE = \sum_{i=1}^{n} \varepsilon_i^2$. Assim, tem-se:

$$SQE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2. \tag{2.5}$$

Os valores $\widehat{\beta}_0$ e $\widehat{\beta}_1$ que minimizam essa expressão serão aqueles que anulam as suas derivadas parciais em ordem a $\widehat{\beta}_0$ e $\widehat{\beta}_1$:

$$\begin{cases} \frac{\partial SQE}{\partial \widehat{\beta}_0} = 0 \\ \\ \frac{\partial SQE}{\partial \widehat{\beta}_1} = 0 \end{cases} \tag{2.6}$$

Tem-se

$$\begin{aligned} \frac{\partial SQE}{\partial \widehat{\beta}_0} &= \sum_{i=1}^{n} 2 \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \times (-1) \\ &= -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \end{aligned}$$

e

$$\frac{\partial SQE}{\partial \widehat{\beta}_1} = \sum_{i=1}^{n} 2 \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \times (-x_i)$$

$$= -2\sum_{i=1}^{n} x_i \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i).$$

De (2.6) obtém-se as *Equações Normais de Mínimos Quadrados*:

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{2.7}$$

Assim,

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{y} - \widehat{\beta}_1 \overline{x}, \tag{2.8}$$

com

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{2.9}$$

e

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.10}$$

Tem-se ainda

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

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

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

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

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

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{2.11}$$

com base nas expressões (2.9) e (2.10) a equação (2.11) pode ser reescrita da seguinte forma:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{yx}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}. \tag{2.12}$$

As equações (2.8) e (2.12) dão-nos os melhores estimadores não enviesados de β_0 e β_1. Por consequência disso, a equação,

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \ i = 1, 2, \cdots, n, \tag{2.13}$$

também nos dá o melhor estimador linear não enviesado de y_i.

2.3.2 Propriedades dos estimadores

Os estimadores $\widehat{\beta}_j, \ j = 0, 1,$ obtidos através do MMQ, devem satisfazer as seguintes propriedades:

1. Serem funções lineares de y_i;

2. Serem estimadores centrados ou não enviesados de $\beta_j, \ j = 0, 1$, isto é, $E(\widehat{\beta}_0) = \beta_0$ e $E(\widehat{\beta}_1) = \beta_1$, como se demonstra abaixo.

   • **Valor esperado de** $\widehat{\beta}_1$

   Como

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

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} - \underbrace{\frac{\sum_{i=1}^{n} (x_i - \overline{x}) \overline{y}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}_{=0}$$

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

obtemos

$$E\left(\widehat{\beta}_1\right) = \frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)E\left(y_i\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$$

$$= \frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)E\left(\beta_0+\beta_1 x_i+\varepsilon_i\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$$

$$= \frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(\beta_0+\beta_1 x_i\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$$

$$= \frac{\beta_0 \overbrace{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)}^{=0} + \beta_1 \sum_{i=1}^{n} x_i\left(x_i-\overline{x}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$$

$$= \frac{\beta_1\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i\overline{x}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$$

$$= \frac{\beta_1\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}.$$

Logo

$$E\left(\widehat{\beta}_1\right) = \frac{\beta_1\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2} = \beta_1 \tag{2.14}$$

• **Valor esperado de $\widehat{\beta}_0$**

Como se pode notar através da equação (2.8),

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \widehat{\beta}_1\sum_{i=1}^{n} x_i\right).$$

Então,

$$E\left(\widehat{\beta}_0\right) = \frac{1}{n}\left(\sum_{i=1}^{n} E\left(y_i\right) - E\left(\widehat{\beta}_1\right)\sum_{i=1}^{n} x_i\right)$$

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

$$= \frac{1}{n}\left(n\beta_0 + \beta_1\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i\right),$$

logo

$$E\left(\widehat{\beta}_0\right) = \beta_0. \tag{2.15}$$

3. Possuírem variância mínima (serem eficientes);

4. Serem estimadores consistentes.

2.4 Estimação da variância residual (σ^2), variância dos estimadores dos parâmetros ($var(\widehat{\beta}_0)$ e $var(\widehat{\beta}_1)$) e covariância ($Cov(\widehat{\beta}_0, \widehat{\beta}_1)$)

2.4.1 Estimação de σ^2

Para se realizar inferência estatística sobre os parâmetros β_0 e β_1 do MRLS é necessário ter-se informação sobre a variância do erro (σ^2). Como esta frequentemente é desconhecida, o que se faz é determinar um estimador pontual não enviesado, que iremos representar por S^2.

No MRLS o referido estimador é dado por:

$$S^2 = \frac{SQE}{n-2} = MQE, \tag{2.16}$$

tal como se demonstra de seguida.

Dado que

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

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

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

como, $\widehat{\beta}_0 = \overline{y}_i - \widehat{\beta}_1 \overline{x}$, obtém-se:

$$SQE = \sum_{i=1}^{n} \left(y_i^2 - 2y_i(\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i) + (\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i)^2 \right).$$

Colocando em evidência $\widehat{\beta}_1$ na expressão $(\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i)^2$ tem-se:

$$SQE = \sum_{i=1}^{n} \left(y_i^2 - 2y_i(\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i) + \left(\overline{y} - \widehat{\beta}_1 (\overline{x} - x_i) \right)^2 \right)$$

$$= \sum_{i=1}^{n} \left(y_i^2 - 2y_i\overline{y} + 2y_i\widehat{\beta}_1\overline{x} - 2y_i\widehat{\beta}_1 x_i + \overline{y}^2 - 2\overline{y}\widehat{\beta}_1(\overline{x} - x_i) + \widehat{\beta}_1^2(\overline{x} - x_i)^2 \right)$$

$$= \sum_{i=1}^{n} \left((y_i^2 - 2y_i\overline{y} + \overline{y}^2) + 2y_i\widehat{\beta}_1\overline{x} - 2y_i\widehat{\beta}_1 x_i - 2\overline{y}\widehat{\beta}_1(\overline{x} - x_i) + \widehat{\beta}_1^2(\overline{x} - x_i)^2 \right).$$

Como $(y_i^2 - 2y_i\overline{y} + \overline{y}^2) = (y_i - \overline{y})^2$ e $(\overline{x} - x_i)^2 = (x_i - \overline{x})^2$ então,

$$SQE = \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\widehat{\beta}_1 \sum_{i=1}^{n}(-y_i\overline{x} + y_i x_i + \overline{y}\overline{x} - x_i\overline{y}) + \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2,$$

e visto que, $(-y_i\overline{x} + y_i x_i + \overline{y}\overline{x} - x_i\overline{y}) = (x_i - \overline{x})(y_i - \overline{y})$, tem-se,

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

$$= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})y_i + \underbrace{2\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})\overline{y}}_{=0} + \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2$$

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

Dado que $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, obtém-se:

$$SQE = \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})y_i + \widehat{\beta}_1 \left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times \sum_{i=1}^{n}(x_i - \overline{x})^2 \right)$$

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

$$= \sum_{i=1}^{n}(y_i - \overline{y})^2 - \widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})y_i. \tag{2.17}$$

Baseando-se na expressão (2.17) e recorrendo às propriedades do valor esperado tem-se,

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

$$= E\left(\sum_{i=1}^{n}(y_i)^2 - n\overline{y}^2 \right) - E\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times \sum_{i=1}^{n}(x_i - \overline{x})y_i \right)$$

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

Pela definição de variância de uma variável aleatória tem-se

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

e

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

Substituindo estes resultados na expressão anterior e visto que, $var(\overline{y}) = \dfrac{\sigma^2}{n}$, temos:

$$E\left(SQE\right) = \sum_{i=1}^{n}\left(\sigma^2 + (\beta_0 + \beta_1 x_i)^2\right) - n\left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 x_i)^2\right) - \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times$$

$$\times \left[var\left(\sum_{i=1}^{n}(x_i - \overline{x})y_i\right) + \left(E\left(\sum_{i=1}^{n}(x_i - \overline{x})y_i\right)\right)^2\right].$$

Como $\sum_{i=1}^{n}\sigma^2 = n\sigma^2$, logo

$$E\left(SQE\right) = n\sigma^2 + \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 x_i)^2 - \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times$$

$$\times \left[var\left(\sum_{i=1}^{n}(x_i - \overline{x})y_i\right) + \left(E\left(\sum_{i=1}^{n}(x_i - \overline{x})y_i\right)\right)^2\right]. \tag{2.18}$$

Visto que

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

e

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

obtém-se:

$$E\left(SQE\right) = n\sigma^2 + \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1\overline{x})^2 - \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times \sum_{i=1}^{n}(x_i - \overline{x})^2 \times \sigma^2 -$$

$$-\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times \left(\sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i)\right)^2$$

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

A partir deste resultudado e tendo em conta que,

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

$$\beta_0 \sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

e

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

obtém-se

$$E\left(SQE\right) = \sigma^2\left(n-2\right) + \sum_{i=1}^{n}(\beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2) - n(\beta_0^2 + 2\beta_0\beta_1\overline{x} + \beta_1^2\overline{x}^2) - \frac{\beta_1^2\left(\sum_{i=1}^{n}(x_i - \overline{x})x_i\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \sigma^2\left(n-2\right) + n\beta_0^2 + 2\beta_0\beta_1\sum_{i=1}^{n}x_i + \beta_1^2\sum_{i=1}^{n}x_i^2 - n\beta_0^2 - 2n\beta_0\beta_1\overline{x} - n\beta_1^2\overline{x}^2 - \beta_1^2\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2.$$

Considerando o facto de que

$$2n\beta_0\beta_1\overline{x} = 2n\beta_0\beta_1\frac{1}{n}\sum_{i=1}^{n}x_i = 2\beta_0\beta_1\sum_{i=1}^{n}x_i;$$

e que

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

colocando em evidência β_1^2 na expressão obtida, isto resulta em:

$$E\left(SQE\right) = \sigma^2\left(n-2\right)+n\beta_0^2+2\beta_0\beta_1\sum_{i=1}^{n}x_i+\beta_1^2\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)-n\beta_0^2-2\beta_0\beta_1\sum_{i=1}^{n}x_i-\beta_1^2\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$$

$$= \sigma^2\left(n-2\right). \tag{2.19}$$

Portanto, obtém-se

$$E\left(S^2\right) = E\left(\frac{SQE}{n-2}\right) = \frac{\sigma^2\left(n-2\right)}{n-2} = \sigma^2.$$

O que significa que

$$S^2 = \frac{SQE}{n-2} = MQE$$

onde S^2 é um estimador centrado de σ^2.

2.4.2 Variância dos estimadores dos parâmetros ($var(\widehat{\beta}_0)$ e $var(\widehat{\beta}_1)$)

• **Variância de $\widehat{\beta}_1$**

Dado que

$$var\left(\widehat{\beta}_1\right) = var\left(\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)y_i}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}\right)$$

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

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

logo

$$var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}. \tag{2.20}$$

• **Variância de $\widehat{\beta}_0$**

Com base na equação (2.8) temos que

$$var\left(\widehat{\beta}_0\right) = var\left(\overline{y} - \widehat{\beta}_1\overline{x}\right)$$

$$= var\left(\overline{y}\right) + \overline{x}^2 var\left(\widehat{\beta}_1\right) - \underbrace{2\overline{x}cov\left(\overline{y}, \widehat{\beta}_1\right)}_{=0}$$

$$= var\left(\overline{y}\right) + \overline{x}^2 var\left(\widehat{\beta}_1\right).$$

Como, $var\left(\overline{y}\right) = var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$ e recorrendo ao resultado da equação (2.20), tem-se

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

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

Logo

$$var\left(\widehat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right). \tag{2.21}$$

2.4.3 Covariância dos estimadores dos parâmetros ($Cov(\widehat{\beta}_0, \widehat{\beta}_1)$)

• **Covariância entre $\widehat{\beta}_0$ e $\widehat{\beta}_1$**

$$cov\left(\widehat{\beta}_0\widehat{\beta}_1\right) = E\left(\widehat{\beta}_0\widehat{\beta}_1\right) - E\left(\widehat{\beta}_0\right)E\left(\widehat{\beta}_1\right)$$

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

$$= E\left[\left(\widehat{\beta}_1\overline{y} - \widehat{\beta}_1^2\overline{x}\right)\right] - \beta_0\beta_1.$$

Como

$$E\left(\widehat{\beta}_1^2\right) = var\left(\widehat{\beta}_1\right) + \left(E\left(\widehat{\beta}_1\right)\right)^2,$$

então

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

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

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

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

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

Uma vez que, $E\left(\varepsilon_i\widehat{\beta}_1\right) = 0$, tem-se,

$$cov\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \frac{-\overline{x}\sigma^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}. \tag{2.22}$$

2.5 Coeficientes de correlação linear e de determinação

Nas secções 2.3 e 2.4 foram apresentados procedimentos associados ao método dos mínimos quadrados que possibilitam obter a reta que melhor expressa a relação linear entre as duas variáveis do MRLS, porém, esta reta não mede o grau de correlação existente entre as duas variáveis.

Para medir o grau de correlação e a qualidade do ajustamento, usam-se medidas como o coeficiente de correlação linear e o coeficiente de determinação.

• **Coeficiente de correlação linear**

O Coeficiente de correlação linear é uma medida que permite quantificar o grau de associação linear entre duas ou mais variáveis. É considerado simples quando mede a relação linear de duas variáveis e múltiplo quando mede a relação linear entre mais de duas variáveis.

Existem vários tipos de coeficientes de correlação, porém, neste trabalho vamos falar do coeficiente de correlação linear de Pearson, simbolicamente representado por R. Os seus valores variam entre -1 e 1. Quanto mais próximo estiverem destes extremos, tanto maior será a correlação existente entre as variáveis (Hill e Hill, 2016). A relação linear é considerada positiva quando o coeficiente de correlação toma valores positivos, negativa quanto toma valores negativos e nula quando é igual a zero. Para classificar a correlação entre as variáveis iremos considerar o critério apresentado na Tabela 2.1.

Tabela 2.1: Interpretação do coeficiente de correlação

Coeficiente de correlação	**Interpretação da correlação**
$0,9 \leq R \leq 1$	Muito forte positiva
$0,7 \leq R < 0,9$	Forte positiva
$0,4 \leq R < 0,7$	Moderada positiva
$0,2 \leq R < 0,4$	Fraca positiva
$0 < R < 0,2$	Muito fraca positiva
$R = 0$	Nula
$-0,2 < R < 0$	Muito fraca negativa
$-0,4 \leq R < -0,2$	Fraca negativa
$-0,7 \leq R < -0,4$	Moderada negativa
$-0,9 \leq R < -0,7$	Forte negativa
$-1 \leq R \leq -0,9$	Muito forte negativa

Fonte: Adaptada do livro de Pestana e Gageiro, (2005)

O sinal positivo ou negativo do coeficiente de correlação de Pearson mostra a direção da relação enquanto que o seu valor em módulo indica a intensidade dessa relação.

Assim, dadas duas variáveis quantitativas X e Y, o coeficiente de correlação linear de Pearson simples entre X e Y é calculado do seguinte modo (Reis, 2017):

$$R = \frac{cov(X,Y)}{S_X S_Y} = \frac{\Sigma_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{2.23}$$

onde o numerador corresponde à covariância entre as duas variáveis e o denominador ao produto dos desvios padrão. Este coeficiente não exprime relações de causalidade, visto que, a correlação entre Y e X é a mesma entre Y e X (Pestana e Gageiro, 2005).

Para efeitos de generalização dos resultados, quando se usa o coeficiente de correlação linear de Pearson, ambas as variáveis devem ter distribuição normal. Se as variáveis não apresentarem distribuição normal, como alternativa ao R, pode-se usar o coeficiente ρ (rô) de Spearman. Este coeficiente calcula-se de forma semelhante ao R de Pearson, substituindo os valores das variáveis pelas respetivas ordens ou postos.

• **Coeficiente de determinação**

Elevando ao quadrado o valor do coeficiente de correlação, obtém-se o coeficiente de determinação, simbolicamente representado por R^2, que é uma medida que mede o poder explicativo da equação de regressão.

O coeficiente de determinação pode ser também calculado através da seguinte expressão:

$$R^2 = \frac{\Sigma_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}, \tag{2.24}$$

onde $\Sigma_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$ e $\Sigma_{i=1}^{n}(y_i - \overline{y})^2$ representam a variação explicada e a variação total de Y, respetivamente. A variação total baseia-se no total de desvio $(y_i - \overline{y})$ que por sua vez resulta da adição dos desvios explicados $(\widehat{y}_i - \overline{y})$ e não explicados $(y_i - \widehat{y}_i)$ pela reta de regressão. A Figura 2.2 mostra a decomposição do desvio total.

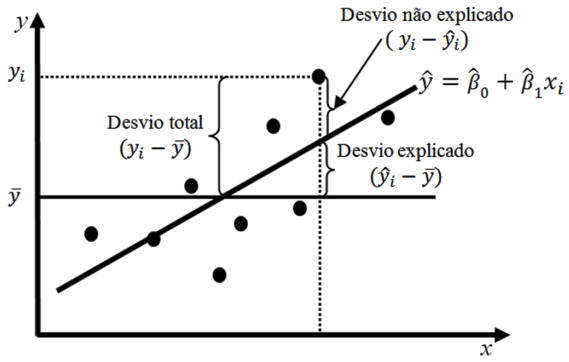

Figura 2.2: Decomposição do desvio total
Fonte: Manual de Estatística Descritiva, (Reis, 2017)

Os valores de R^2 variam entre 0 e 1 e medem a proporção da variabilidade total da variável dependente, que é explicada pelo modelo de regressão. Quando $R^2 = 0$ considera-se que o modelo não se ajusta aos dados, e quando $R^2 = 1$ o ajustamento é considerado perfeito. Na prática, o valor de R^2 que se considera produzir um ajustamento adequado é tido como algo subjetivo (Marôco, 2014).

O R^2 é geralmente interpretado em termos percentuais, cujo valor indica a proporção da variação de Y explicada pela presença da variável X.

2.6 Distribuição amostral dos estimadores

Ao efetuar inferências sobre $\widehat{\beta}_0$ e $\widehat{\beta}_1$, deve-se ter sempre em conta as suas distibuições amostrais, que são apresentadas a seguir.

Como $\widehat{\beta}_0$ e $\widehat{\beta}_1$ são combinações lineares de $y_i, i = 1, \cdots, n$, uma vez que y_i, possui uma distribuição normal, com média $\beta_0 + \beta_1 x_i$ e variância constante σ^2,

$$y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right), i = 1, \cdots, n,$$

então pode-se concluir que $\widehat{\beta}_0$ e $\widehat{\beta}_1$ também seguem uma distribuição normal.

Entretanto, considerando os valores esperados e as variâncias de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ obtidos anteriormente, podemos expressar as distribuições amostrais dos estimadores, $\widehat{\beta}_0$ e $\widehat{\beta}_1$, da seguinte forma:

• **Distribuição amostral de $\widehat{\beta}_0$**

A distribuição amostral de $\widehat{\beta}_0$ é dada por

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right)\right) \tag{2.25}$$

• **Distribuição amostral de $\widehat{\beta}_1$**

A distribuição amostral de $\widehat{\beta}_1$ é dada por

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right). \tag{2.26}$$

2.7 Intervalos de confiança e testes relativos aos parâmetros do modelo

Quando se efetua uma análise estatística utilizando um modelo de regressão linear, depois de ajustar a função de regressão aos dados, a seguir o interesse dos investigadores passa por testar estatisticamente o efeito da variável independente sobre a variável dependente e construir intervalos de confiança para os verdadeiros parâmetros (à partida desconhecidos) com base nas distribuições de probabilidade dos seus estimadores (Caiado, 2016, p. 140).

A construção dos intervalos de confiança e realização de testes de hipóteses para os parâmetros, fazem parte de um conjunto de métodos inferenciais, que são utilizados após a estimação dos parâmetros, para se calcular os intervalos onde os mesmos podem ser considerados válidos e aferir a sua significância no modelo.

2.7.1 Intervalos de confiança

As estimativas pontuais obtidas a partir dos estimadores do modelo de regressão linear, não contêm informações sobre a precisão dos valores obtidos, o que justifica a necessidade de se

proceder à estimação dos seus valores através da construção de intervalos de confiança.

• **Intervalo de confiança para** β_1

Para se determinar o intervalo de confiança para β_1, é necessário recordar a distribuição amostral deste parâmetro, dada pela expressão (2.26) e o estimador pontual de σ^2, $S^2 = \frac{SQE}{n-2}$, dado pela expressão (2.16).

Assim, o estimador não enviesado da variância de β_1 será dado por

$$S^2_{\widehat{\beta}_1} = \frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

e

$$S_{\widehat{\beta}_1} = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \tag{2.27}$$

Logo a estatística de teste para o parâmetro β_1 é dada por

$$T_1 = \frac{\widehat{\beta}_1 - \beta_1}{S_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \sim t_{(n-2)}, \tag{2.28}$$

que segue uma distribuição $t\ de\ Student$ com $(n-2)$ graus de liberdade, $t_{(n-2)}$. $(n-2)$ é também o número de graus de liberdade associados ao estimador não enviesado de σ^2, representado por S^2.

Seja $t_{1-\frac{\alpha}{2},n-2}$, o quantil de probabilidade de ordem $1-\frac{\alpha}{2}$, de uma distribuição t com $(n-2)$ graus de liberdade. O intervalo de confiança a $(1-\alpha)\times 100\%$, para o parâmetro β_1 será dado por:

$$\left[\widehat{\beta}_1 - t_{1-\frac{\alpha}{2},n-2}\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}; \widehat{\beta}_1 + t_{1-\frac{\alpha}{2},n-2}\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right]. \tag{2.29}$$

• **Intervalo de confiança para** β_0

O procedimento para a obtenção do intervalo de confiança para o parâmetro β_0 é semelhante ao da obtenção do intervalo de confiança para β_1. Assim, uma vez que $\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)\right)$, tem-se o estimador não enviesado da variância de β_0 dado por:

$$S^2_{\widehat{\beta}_0} = \frac{S^2}{n} + \frac{S^2\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

e

$$S_{\widehat{\beta}_0} = S\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right)}, \tag{2.30}$$

deste modo, a estatística do teste para o parâmetro β_0 continua a seguir uma distribuição $t\ de\ Student$ com $(n-2)$ graus de liberdade e é dada por:

$$T_0 = \frac{\widehat{\beta}_0 - \beta_0}{S_{\widehat{\beta}_0}} = \frac{\widehat{\beta}_0 - \beta_0}{S\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}} \sim t_{(n-2)}, \tag{2.31}$$

Assim, pode-se concluir que o intervalo de confiança a $(1-\alpha)\times 100\%$, para o parâmetro β_0 será dado por:

$$\left[\widehat{\beta}_0 - t_{1-\frac{\alpha}{2},n-2}S\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right)}; \widehat{\beta}_0 + t_{1-\frac{\alpha}{2},n-2}S\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right)}\right]. \tag{2.32}$$

2.7.2 Testes relativos aos parâmetros do modelo

• **Testes de hipóteses sobre** β_1

Normalmente está-se interessado em saber se a relação estatística entre x_i e y_i de facto existe, isto é, se existe alguma razão para se considerar a regressão. Para isso, é necessário realizar-se um teste de hipóteses adequado, por forma a permitir a tomada de uma determinada decisão.

Para se saber a significância do parâmetro β_1 no modelo testam-se as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}. \tag{2.33}$$

Quando $\beta_1 = 0$, o conhecimento de X não dá nenhuma informação sobre Y, concluindo-se assim, que não existe uma relação significativa entre X e Y. Assim, um MRLS só pode ser considerado como estatisticamente significativo quando se rejeita a hipótese nula (H_0) referente ao parâmetro β_1.

A estatística de teste que surge como consequência do processo desenvolvido para obtenção do intervalo de confiança para β_1 é dada por:

$$T_1 = \frac{\widehat{\beta}_1}{S_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \sim t_{(n-2)},$$

visto que, para H_0, se tem $\beta_1 = 0$.

No caso mais geral, em que se pretende testar

$$\begin{cases} H_0 : \beta_1 = \beta_1' \\ H_1 : \beta_1 \neq \beta_1' \quad (\beta_1 > \beta_1' \; ou \; \beta_1 < \beta_1'), \end{cases} \tag{2.34}$$

usa-se a estatística de teste obtida com a expressão (2.28),

$$T_1 = \frac{\widehat{\beta}_1 - \beta_1'}{S_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1 - \beta_1'}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \sim t_{(n-2)}.$$

Nos dois casos, rejeita-se H_0, para um nível de significância α, quando $|t_{obs}| > t_{1-\frac{\alpha}{2},n-2}$, onde t_{obs} representa o valor observado da estatística de teste.

• **Testes de hipóteses sobre** β_0

Quanto ao teste de hipóteses sobre β_0, têm-se como hipóteses:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases} . \tag{2.35}$$

Na eventualidade de se admitir a veracidade de H_0, o modelo passa a ter características de um modelo de regressão sem termo independente, cuja reta de regressão passa pela origem das coordenadas.

No caso mais geral em que se testem as hipóteses:

$$\begin{cases} H_0 : \beta_0 = \beta_0' \\ H_1 : \beta_0 \neq \beta_0' \quad (\beta_0 > \beta_0' \; ou \; \beta_0 < \beta_0') \end{cases} \tag{2.36}$$

usa-se a estatística

$$T_0 = \frac{\widehat{\beta}_0 - \beta_0'}{S_{\widehat{\beta}_0}} = \frac{\widehat{\beta}_0 - \beta_0'}{S\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}} \sim t_{(n-2)}. \tag{2.37}$$

Também nestes dois casos, se rejeita H_0, com um nível de significância α, quando $|t_{obs}| > t_{1-\frac{\alpha}{2},n-2}$.

2.8 Conclusão

Termina aqui a primeira parte da abordagem que se reservou para os aspetos teóricos essenciais do MRLS. No Capítulo 4 iremos falar da validação dos pressupostos impostos ao erro deste tipo de modelo. Acredita-se que o conteúdo apresentado neste tema poderá de algum modo ajudar os estudantes do ensino secundário e superior a aprofundar os seus conhecimentos sobre esta temática.

Como se pode notar, no MRLS só se utiliza uma única variável independente, ou seja, um fator condicionante da variável dependente ou resposta. Porém, na prática, raramente a resposta de um determinado fenómeno que se esteja interessado em estudar, depende de apenas um fator. Nesse caso é importante medir o efeito dos demais fatores, cuja informação esteja disponível. Assim sendo, na análise estatística com recurso à regressão linear já não se deve usar o MRLS, mas sim, o Modelo de Regressão Linear Múltipla onde a variável dependente é influenciada por mais de uma variável independente. É sobre este modelo que nos iremos debruçar no próximo Capítulo.

Capítulo 3

Modelo de Regressão Linear Múltipla - MRLM

Ao estudar um determinado fenómeno e caso estejam disponíveis várias variáveis independentes, em vez de se estudar o efeito de cada uma das variáveis sobre a variável dependente de forma individual, é aconselhado que se incluam no modelo de regressão linear as variáveis independentes disponíveis. Deste modo, utiliza-se o Modelo de Regressão Linear Múltipla (MRLM), que é uma extensão do MRLS, em que se considera mais do que uma variável independente.

Neste Capítulo vão ser abordados os principais aspetos deste modelo, nomeadamente, a equação e os pressupostos, a estimação dos parâmetros, a definição dos intervalos de confiança e a realização de testes relativos aos parâmetros, a ANOVA de regressão, os métodos de seleção das variáveis independentes e o uso de variáveis qualitativas.

3.1 Equação do modelo de regressão linear múltipla

Modelo de Regressão Linear Múltipla é o modelo que descreve uma relação linear entre um conjunto de variáveis independentes (variáveis preditoras), $X_j,\ j = 1, 2, \cdots, k,$ e uma variável dependente (variável explicada ou resposta) Y. É representado pela seguinte equação:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,\ i = 1, 2, \cdots, n, \tag{3.1}$$

onde
y_i - representa a i-ésima observação da variável dependente;
$x_{i1}, \cdots, x_{ik}$ - representam as i-ésimas observações das variáveis independentes;
$\beta_0, \beta_1, \cdots, \beta_k$ - representam os coeficientes ou parâmetros do modelo (à partida desconhecidos);
ε_i - variável aleatória residual (erro) que descreve os efeitos em y_i, não explicados pelas variáveis $x_{ij},\ j = 1, 2, \cdots, k$;
n - é o número de observações.

3.1.1 Representação matricial

As n igualdades da equação (3.1) podem ser representadas em notação matricial da seguinte forma:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}}_{\mathbf{X}} \times \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

da qual resulta o MRLM escrito da forma,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.2}$$

onde:
$\mathbf{Y}$ - é o vetor das n observações aleatórias da variável dependente;
$\mathbf{X}$ - é a matriz das observações das variáveis independentes;
$\boldsymbol{\beta}$ - é o vetor dos coeficientes de regressão;
$\boldsymbol{\varepsilon}$ - é o vetor correspondente à variável residual (erro).

3.1.2 Representação gráfica do MRLM

Graficamente o Modelo de RLM é representado por um hiperplano k-dimensional com base nas *k* variáveis independentes do Modelo. A Figura 3.1 apresenta o exemplo de um caso tridimensional.

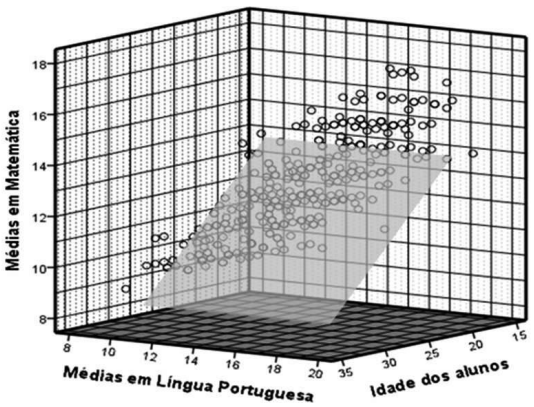

Figura 3.1: Hiperplano tridimensional

3.1.3 Equação do MRLM com variáveis de interação

Na análise do modelo de regressão linear múltipla, em determinadas situações, os modelos podem tornar-se cada vez mais complexos. Em vez de terem a configuração habitual, na equação

do modelo podem aparecer, por exemplo, termos das variáveis independentes que são o quadrado de uma ou mais variáveis independentes e outros termos das variáveis independentes que são o produto de duas variáveis independentes distintas. Um dos exemplos é o caso da seguinte equação:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_4 x_{i4} + \beta_1 x_{i1}^2 + \ldots + \beta_4 x_{i4}^2 + \beta_5 x_{i1} x_{i2} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{3i} x_{i4} + \varepsilon_i. \quad (3.3)$$

Neste tipo de modelo, o termo $x_{ij}x_{il}$, $j, l = 1, \ldots, 4$, por exemplo, representa a interação entre as variáveis X_j e X_l. Se a interação, entre as duas variáveis for significativa, o efeito de x_{ij} no valor esperado da variável dependente fica condicionado pelo valor de x_{il} e analogamente o efeito de x_{il} no valor esperado da mesma variável depende do valor de x_{ij}.

É importante referir que, devido a sua complexidade, este tipo de modelos não vão ser abordados neste trabalho.

3.2 Pressupostos do modelo

Os pressupostos para a utilização do MRLM são os seguintes:

1. Linearidade do fenómeno em análise;
2. O valor esperado dos resíduos do modelo deve ser nulo, $E(\varepsilon_i) = 0$;
3. Os resíduos das observações devem ser mutuamente independentes;
4. Os resíduos devem ser homoscedásticos, o que significa que a sua variância deve ser constante, $var(\varepsilon_i) = \sigma^2 > 0$, $i = 1, \cdots, n$;
5. Os resíduos devem ter uma distribuição normal com média zero e variância constante, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \cdots, n$;
6. Não devem existir situações de colinearidade e/ou multicolinearidade entre as variáveis independentes do modelo.

O MRLM só é considerado válido para explicar ou prever certo fenómeno em estudo, caso verifique estes pressupostos.

Tal como se afirmou no capítulo anterior, para qualquer modelo de regressão linear não se obtêm diretamente os valores de y_i, dados os valores de x_{ij} $j = 1, \cdots, k$, já que a equação do modelo contém a variável ε_i, que representa quantidades aleatórias difíceis de quantificar. O que se pode obter é o valor esperado de y_i.

No caso do MRLM o valor esperado de y_i é dado por,

$$E\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + E(\varepsilon_i)$$

$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},$$

visto que, $E(\varepsilon_i) = 0$.

Neste caso, a função de regressão (ou função resposta) que relaciona o valor esperado de y_i com o valor das variáveis independentes, $X_1, \cdots, X_k$, é

$$\widehat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},\ i = 1, 2, \cdots, n. \tag{3.4}$$

Com a obtenção da equação (3.4), é possível estimar o valor da variável residual, ε_i, através da expressão:

$$\varepsilon_i = y_i - \widehat{y}_i,\ i = 1, 2, \cdots, n.$$

Tal como acontece na RLS, na RLM também as estimativas dos resíduos são usadas para validar os pressupostos do modelo e estimar os parâmetros do modelo. Assim, os parâmetros $\beta_0, \beta_1, \cdots, \beta_k$, do modelo são estimados por meio dos pontos experimentais obtidos através da amostra, resultando na equação estimada do modelo, que é representada da seguinte forma:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}, \tag{3.5}$$

onde $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k$ correspondem às estimativas dos parâmetros $\beta_0, \beta_1, \cdots, \beta_k$.

3.3 Estimação dos parâmetros do modelo

Para se estimar os parâmetros do MRLM é necessário ter uma amostra aleatória, da qual se vão obter os estimadores não enviesados para a equação ajustada,

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}.$$

Assim, os parâmetros $(\beta_0, \beta_1, \cdots, \beta_k)$ do MRLM, são estimados a partir de um conjunto de observações de uma amostra, recorrendo ao Método dos Mínimos Quadrados (MMQ).

3.3.1 Método dos mínimos quadrados

Tal como acontece no MRLS, o método dos mínimos quadrados no MRLM permite obter os estimadores não enviesados dos parâmetros do modelo.

O estimador dos mínimos quadrados é aquele que minimiza a soma dos quadrados dos resíduos.

Neste caso se tem

$$SQE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - \widehat{y}_i]^2$$

$$= \sum_{i=1}^{n} \left[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}) \right]^2 . \tag{3.6}$$

Os valores de $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k$ que minimizam este resultado, são aqueles que anulam as suas derivadas parciais em ordem aos respetivos estimadores

$$\begin{cases} \frac{\partial SQE}{\partial \widehat{\beta}_0} = 0 \\ \\ \frac{\partial SQE}{\partial \widehat{\beta}_1} = 0 \\ \quad \vdots \\ \frac{\partial SQE}{\partial \widehat{\beta}_k} = 0 \end{cases} .$$

Assim tem-se,

$$\begin{aligned} \frac{\partial SQE}{\partial \widehat{\beta}_0} &= \sum_{i=1}^{n} 2 \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}) \times (-1) \\ &= -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}), \end{aligned}$$

$$\begin{aligned} \frac{\partial SQE}{\partial \widehat{\beta}_1} &= \sum_{i=1}^{n} 2 \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}) \times (-x_{i1}) \\ &= -2 \sum_{i=1}^{n} x_{i1} \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}), \end{aligned}$$

$(\cdots)$

$$\begin{aligned} \frac{\partial SQE}{\partial \widehat{\beta}_k} &= \sum_{i=1}^{n} 2 \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}) \times (-x_{ik}) \\ &= -2 \sum_{i=1}^{n} x_{ik} \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}). \end{aligned}$$

Destes resultados deduz-se o seguinte sistema de equações:

$$\begin{cases} -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}) = 0 \\ \\ -2 \sum_{i=1}^{n} x_{i1} \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}) = 0 \\ \quad \vdots \\ -2 \sum_{i=1}^{n} x_{ik} \times (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}) = 0 \end{cases} ,$$

de onde se obtêm as equações normais dos Mínimos Quadrados para o MRLM que são apresentadas a seguir:

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i \end{cases} . \tag{3.7}$$

Isolando $\widehat{\beta}_0$ da primeira equação do sistema obtém-se:

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} - \cdots - \widehat{\beta}_k \frac{1}{n}\sum_{i=1}^{n} x_{ik}, \tag{3.8}$$

com base nas expressões (2.9) e (2.10) do capítulo anterior a expressão (3.8) pode ser reescrita na forma:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}_1 - \cdots - \widehat{\beta}_k\overline{x}_k.$$

Depois de se obter o resultado de $\widehat{\beta}_0$ substituindo-o nas restantes equações obtém-se um sistema de equações como o seguinte:

$$\begin{cases} \widehat{\beta}_1 S_{x_1x_1} + \widehat{\beta}_2 S_{x_1x_2} + \cdots + \widehat{\beta}_k S_{x_1x_k} = S_{x_1y} \\ \widehat{\beta}_1 S_{x_2x_1} + \widehat{\beta}_2 S_{x_2x_2} + \cdots + \widehat{\beta}_k S_{x_2x_k} = S_{x_2y} \\ \vdots \\ \widehat{\beta}_1 S_{x_kx_1} + \widehat{\beta}_2 S_{x_kx_2} + \cdots + \widehat{\beta}_k S_{x_kx_k} = S_{x_ky} \end{cases} , \tag{3.9}$$

onde

$$S_{x_jx_j} = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2, j = 1, \cdots, k, \tag{3.10}$$

$$S_{x_jx_l} = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)\left(x_{il} - \overline{x}_l\right), j = 1, \cdots, k, l = 1, \cdots, k, \tag{3.11}$$

e

$$S_{x_iy} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)\left(x_{ij} - \overline{x}_j\right), j = 1, \cdots, k. \tag{3.12}$$

Substituindo os resultados das somas de quadrados e de produtos cruzados em cada uma das equações que compõem o sistema de equações e resolvendo-o, obtêm-se as equações que nos permitem calcular os estimadores $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\cdots$, $\widehat{\beta}_k$, dos parâmetros.

Os estimadores $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\cdots$, $\widehat{\beta}_k$ dão as melhores estimativas não enviesadas de β_1, β_2, $\cdots$, β_k e como consequência disso, a equação,

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik},\ i = 1, \cdots, n,$$

também dá a melhor estimativa linear não enviesada de $\widehat{y}_i$.

3.3.2 Estimação dos parâmetros a partir da representação matricial do MRLM

Os parâmetros do MRLM também podem ser estimados a partir da sua representação matricial apresentada em 3.1.1. Caso o vetor desconhecido β da expressão (3.2) seja substituído pela sua estimativa $\widehat{\beta}$, defini-se um vetor de resíduos ε , tendo-se, ver Johnston e DiNardo (1997),

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}, \tag{3.13}$$

onde

$$\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}. \tag{3.14}$$

A aplicação do MMQ neste caso consiste em escolher a estimativa $\widehat{\beta}$, que minimiza a soma dos quadrados dos resíduos, $\varepsilon'\varepsilon$. Tem-se

$$\begin{aligned} SQE &= \varepsilon'\varepsilon \\ &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{Y'Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X'Y} - \boldsymbol{Y'X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X'X}\widehat{\boldsymbol{\beta}}, \\ &= \boldsymbol{Y'Y} - 2\widehat{\boldsymbol{\beta}}'\boldsymbol{X'Y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X'X}\widehat{\boldsymbol{\beta}}, \end{aligned}$$

dado que $\boldsymbol{Y'X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y'X}\widehat{\boldsymbol{\beta}})' = \widehat{\boldsymbol{\beta}}'\boldsymbol{X'Y}$. Assim

$$\frac{\partial SQE}{\partial \widehat{\boldsymbol{\beta}}} = 0$$

$$\frac{\partial(\boldsymbol{Y'Y} - 2\widehat{\boldsymbol{\beta}}'\boldsymbol{X'Y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X'X}\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}} = 0$$

$$-2\boldsymbol{X'Y} + 2\boldsymbol{X'X}\widehat{\boldsymbol{\beta}} = 0^{2},$$

e a resolução desta equação dá o seguinte resultado:

$$(\boldsymbol{X'X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X'Y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}, \tag{3.15}$$

onde $(\boldsymbol{X'X})^{-1}$ é a matriz inversa da matriz $\boldsymbol{X'X}$.

[2] Visto que, na diferenciação matricial caso se tenha $\boldsymbol{a}$ e $\boldsymbol{b}$ dois vetores e $\boldsymbol{A}$ uma matriz simétrica tais que $\boldsymbol{b'a}$ e $\boldsymbol{b'Ab}$ existam, verificam-se as seguintes regras, ver (Oliveira, Santos e Fortuna, 2011, p. 29):

$$\frac{\partial(\boldsymbol{b'a})}{\partial \boldsymbol{b}} = \boldsymbol{a} \quad \text{e} \quad \frac{\partial(\boldsymbol{b'Ab})}{\partial \boldsymbol{b}} = 2\boldsymbol{Ab}\ .$$

Através do resultado do cálculo matricial tem-se:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} = \begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i \end{cases} .$$

Este sistema de equações é idêntico ao sistema (3.7) obtido anteriormente, a partir do qual se consegue obter as equações que possibilitam calcular os estimadores não enviesados dos parâmetros do modelo.

3.3.3 Exemplo da estimação dos parâmetros de um modelo de RLM com três parâmetros

Caso se tenha um problema que envolva a estimação de três parâmetros ($\beta_0, \beta_1 \ e \ \beta_2$), para se obter as equações que permitem calcular os valores dos estimadores ($\widehat{\beta}_0, \widehat{\beta}_1 \ e \ \widehat{\beta}_2$) dos referidos parâmetros, pode-se proceder do seguinte modo:

Em primeiro lugar defini-se o MRLM, que neste caso corresponde a:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \ i = 1, 2, \cdots, n, \tag{3.16}$$

a partir do qual, recorrendo ao método dos mínimos quadrados, se obtêm as seguintes equações normais:

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2}y_i \end{cases} .$$

Na primeira equação isolando $\widehat{\beta}_0$ obtém-se a seguinte equação

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} - \widehat{\beta}_2 \frac{1}{n}\sum_{i=1}^{n} x_{i2},$$

que pode ser reescrita da forma

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}_1 - \widehat{\beta}_2 \overline{x}_2, \tag{3.17}$$

visto que

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

e

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ j = 1, 2. \tag{3.18}$$

Assim, para se obter as equações que permitem calcular os estimadores $\widehat{\beta}_1$ e $\widehat{\beta}_2$, vai-se estabelecer o seguinte sistema de equações:

$$\begin{cases} \widehat{\beta}_1 S_{x_1^2} + \widehat{\beta}_2 S_{x_1 x_2} = S_{x_1 y} \\ \\ \widehat{\beta}_1 S_{x_2 x_1} + \widehat{\beta}_2 S_{x_2^2} = S_{x_2 y} \end{cases}, \tag{3.19}$$

onde

$$S_{x_j^2} = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2 = \sum_{i=1}^{n} x_{ij}^2 - n\overline{x}_j^2,\ j = 1, 2. \tag{3.20}$$

$$S_{x_j x_l} = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)\left(x_{il} - \overline{x}_l\right) = \sum_{i=1}^{n} x_{ij} x_{il} - n\overline{x}_j\overline{x}_l,\ j, l = 1, 2,\ j \neq l \tag{3.21}$$

e

$$S_{x_j y} = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)\left(y_i - \overline{y}\right) = \sum_{i=1}^{n} x_{ij} y_i - n\overline{x}_j\overline{y},\ j = 1, 2. \tag{3.22}$$

Com base nos resultados destas expressões o sistema de equações (3.19) pode ser reescrito da seguinte forma:

$$\begin{cases} \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2\right) + \widehat{\beta}_2 \left(\sum_{i=1}^{n} x_{i1} x_{i2} - n\overline{x}_1\overline{x}_2\right) = \sum_{i=1}^{n} x_{i1} y_i - n\overline{x}_1\overline{y} \\ \\ \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_{i2} x_{i1} - n\overline{x}_2\overline{x}_1\right) + \widehat{\beta}_2 \left(\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2\right) = \sum_{i=1}^{n} x_{i2} y_i - n\overline{x}_2\overline{y} \end{cases}. \tag{3.23}$$

Isolando $\widehat{\beta}_1$ na primeira equação obtém-se,

$$\widehat{\beta}_1 \left(\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2\right) = \sum_{i=1}^{n} x_{i1} y_i - n\overline{x}_1\overline{y} - \widehat{\beta}_2 \left(\sum_{i=1}^{n} x_{i1} x_{i2} - n\overline{x}_1\overline{x}_2\right)$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_{i1} y_i - n\overline{x}_1\overline{y} - \widehat{\beta}_2 \left(\sum_{i=1}^{n} x_{i1} x_{i2} - n\overline{x}_1\overline{x}_2\right)}{\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2}. \tag{3.24}$$

Substituindo o resultado de $\widehat{\beta}_1$ na segunda equação do sistema, com o propósito de isolar $\widehat{\beta}_2$ obtém-se:

$$\widehat{\beta}_1 \left(\sum_{i=1}^{n} x_{i2} x_{i1} - n\overline{x}_2\overline{x}_1\right) + \widehat{\beta}_2 \left(\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2\right) = \sum_{i=1}^{n} x_{i2} y_i - n\overline{x}_2\overline{y}$$

$$\left(\frac{\sum_{i=1}^{n} x_{i1}y_i - n\overline{x}_1\overline{y} - \widehat{\beta}_2\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\overline{x}_1\overline{x}_2\right)}{\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2}\right)\times\left(\sum_{i=1}^{n} x_{i2}x_{i1} - n\overline{x}_2\overline{x}_1\right)+\widehat{\beta}_2\left(\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2\right) = \sum_{i=1}^{n} x_{i2}y_i - n\overline{x}_2\overline{y}$$

$$\widehat{\beta}_2\left(\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2\right) = \sum_{i=1}^{n} x_{i2}y_i - n\overline{x}_2\overline{y} - \left(\frac{\sum_{i=1}^{n} x_{i1}y_i - n\overline{x}_1\overline{y} - \widehat{\beta}_2\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\overline{x}_1\overline{x}_2\right)}{\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2}\right)\times\left(\sum_{i=1}^{n} x_{i2}x_{i1} - n\overline{x}_2\overline{x}_1\right)$$

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{n} x_{i2}y_i - n\overline{x}_2\overline{y} - \left(\frac{\sum_{i=1}^{n} x_{i1}y_i - n\overline{x}_1\overline{y} - \widehat{\beta}_2\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\overline{x}_1\overline{x}_2\right)}{\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2}\right)\times\left(\sum_{i=1}^{n} x_{i2}x_{i1} - n\overline{x}_2\overline{x}_1\right)}{\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2}$$

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{n} x_{i2}y_i - n\overline{x}_2\overline{y} - \left(\frac{\left(\sum_{i=1}^{n} x_{i1}y_i - n\overline{x}_1\overline{y}\right)\left(\sum_{i=1}^{n} x_{i2}x_{i1} - n\overline{x}_2\overline{x}_1\right) - \widehat{\beta}_2\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\overline{x}_1\overline{x}_2\right)^2}{\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2}\right)}{\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2}$$

$$\vdots$$

$$\widehat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} x_{i2}y_i - n\overline{x}_2\overline{y}\right)\left(\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2\right) - \left(\sum_{i=1}^{n} x_{i1}y_i - n\overline{x}_1\overline{y}\right)\left(\sum_{i=1}^{n} x_{i2}x_{i1} - n\overline{x}_2\overline{x}_1\right)}{\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\overline{x}_1\overline{x}_2\right)^2 + \left(\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2\right)\left(\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2\right)}. \tag{3.25}$$

Deste modo, caso se esteja a analisar um problema que envolva três parâmetros com recurso à regressão linear múltipla, pode-se utilizar as equações que vêm abaixo para se obter os estimadores $\widehat{\beta}_0$, $\widehat{\beta}_1$ e $\widehat{\beta}_2$:

1. $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}_1 - \widehat{\beta}_2\overline{x}_2$,

2. $\widehat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_{i1}y_i - n\overline{x}_1\overline{y} - \widehat{\beta}_2\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\overline{x}_1\overline{x}_2\right)}{\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2}$,

3. $\widehat{\beta}_2 = \dfrac{\left(\sum_{i=1}^{n} x_{i2}y_i - n\overline{x}_2\overline{y}\right)\left(\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2\right) - \left(\sum_{i=1}^{n} x_{i1}y_i - n\overline{x}_1\overline{y}\right)\left(\sum_{i=1}^{n} x_{i2}x_{i1} - n\overline{x}_2\overline{x}_1\right)}{\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\overline{x}_1\overline{x}_2\right)^2 + \left(\sum_{i=1}^{n} x_{i2}^2 - n\overline{x}_2^2\right)\left(\sum_{i=1}^{n} x_{i1}^2 - n\overline{x}_1^2\right)}$.

Visto que, estas equações dão as melhores estimativas não enviesadas de β_0, β_1 e β_2, então a equação,

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2},\ i = 1, \cdots, n, \tag{3.26}$$

também dá a melhor estimativa linear não enviesada de y_i.

3.3.4 Propriedades dos estimadores

Tal como no MRLS, no MRLM os estimadores $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\cdots$, $\widehat{\beta}_k$ obtidos através do método dos mínimos quadrados devem satisfazer as seguintes propriedades:

1. Devem ser funções lineares de y_i;

2. Devem ser estimadores centrados ou não enviesados de β como se demonstra a seguir. Dado que,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y},$$

onde

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

e sabendo que,

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I_{k+1}},$$

onde $\boldsymbol{I_{k+1}}$ corresponde à matriz identidade de ordem k+1. Deste modo o valor esperado de $\widehat{\beta}$ é dado por,

$$\begin{aligned}
E(\widehat{\boldsymbol{\beta}}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon})) \\
&= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'\varepsilon}) \\
&= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta}) + E((\boldsymbol{X'X})^{-1}\boldsymbol{X'\varepsilon}) \\
&= E(\boldsymbol{I_{k+1}} \times \boldsymbol{\beta}) + E((\boldsymbol{X'X})^{-1}\boldsymbol{X'\varepsilon}) \\
&= E(\boldsymbol{I_{k+1}} \times \boldsymbol{\beta}) + (\boldsymbol{X'X})^{-1}\boldsymbol{X'} \times E(\boldsymbol{\varepsilon}) \\
&= E(\boldsymbol{I_{k+1}} \times \boldsymbol{\beta}) + (\boldsymbol{X'X})^{-1}\boldsymbol{X'} \times \mathbf{0_n},
\end{aligned}$$

onde $\mathbf{0_n}$ representa o vetor nulo com n componentes. Logo tem-se

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{3.27}$$

Este resultado significa que, se fosse possível obter muitas observações particulares do vetor $\boldsymbol{Y}$, para a mesma matriz $\boldsymbol{X}$, se obteriam outras tantas estimativas de $\widehat{\boldsymbol{\beta}}$, que em média, tenderiam para o verdadeiro valor do vetor dos coeficientes $\boldsymbol{\beta}$. Assim, pode-se afirmar que o não enviesamento de $\widehat{\boldsymbol{\beta}}$ garante que este estimador é centrado (Murtela, Ribeiro, Silva, Pimenta e Pimenta, 2015).

3. Os estimadores devem possuir variância mínima, ou seja, devem ser eficientes.

A variâcia de $\widehat{\beta}$ é dada por:

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

tal como se demonstra a seguir.

A matriz de variâncias e covariâncias de dimensão $(k+1) \times (k+1)$, do vetor $\widehat{\boldsymbol{\beta}}$ é dada por:

$$Var(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix}
Var(\widehat{\beta}_0) & Cov(\widehat{\beta}_0\widehat{\beta}_1) & \cdots & Cov(\widehat{\beta}_0\widehat{\beta}_k) \\
Cov(\widehat{\beta}_1\widehat{\beta}_0) & Var(\widehat{\beta}_1) & \cdots & Cov(\widehat{\beta}_1\widehat{\beta}_k) \\
\cdots & \cdots & \cdots & \cdots \\
Cov(\widehat{\beta}_k\widehat{\beta}_0) & Cov(\widehat{\beta}_k\widehat{\beta}_1) & \cdots & Var(\widehat{\beta}_k)
\end{bmatrix}$$

Com base nos resultados desta matriz, e uma vez que

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

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

tem-se

$$Var(\widehat{\boldsymbol{\beta}}) = E\left[(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}))(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}))'\right]$$

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

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

$$= \sigma^2\boldsymbol{I_{k+1}}(\boldsymbol{X'X})^{-1}$$

dado que $E(\boldsymbol{\varepsilon\varepsilon'}) = \sigma^2$ e $(\boldsymbol{X'X})^{-1}\boldsymbol{X'X} = \boldsymbol{I_{k+1}}$. Logo,

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X'X})^{-1}. \tag{3.28}$$

4. Os estimadores devem ser consistentes, ou seja, à medida que o tamanho da amostra aumenta, os seus valores esperados devem convergir para os verdadeiros valores dos parâmetros e as suas variâncias devem tender para o valor nulo.

3.3.5 Estimador da variância residual σ^2

Tal como se pode notar na expressão (3.28), a variância dos estimadores dos coeficientes de um MRLM é calculada através da expressão, $\sigma^2(\boldsymbol{X'X})^{-1}$. Uma vez que o valor da variância da variável residual, σ^2, é frequentemente desconhecido, para que se possa obter um valor numérico da mesma, considera-se um estimador desta variável, que é obtido através da seguinte expressão:

$$S^2 = \frac{SQE}{n-(k+1)}, \tag{3.29}$$

tal como se mostra a seguir.

Considerando a representação do modelo ajustado na forma matricial vista anteriormente, tem-se $\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \varepsilon$, donde se obtém $\varepsilon = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Para $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\omega}$, vem

$$\boldsymbol{\varepsilon} = \boldsymbol{X\beta} + \boldsymbol{\omega} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Substituindo $\widehat{\boldsymbol{\beta}}$, conforme a expressão (3.15), obtém-se:

$$\boldsymbol{\varepsilon} = \boldsymbol{X\beta} + \boldsymbol{\omega} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y},$$

e substituindo novamente $\mathbf{Y}$ por $\boldsymbol{X\beta} + \boldsymbol{\omega}$, tem-se:

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

o que resulta em

$$\boldsymbol{\varepsilon = X\beta + \omega - X\underbrace{(X'X)^{-1}X'X}_{=I_{k+1}}\beta - X(X'X)^{-1}X'\omega}$$

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

$$= \boldsymbol{\omega - X(X'X)^{-1}X'\omega},$$

$$= \boldsymbol{\left(I_n - X(X'X)^{-1}X'\right)\omega}.$$

Considerando,

$$\mathbf{M} = \left(\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right), \tag{3.30}$$

tem-se

$$\boldsymbol{\varepsilon = \mathbf{M} \times \omega}. \tag{3.31}$$

Segundo Oliveira, Santos e Fortuna (2011), a matriz $\mathbf{M}$ é frequentemente utilizada como um recurso para dar solução a problemas de estimação em estatística e possui as seguintes características:

1. É uma matriz simétrica como se mostra de seguida:

$$\mathbf{M}' = \left(\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)'$$

$$= (\mathbf{I_n})' - \mathbf{X}'\left((\mathbf{X}'\mathbf{X})^{-1}\right)'(\mathbf{X}')'$$

$$= \left(\mathbf{I_n} - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\right)$$

$$= \left(\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) = \mathbf{M}. \tag{3.32}$$

2. É uma matriz idempotente:

$$\mathbf{M}^2 = \left[\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \times \left[\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$$

$$= \boldsymbol{I_n - X(X'X)^{-1}X'I_n - X(X'X)^{-1}X'I_n + X\underbrace{(X'X)^{-1}X'X}_{I_{k+1}}(X'X)^{-1}X'}$$

$$= \mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$= \left(\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) = \mathbf{M}. \tag{3.33}$$

3. O seu traço (soma dos elementos da diagonal principal de uma matriz quadrada) é igual a $tr(\mathbf{M}) = \Sigma_{i=1}^{n} m_{ii} = n - (k+1)$, como podemos ver de seguida

$$tr(\mathbf{M}) = tr\left(\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)$$

$$= tr\left(\mathbf{I_n}\right) - tr\left(\mathbf{I_{k+1}}\right) = n - (k+1), \tag{3.34}$$

dado que, $tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$ e $tr(\mathbf{ABC}) = tr(\mathbf{BCA}) = tr(\mathbf{CAB})$.

4. $\mathbf{MX} = \mathbf{0}$.

$$\mathbf{MX} = \left(\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{X}$$

$$= \mathbf{X} - \mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}}_{\mathbf{I_{k+1}}}$$

$$= \mathbf{X} - \mathbf{X} = 0. \tag{3.35}$$

Entretanto, dando sequência à estimação de σ^2, e uma vez que a soma de quadrados dos resíduos é dada por:

$$\Sigma_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon'\varepsilon}, \tag{3.36}$$

com base na expressão (3.31) tem-se

$$\boldsymbol{\varepsilon'\varepsilon} = (\boldsymbol{M\omega})'(\boldsymbol{M\omega}) = \boldsymbol{M'\omega'M\omega}.$$

Como se verificou anteriormente, a matriz $\mathbf{M}$ é simétrica e idempotente, ou seja, $\mathbf{M} = \mathbf{M}'$ e $\mathbf{M} \times \mathbf{M} = \mathbf{M}$. Aplicando estas propriedades, tem-se

$$\Sigma_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon'\varepsilon} = \boldsymbol{\omega'M\omega}.$$

Fazendo $\boldsymbol{\omega}' = \boldsymbol{\varepsilon}'$ e $\boldsymbol{\omega} = \boldsymbol{\varepsilon}$, multiplicando as matrizes, obtém-se,

$$\boldsymbol{e'Me} = m_{11}\varepsilon_1^2 + \ldots + m_{nn}\varepsilon_n^2 + 2m_{12}\varepsilon_1\varepsilon_2 + \ldots + 2m_{n-1,n}\varepsilon_{n-1}\varepsilon_n.$$

Calculando o valor esperado da soma de quadrados dos resíduos, e uma vez que $E(\varepsilon_i^2) = \sigma^2$ e $E(\varepsilon_i\varepsilon_j) = 0,\ i \neq j$ e $tr(\mathbf{M}) = \Sigma_{i=1}^{n} m_{ii} = n - (k+1)$, obtém-se,

$$E(\boldsymbol{\varepsilon'\varepsilon}) = \sigma^2(m_{11} + m_{22} + \ldots + m_{nn}) = \sigma^2 tr(\mathbf{M}).$$

Substituindo neste resultado a matriz $\mathbf{M}$ pela sua expressão, obtém-se:

$$E(\boldsymbol{\varepsilon'\varepsilon}) = \sigma^2 tr\left(\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\right)$$

$$= \sigma^2\left[tr\left(\mathbf{I_n}\right) - tr\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\right)\right]$$

$$= \sigma^2\left[n - tr\left(\underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'}_{\mathbf{I_{k+1}}}\right)\right]$$

$$= \sigma^2\left[n - tr\left(\mathbf{I_{k+1}}\right)\right]$$

$$E(\boldsymbol{\varepsilon'\varepsilon}) = \sigma^2\left(n - (k+1)\right).$$

Escrito em ordem à variância residual vem:

$$\sigma^2 = \frac{E(\boldsymbol{\varepsilon'\varepsilon})}{n - (k+1)},$$

logo conclui-se que o estimador não enviesado de σ^2 é dado por:

$$S^2 = \frac{SQE}{n - (k+1)}.$$

3.4 Previsão pontual e por intervalo

Uma das aplicações usuais dos modelos de regressão é a estimação do valor médio ou valor esperado da variável dependente y_i, dada em função de um vetor de observações das variáveis independentes. A reta de regressão permite predizer o valor estimado, ou seja, efetuar uma previsão pontual para y_i. Entretanto, em muitas circunstâncias, a aleatoriedade do modelo não garante que o valor previsto esteja sempre sobre a reta ajustada. Neste caso a qualidade das previsões passa por estabelecer intervalos de confiança para o valor esperado, ou seja, efetuar uma previsão por intervalo.

Deve-se ter em conta que, o modelo ajustado fornece apenas uma estimativa pontual da variável dependente com base nos dados da amostra onde o mesmo foi obtido. Contudo, a variável dependente é afetada pelo erro aleatório. Assim, no caso da previsão de novos valores da variável dependente que não fazem parte da amostra inicial de onde foi extraído o modelo de regressão, deve-se saber que a variância do valor esperado de y_i obtida com a amostra inicial não serve para efetuar inferências acerca deste novo valor da variável dependente. Por isso, é necessário quantificar qual é o erro associado a cada nova previsão, a partir do qual podem ser estabelecidos intervalos para novas previsões.

Os intervalos de confiança para a previsão dão uma margem de erro sobre o ponto ou intervalo de pontos previstos, servindo de guias de ajuda à construção dos modelos. Uma vez que os valores observados não pertencerem aos respetivos intervalos de previsão e cumulativamente se os valores estimados não forem reais, então o modelo não se comporta satisfatoriamente, devendo-se rever a sua estrutura (Pestana e Gageiro, 2014, p. 651).

3.5 Intervalos de confiança e testes relativos aos parâmetros do modelo

À semelhança do MRLS, no MRLM depois de se ajustar a função de regressão aos dados, a análise deve prosseguir com a determinação dos intervalos de confiança e realização de testes de hipóteses para os parâmetros do modelo.

3.5.1 Intervalos de confiança

Se os estimadores dos parâmetros do modelo forem funções lineares de y_i, centrados, eficientes e consistentes e se o erro do modelo possuir distribuição normal, então é possível definir as distribuições dos estimadores, elementos necessários para a definição das estatísticas do teste t, utilizadas na definição dos intervalos de confiança e nos testes de hipóteses para os parâmetros do modelo.

As distribuições amostrais dos estimadores dos parâmetros $\beta_j, j = 0, \ldots, k$, do MRLM são:

$$\begin{aligned}
\widehat{\beta}_0 &\sim N\left(\beta_0, var(\widehat{\beta}_0)\right); \\
\widehat{\beta}_1 &\sim N\left(\beta_1, var(\widehat{\beta}_1)\right); \\
&\vdots \\
\widehat{\beta}_k &\sim N\left(\beta_k, var(\widehat{\beta}_k)\right).
\end{aligned} \tag{3.37}$$

As variâncias, $var(\widehat{\beta}_1), \ldots, var(\widehat{\beta}_k)$ são definidas a partir da matriz de variâncias-covariâncias apresentada em 3.3.4, onde se obteve a expressão (3.28) que permite calcular a variância de cada estimador.

Com base nas distribuições amostrais e no estimador da variância residual S^2 (ver expressão (3.29)), podem-se definir as estatísticas para os parâmetros:

$$\begin{aligned}
T_0 &= \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{var(\widehat{\beta}_0)}} \sim t_{(n-k-1)}, \\
T_1 &= \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{var(\widehat{\beta}_1)}} \sim t_{(n-k-1)}, \\
&\vdots \\
T_k &= \frac{\widehat{\beta}_k - \beta_k}{\sqrt{var(\widehat{\beta}_k)}} \sim t_{(n-k-1)},
\end{aligned} \tag{3.38}$$

Com estas expressões é possível definir os intervalos de confiança e realizar testes de hipóteses para os parâmetros de regressão $\beta_j,\ j = 0, \cdots, k$.

• **Intervalo de confiança para os parâmetros do modelo**

Os procedimentos envolvidos na definição de intervalos de confiança para os parâmetros do MRLM são semelhantes aos considerados no modelo de regressão linear simples.

Assim, para os parâmetros $\beta_j\ j = 0, \cdots, k$, os intervalos de confiança bilaterais a $(1-\alpha) \times 100\%$ derivam da expressão:

$$\left[\widehat{\beta}_j - t_{1-\frac{\alpha}{2}, n-k-1} \times \sqrt{var(\widehat{\beta}_j)}; \widehat{\beta}_j + t_{1-\frac{\alpha}{2}, n-k-1} \times \sqrt{var(\widehat{\beta}_j)}\right], j = 0, \cdots, k, \tag{3.39}$$

em que $t_{1-\frac{\alpha}{2}, n-k-1}$, representa o quantil de ordem $(1 - \frac{\alpha}{2})$, da distribuição $t - Student$ com $(n - k - 1)$ graus de liberdade.

3.5.2 Testes relativos aos Parâmetros do MRLM

Na análise de regressão através do MRLM é necessário saber a influência que possui cada uma das variáveis independentes inseridas no modelo sobre a variável dependente, visto que, o modelo pode ser mais eficaz com a inclusão ou exclusão de determinadas variáveis. Para medir esta influência, são realizados os testes de hipóteses individuais aos parâmetros do modelo.

As hipóteses para testar a significância dos parâmetros do modelo de forma individual são as seguintes:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \quad j = 0, \cdots, k. \end{cases} \tag{3.40}$$

Quanto ao teste de hipóteses sobre β_0, caso não se rejeite H_0 o modelo passa a ter características de um modelo de regressão linear sem termo independente. Para os restantes parâmetros β_j, $j = 1, 2, \cdots, k$, caso não se rejeite H_0 a variável x_{ij} associada ao parâmetro não influência a resposta do modelo, por isso, deve ser excluída do mesmo.

Para averiguar a significância de cada parâmetro individualmente considera-se a estatística t. Caso se queira testar a significância de todos os parâmetros do modelo em simultâneo, recorre-se à ANOVA (Análise de variância) de regressão, que vamos abordar na secção a seguir.

3.6 Análise da variância

A análise de variância (ANOVA), neste contexto também designada por ANOVA de regressão, permite medir a qualidade de um determinado modelo de regressão, através da estatística do teste $F - Snedecor$, analisando a significância conjunta dos parâmetros do modelo.

Quando se efetua uma análise usando o MRLS, utiliza-se apenas o teste $t-Student$ para avaliar a significância do modelo, visto que, para o referido modelo as estatísticas dos testes $t - Student$ e $F - Snedecor$ dão o mesmo resultado. O mesmo não acontece com o MRLM, onde as variáveis independentes podem contribuir de forma diferente para o modelo. Assim, quando se trabalha com o MRLM usa-se o teste F para medir a significância conjunta dos parâmetros, pelo facto de o mesmo ter a possibilidade de produzir uma estatística de teste geral sobre a significância do modelo.

O teste F pode produzir uma estatística de teste que confirma a significância do modelo ajustado e ainda assim, existirem no mesmo variáveis não significativas. Nestes casos, faz-se recurso às estatísticas do teste t para identificar quais são as variáveis significativas e não significativas que o modelo ajustado possui.

Na regressão linear, a análise de variância consiste em dividir a variação total da variável dependente (y_i) em componentes com significado estatístico, que são tratadas de forma sistemática. Ou seja, trata de relacionar a soma de quadrados dos desvios totais (SQT), com a soma de qua-

drados dos desvios explicados pela regressão (SQR) e com a soma de quadrados dos resíduos (SQE). Esta relação é representada através da expressão:

$$SQT = SQR + SQE. \tag{3.41}$$

A seguir apresenta-se a tabela da ANOVA que sintetiza os resultados mais importantes desta técnica estatística:

Tabela 3.1: ANOVA

Fontes de variação	Variações	Graus de Liberdade	Desvios	Estatística de teste F
Regressão	$SQR = \Sigma_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$	k	$MQR = \frac{SQR}{k}$	$F = \frac{MQR}{MQE}$
Residual	$SQE = \Sigma_{i=1}^{n} (y_i - \widehat{y}_i)^2$	$n-k-1$	$MQE = \frac{SQE}{n-k-1}$	
Total	$SQT = \Sigma_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$		

No MRLM a ANOVA testa as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0 \\ H_1 : \exists_j : \beta_j \neq 0 \qquad\qquad j = 0, 1, \cdots, k \end{cases}. \tag{3.42}$$

Entretanto, para testar estas hipótese utiliza-se a estatística de teste,

$$F = \frac{\frac{SQR}{k}}{\frac{SQE}{n-k-1}} = \frac{MQR}{MQE} \tag{3.43}$$

com $\frac{SQR}{\sigma^2} \sim \chi^2_k$ e $\frac{SQE}{\sigma^2} \sim \chi^2_{n-k-1}$. Quando H_0 se verifica, a estatística F segue uma distribuição F central com k e $n-k-1$ graus de liberdade, $F_{(k,n-k-1)}$ (Murteira et al., 2015).

Com base nos resultados da estatística do teste F, rejeita-se a hipótese H_0, se $F_{obs} > f_{(1-a,n-k-1)}$, onde F_{obs} é o valor observado da estatística e $f_{(1-a,n-k-1)}$ o quantil $1-\alpha$ da distribuição F central com k e $n-k-1$ graus de liberdade. Quando se rejeita H_0, conclui-se que pelo menos uma das variáveis independentes inseridas no modelo contribui significativamente para o mesmo. Outra forma de verificar a significância das variáveis independentes, é através da análise do $p-value$ associado à estatística do teste F obtida. A regra consiste em rejeitar H_0 se $p-value \leqslant \alpha$, onde α é o nível de significância estabelecido para o teste.

3.7 Coeficiente de determinação

Tal como foi abordado na secção 2.5 do capítulo anterior, quando se está a trabalhar com modelos de regressão, uma das preocupações de qualquer investigador é de saber qual é o efeito que as variáveis independentes (ou variáveis explicativas) exercem sobre a variável dependente (ou variável resposta), ou seja, saber se o modelo proposto possui uma boa qualidade de ajustamento. Uma das ferramentas estatísticas utilizada habitualmente para analisar a qualidade de ajustamento é o Coeficiente de Determinação, simbolicamente representado por R^2, que pode ser também calculado através da expressão:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}, \tag{3.44}$$

que é análoga à expressão (2.24). O resultado desta expressão é interpretado da mesma maneira que se interpreta a expressão apresentada no capítulo anterior.

Na RLM o uso do R^2 deve ser acautelado, já que de modo geral, a inserção de mais uma variável independente no modelo tende a aumentar o valor de R^2 mesmo que esta variável possua um efeito reduzido sobre a variável dependente, Marôco (2014). Como alternativa ao R^2 deve usar-se o Coeficiente de Determinação Ajustado, simbolicamente representado por R^2_a, que é calculado através da expressão:

$$R^2_a = 1 - \frac{MQE}{MQT} = 1 - \frac{n-1}{n-k-1} \times \left(1 - R^2\right) = R^2 - \frac{k\left(1 - R^2\right)}{n-k-1}. \tag{3.45}$$

O coeficiente de determinação ajustado pode ser utilizado como o melhor estimador da qualidade de ajustamento e uma boa medida da dimensão do efeito das variáveis independentes sobre a variável dependente. Enquanto que o R^2 tende a aumentar com a inserção de uma variável independente, o R^2_a só aumenta, se esta variável conduzir a um melhor ajustamento do modelo aos dados, isto é, quando a variância do erro (MQE) diminui em relação à variância total (MQT).

Por outro lado, R^2_a é tido como o melhor estimador do coeficiente de determinação na população do que R^2. Deste modo, R^2_a dá a percentagem de variabilidade da variável dependente que é explicada pelo modelo, como se o modelo tivesse sido obtido a partir da população, em alternativa a uma amostra representativa desta população. O R^2_a também pode ser interpretado como uma medida da capacidade de generalização do modelo.

3.8 Métodos de seleção das variáveis independentes para o MRLM

Quando se trabalha com o MRLM, por vezes o investigador dispõe de um número elevado de variáveis independentes que hipoteticamente podem influenciar a variável dependente. Por vários motivos, depois da introdução de todas elas no modelo, pode notar-se que parte delas não são significativas, ou mesmo até, antes de introduzi-las, se tenha a necessidade de escolher apenas parte destas variáveis para serem incluídas no modelo. Quais variáveis independentes podem ser incluídas? A resposta depende do investigador. Porém, em muitas ocasiões recorre-se aos métodos de seleção das variáveis independentes, que se baseiam nos testes individuais de significância aos submodelos de regressão com diferentes números de variáveis independentes, para escolher o melhor modelo que se ajusta aos dados.

Entretanto, ao se efetuar a análise com o objetivo de se selecionar variáveis independentes para um determinado MRLM, é aconselhável utilizar mais do que um método e depois identificar quais são as variáveis que terão sido escolhidas por alguns ou por todos os métodos em simultâneo. Estas variáveis poderão definir o modelo adequado, enquanto que as restantes devem ser eliminadas. Aconselha-se o uso do maior número possível de métodos existentes, pelo facto de a experiência mostrar que entre eles, não existe nenhum que é melhor que outro (Marôco, 2014, p. 718).

A vantagem que se tem ao utilizá-los é que em vez de o investigador decidir por um critério qualquer a retirada de uma ou outra variável do modelo, estes métodos, com base num critério baseado em testes estatísticos, indicam quais são as variáveis independentes que apresentam relações significativas com a variável dependente, escolhendo-as para o modelo, enquanto que as restantes são excluídas do mesmo.

De seguida é feita uma descrição dos três métodos mais utilizados para se efetuar a seleção das variáveis independentes.

• **Método *Forward***

Quando se usa o método *Forward* (ou Progressivo) para se efetuar a seleção das variáveis independentes, a equação inicial do modelo começa apenas com a constante. A seguir são adicionadas de forma progressiva as variáveis independentes $x_{ij},\ j = 1, \cdots, k$, por ordem dos seus graus de correlação (em valor absoluto) com a variável dependente, começando por aquelas que tiverem maior grau de correlação, desde que satisfaçam os critérios estatísticos de entrada previamente estabelecidos. Este procedimento continua até que todas as variáveis sejam inseridas no modelo, ou quando uma (ou mais variáveis) seja excluída do modelo por não satisfazer os critérios de entrada.

• **Método *Backward***

Neste método, começa-se por incluir no modelo todas as variáveis independentes. Após um teste individual de significância, são retiradas do modelo, uma a uma, as variáveis cuja pre-

sença não contribui para explicar uma proporção significativa da variação total da variável dependente. O processo de exclusão das variáveis independentes não significativas consiste em retirar em primeiro lugar a variável menos significativa e assim sucessivamente, mantendo no modelo apenas as que forem significativas.

• **Método *Stepwise***

Este método funde alguns procedimentos dos dois métodos anteriores. Começa tal como no método *Forward*, com a inserção de apenas uma variável independente, porém, as outras variáveis a serem inseridas no modelo, são testadas como no método *Backward*. A vantagem deste método consiste em permitir a remoção de uma variável cuja importância no modelo é reduzida por causa da adição das novas variáveis. Este procedimento termina quando nenhuma das variáveis independentes que estiver fora consegue entrar, por não possuir um valor da estatística F maior do que o valor crítico pré-definido para entrada no modelo, e nenhuma das variáveis independentes presentes no modelo for retirada. Este método é recomendável quando existem suspeitas de correlações significativas entre as variáveis independentes.

Como se pode notar, cada um dos três métodos apresentados utiliza critérios de seleção distintos dos outros, o que faz com que, ao longo do processo de seleção, uma variável importante para o modelo, possa ser penalizada por um ou outro método, dai a recomendação da utilização de todos e depois escolher o melhor modelo proposto por um ou em simultâneo por alguns ou todos os métodos.

3.9 Breve introdução ao uso de variáveis qualitativas no MRLM

Quando se utiliza o MRLM, os investigadores são muitas vezes confrontados com dados, cuja natureza das variáveis é qualitativa (nominal ou ordinal) em que a inclusão destas variáveis no modelo é imprescindível, para uma melhor explicação do fenómeno em análise. Se eventualmente estas variáveis forem independentes, é possível incluí-las no modelo com recurso às variáveis auxiliares indicadoras, também conhecidas como variáveis artificiais, variáveis zero e um ou ainda variáveis *dummies* (como vão ser tratadas a seguir).

As variáveis *dummies* podem ser usadas em modelos que contenham dados seccionais ou cronológicos. Com qualquer tipo de dados em que se estiver a trabalhar, a construção de modelos de regressão linear com variáveis *dummies* deve atender a determinadas situações como é o caso do ajustamento de modelos com variações contínuas ou descontínuas nos parâmetros e a análise de sazonalidade (Valle e Rebelo, 2002, p. 21).

Uma variável nominal (ou até mesmo ordinal) com n categorias, para ser inserida no modelo é necessário construir $n-1$ variáveis *dummies*. Por exemplo, para um modelo que contenha a variável ordinal **nível académico** que possui três categorias (básico, secundário e superior), devem ser inseridas neste modelo duas variáveis *dummies*, que representam duas das três categorias da variável nominal e a terceira passa a ser considerada categoria de referência.

Assim, para modelar uma variável qualitativa (nominal ou ordinal) com duas categorias, basta para o efeito definir uma variável *dummy* (que por convenção assume os valores 0 e 1) associada a um determinado acontecimento D da seguinte forma:

$$d_i = \begin{cases} 1, & se\, D\, se\, verifica \\ 0, & se\, D\, não\, se\, verifica \end{cases} . \tag{3.46}$$

Neste caso a variável d_i é o exemplo de uma variável *dummy*. A escolha dos valores 0 e 1 é arbitrária, mas por convenção, o nome da variável binária deve ser o nome da categoria a que corresponde o valor 1. Por exemplo, se o conjunto é $D = \{Estudantes\ do\ sexo\ masculino\}$ então, tem-se: $d_i = 1$ (*se é estudante do sexo masculino*) e $d_i = 0$ (se é estudante do sexo feminino).

Como se afirmou acima, a informação qualitativa trazida pela variável *dummy* só pode ser introduzida num MRLM caso a variável que a contém seja uma variável independente. Por exemplo, no caso em que se considere o MRLM dado por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

onde a variável x_{i2} representa a variável *dummy* (d_{i2}) associada ao acontecimento D com duas categorias, o modelo de regressão com a variável *dummy* define-se do seguinte modo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \delta_2 d_{i2} + \varepsilon_i, \tag{3.47}$$

onde δ_2 é a diferença de termos independentes ($\delta_2 = \beta_2 - \beta_0$) e d_{i2} é a variável *dummy*, que assume o valor 1 quando se verifica o acontecimento D, e o valor 0 quando não se verifica esse acontecimento. Como resultado, obtêm-se os seguintes submodelos:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i,\ para\, d_{i2} = 0 \\ \\ y_i = (\beta_0 + \delta_2) + \beta_1 x_{i1} + \varepsilon_i,\ para \quad d_{i2} = 1 \end{cases} . \tag{3.48}$$

No modelo (3.47) constata-se que a variável qualitativa tem efeito apenas no termo independente e não há interação entre as variáveis independentes. Além disso, a interpretação dos parâmetros β_0 e $(\beta_0 + \delta_2)$ é diferente daquela que se faz habitualmente, visto que β_0 representa o termo independente caso o acontecimento D não se verifique e $(\beta_0 + \delta_2)$ é o termo independente caso este acontecimento se verifique.

Deste modo o parâmetro δ_2 representa a variação em média de y_i, quando se passa de $\overline{D}(\ d_{i2} = 0)$ para $D\ (d_{i2} = 1)$. Assim, se $\delta_2 > 0$, y_i cresce; se $\delta_2 = 0$, y_i não varia e se $\delta_2 < 0$, y_i decresce.

O modelo de regressão linear com variável *dummy* apresentado é apenas um dos tipos de modelo que ilustra o uso destas variáveis. Como se pode notar neste modelo, ajustaram-se duas retas

com o mesmo declive (β_1) e intercetos ou termos independentes diferentes, (β_0 quando $d_{i2} = 0$ e $(\beta_0 + \delta_2)$ quando $d_{i2} = 1$). A estimação do modelo $y_i = \beta_0 + \beta_1 x_{i1} + \delta_2 d_{i2} + \varepsilon_i$, em alternativa à estimação separada dos dois submodelos, definidos em (3.48) tem por consequência garantir que o coeficiente β_1, comum nos dois submodelos, seja estimado de forma única utilizando toda a informação disponível. Este tipo de modelo é conhecido como modelo de regressão linear com variável *dummy* de interceto. De seguida tem-se a representação geométrica da estrutura estimada do mesmo (Figura 3.2).

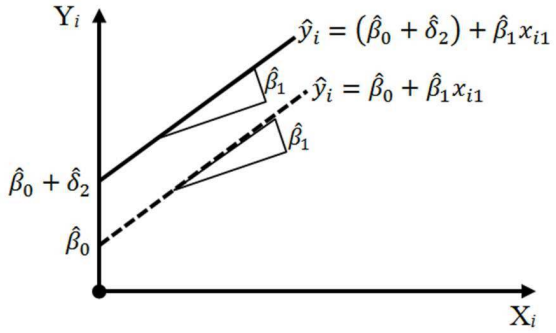

Figura 3.2: Representação geométrica da estrutura estimada do modelo de regressão linear com uma variável *dummy*, com efeito no termo independente

Porém, esta não é a única forma que existe de incluir as variáveis independentes qualitativas num modelo de regressão linear. Existem situações onde o efeito do fator qualitativo se dá apenas nos coeficientes das variáveis independentes quantitativas (declives) β_j, $(j = 1, 2, , \cdots, k)$, e o termo independente, β_0, se mantém fixo. Nestes casos, por exemplo, obtém-se o seguinte modelo:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \delta d_{i1} x_{i1} + \varepsilon_i \\ &= \beta_0 + (\beta_1 + \delta d_{i1}) x_{i1} + \varepsilon_i, \end{aligned} \tag{3.49}$$

em que d_{i1} (variável *dummy*) assume o valor 1 se se verifica o acontecimento D e o valor 0 no caso contrário. Neste caso $d_i x_{i1}$ é considerada a variável de interação. Isto resulta nos seguintes submodelos,

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \; para \, d_{i1} = 0 \\ \\ y_i = \beta_0 + (\beta_1 + \delta) x_{i1} + \varepsilon_i, \; para \quad d_{i1} = 1 \end{cases} . \tag{3.50}$$

Neste modelo de regressão linear, ajustaram-se duas retas com o mesmo valor do termo independente, β_0, e declives diferentes, β_1 e $(\beta_1 + \delta)$. Este tipo de modelo é conhecido como modelo de regressão linear com variável *dummy* de inclinação. A Figura 3.3 mostra a representação geométrica da estrutura estimada deste tipo de modelo.

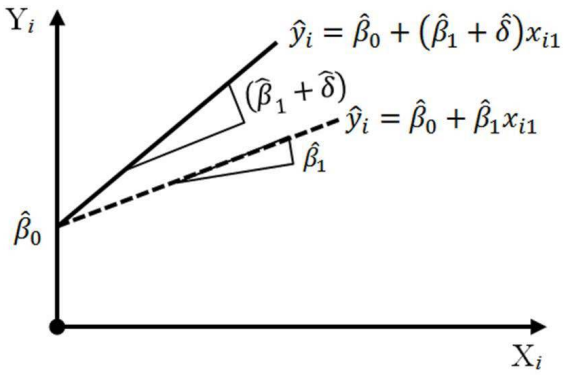

Figura 3.3: Representação geométrica da estrutura estimada do modelo de regressão linear com uma variável *dummy*, com efeito no declive

Os modelos de regressão linear com variáveis *dummies* de inclinação e de interceto, podem ser generalizados para casos em que simultaneamente se podem registar alterações dos valores dos termos independentes e dos declives como, por exemplo, no seguinte caso:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \delta d_{i1} + \gamma x_{i1} d_{i1} + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + (\delta + \gamma x_{i1}) d_{i1} + \varepsilon_i. \end{aligned} \tag{3.51}$$

Neste caso, quando d_{i1} (variável *dummy*) assume o valor 1 verifica-se o acontecimento D e quando assume o valor 0, não se verifica esse mesmo acontecimento. Isto resulta nos seguintes submodelos:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \ para\, d_{i1} = 0 \\ \\ y_i = (\beta_0 + \delta) + (\beta_1 + \gamma) x_{i1} + \varepsilon_i, \quad para\ d_{i1} = 1. \end{cases} \tag{3.52}$$

Geometricamente, este tipo de modelo possui a seguinte estrutura estimada:

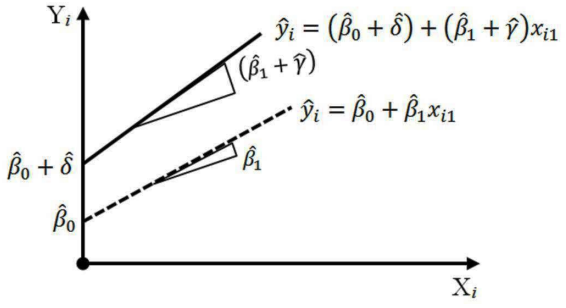

Figura 3.4: Representação geométrica da estrutura estimada do modelo de regressão linear com variável *dummy*, com efeito simultâneo no declive e no termo independente

As três maneiras apresentadas de como se pode definir modelos de regressão linear com variáveis *dummy*, não são as únicas que existem. Algumas delas dependem do tipo de dados que estiverem a ser analisados. No entanto, para este trabalho, termina aqui a breve resenha que se reservou para este conteúdo.

3.10 Conclusão

Os conteúdos que acabamos de apresentar sobre o modelo de regressão linear múltipla não encerram toda a teoria que aborda esta metodologia estatística, porém, na nossa opinião, acredita-se que o mesmo possui o indispensável para ajudar os alunos e professores do ensino secundário e universitário a compreenderem os aspetos essenciais relacionados a este tópico.

Entretanto, para além da estimação dos parâmetros, dos intervalos de confiança, dos testes de hipóteses, dos métodos de seleção de variáveis independentes, da inserção de variáveis *dummies* e outros tópicos não menos importantes já apresentados, a abordagem reservada neste trabalho sobre o MRLM abarca ainda conteúdos ligados à validação dos pressupostos impostos ao erro do modelo, que vão ser abordados no capítulo a seguir.

Capítulo 4

Validação dos pressupostos do modelo de regressão

Depois da estimação dos coeficientes do modelo de regressão linear, a análise de regressão deve continuar, com a validação dos pressupostos do modelo. Caso estes não sejam cumpridos, o modelo de regressão estimado é considerado impróprio para explicar ou predizer o valor da variável dependente em função das variáveis independentes que o compõem.

Neste capítulo apresentam-se as técnicas que permitem verificar a normalidade, a homoscedasticidade (ou variância constante) e a independência dos resíduos. Far-se-á também uma abordagem sobre a existência de colinearidade e/ou multicolinearidade e o diagnóstico de *outliers* e observações influentes.

4.1 Análise de resíduos

Nos capítulos anteriores afirmou-se que os valores das estimativas dos resíduos têm um papel importante na avaliação final da qualidade do modelo ajustado, visto que, para além de serem usados na estimação dos parâmetros do modelo, são usados também para verificarem os pressupostos impostos ao erro do modelo e deteção de *outliers* e observações influentes.

As técnicas de diagnóstico dos pressupostos impostos ao erro do modelo usam diferentes tipos de resíduos, tipos esses que passamos a mencionar a seguir.

4.1.1 Tipos de resíduos

Os tipos de resíduos usados habitualmente são:

- **Resíduos originais ou resíduos não estandardizados**: no caso geral da análise dos modelos de RLS e RLM são obtidos pela expressão $\varepsilon_i = y_i - \widehat{y}_i,\ i = 1, 2, \cdots, n$. Este tipo de resíduos, para além de serem usados na estimação dos parâmetros do modelo, tal como se mostrou nos capítulos anteriores, são também usados, por exemplo, na análise da independência e da homoscedasticidade dos resíduos. É a partir destes que se obtêm outros tipos de resíduos.

- **Resíduos estandardizados**: a padronização ou estandardização dos resíduos tem sido utilizada com o objetivo de eliminar as diferenças entre as variações das variáveis, permitindo assim que se possam comparar diretamente os parâmetros de regressão estimados. Os resíduos padronizados para cada observação i são obtidos através da expressão:

$$r_i = \frac{\varepsilon_i}{S}, \quad i = 1, 2, \cdots, n, \tag{4.1}$$

onde S é a raiz quadrada do estimador S^2 apresentado em (3.29). Este tipo de resíduos é usado, por exemplo, para verificar o pressuposto da normalidade.

- **Resíduos estudantizados**: os resíduos estudantizados são usados com o objetivo de verificar se os valores das observações estão adequadamente ajustados ao modelo e mostrar quais as observações cujos valores não estão em concordância com os demais, os chamados *outliers*. São definidos por (Montgomery, Peck e Vining, 2012, p. 130):

$$t_i = \frac{\varepsilon_i}{S\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \cdots, n, \tag{4.2}$$

onde S é a raiz quadrada do estimador S^2 e h_{ii} representa o i-ésimo elemento da diagonal da matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$.

- **Resíduos excluídos**: estes resíduos são usados na análise de resíduos para detetar *outliers*. Os resíduos excluídos $\varepsilon_{(-i)}$, obtêm-se estimando o valor de y_i, quando se retira do modelo a observação i (Fernandes, 1999),

$$\varepsilon_{(-i)} = \frac{\varepsilon_i}{1 - h_{ii}}, \; i = 1, 2, \cdots, n. \tag{4.3}$$

- **Resíduos estudantizados excluídos**: À semelhança dos resíduos excluídos estes resíduos também são usados com a finalidade de detetar *outliers*. São obtidos através da expressão:

$$t_{(-i)} = \frac{\varepsilon_i}{S_{(-i)}\sqrt{1 - h_{ii}}}, \; i = 1, 2, \cdots, n, \tag{4.4}$$

onde $S_{(-i)}$ é a raiz quadrada do estimador S^2 calculado sem o resíduo i.

4.1.2 Verificação dos pressupostos impostos ao erro do modelo de regressão

A verificação dos pressupostos impostos ao erro de um modelo de regressão comporta a realização do diagnóstico da normalidade, homoscedasticidade e independência dos resíduos.

4.1.2.1 Diagnóstico da normalidade

A normalidade dos resíduos permite obter a distribuição dos estimadores de Mínimos Quadrados, utilizada na estimação dos intervalos de confiança e realização dos testes de hipóteses sobre os parâmetros do modelo.

Para diagnosticar a normalidade são empregues métodos gráficos, tais como, os gráficos Q-Q ou P-P dos resíduos e o histograma dos resíduos estandardizados e os testes não paramétricos de aderência, nomeadamente os testes de Kolmogorov-Smirnov (K-S), Shapiro-Wilk (S-W), entre outros. A seguir, vamos descrever a forma como parte destas técnicas podem ser usadas para diagnosticar a normalidade dos resíduos.

• Os gráficos Q-Q e P-P dos resíduos

O gráfico Q-Q, representa o quantil de probabilidade esperado se a distribuição é normal em função dos valores observados. Neste gráfico comparam-se os valores observados com os valores esperados de uma distribuição normal, representados por uma reta diagonal. Quando os resíduos possuem uma distribuição normal, as observações apresentam-se próximo dessa reta diagonal, sem afastamentos consideráveis.

Já o gráfico P-P, representa a probabilidade acumulada esperada caso a distribuição seja normal, em função da probabilidade observada acumulada dos resíduos. Também neste gráfico, caso exista uma boa aderência dos resíduos à distribuição Normal, os pontos aparecem ao redor da reta de referência apresentada no gráfico.

Nas figuras apresentadas abaixo podem-se observar exemplos destes dois gráficos, em que não se rejeita a normalidade dos resíduos.

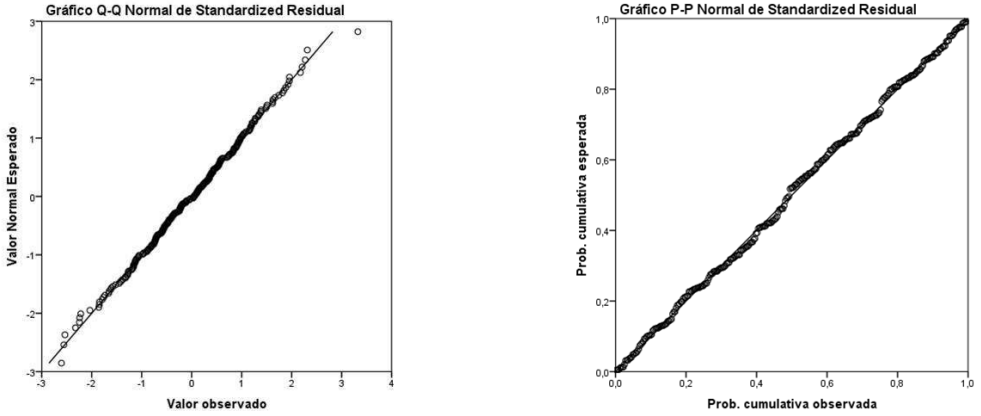

Figura 4.1: Gráficos normal Q-Q e normal P-P dos resíduos

• Histograma dos resíduos estandardizados

O histograma dá uma perceção clara se a forma de sino, característica da distribuição Normal, está presente no gráfico ou não. Caso as barras se ajustem bem à forma de sino, conclui-se que os dados possuem uma distribuição aproximadamente normal, caso contrário, essa conclusão não pode ser retirada. Entretanto, a utilização do histograma é recomendável apenas para amostras de dimensão elevada, já que em amostras de pequena dimensão existem sempre algumas indefinições.

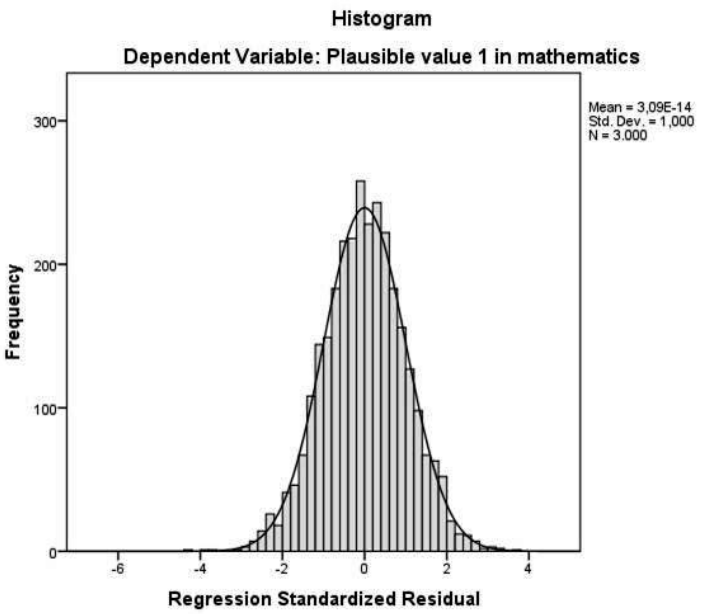

Figura 4.2: Histograma dos resíduos padronizados

A análise gráfica é em algumas situações suficiente para se tomar a decisão sobre a normalidade ou não dos resíduos. Porém, os métodos gráficos têm sempre algumas desvantagens pelo facto de serem um tanto ou quanto subjetivos. Por isso mesmo, muitas vezes para se decidir quanto à existência da normalidade, usam-se testes não paramétricos de aderência à distribuição Normal, já mencionados anteriormente. Vamos considerar a seguir dois destes testes.

• **Kolmogorov-Smirnov (K-S)**

O teste de Kolmogorov-Smirnov ou (K-S) é um teste de aderência que verifica o grau de concordância entre distribuições, considerando um conjunto de valores. Tem como objetivo identificar se os dados seguem uma determinada distribuição. Este teste utiliza a distribuição de frequência acumulada, que ocorre dada a descrição teórica, e compara-a com a distribuição de frequência acumulada observada.

O teste não paramétrico de aderência à normalidade de Kolmogorov-Smirnov testa as seguintes hipóteses:

$H_0 : X$ tem distribuição normal.
$H_1 : X$ não tem distribuição normal.

Para testar as hipóteses, pode-se recorrer a softwares estatísticos como, por exemplo, o $SPSS$ e R, ou consultar tabelas de valores críticos da distribuição da estatística,

$$D = \underbrace{Supremo}_{x} \mid S(x) - F_0(x) \mid \tag{4.5}$$

onde $F_0(x)$ é a função de distribuição populacional e $S(x)$ a função de distribuição da amostra. Por exemplo, com um nível de significância de $\alpha = 0,05$ para amostras que vão de 1 a 40

observações, a estatística de teste apresenta valores críticos que variam entre 0.975 a 0.189, enquanto que para amostras com mais de 40 observações, os valores críticos são dados por, $\frac{1.36}{\sqrt{n}}$ (ver Guimarães e Cabral (1998) e Fernandes (1999)).

• **Shapiro-Wilk (S-W)**

Este teste deve ser usado em substituição do K-S quando se tem uma amostra de pequena dimensão ($n < 30$). O teste de Shapiro-Wilk é dado pela estatística (Marôco, 2014),

$$W = \frac{\left(\Sigma_{i=1}^{n} a_i x_i\right)^2}{\Sigma_{i=1}^{n} \left(x_i - \overline{x}\right)^2}, \tag{4.6}$$

onde x_i representa os valores das observações da variável X ordenados de maneira crescente, $\overline{x}$ é a média destes valores e a_i são as constantes geradas a partir da média, variância e covariância de n ordens com distribuição normal, $N(0,1)$. Os valores de a_i vêm expressos em tabelas.

O mesmo testa as seguintes hipóteses:

$H_0 : X$ tem distribuição normal.
$H_1 : X$ não tem distribuição normal.

Para validar as hipóteses é necessário observar os valores da estatística W. Valores pequenos de W indicam que a distribuição da variável em estudo não é Normal. Os valores críticos para W vêm também em tabelas publicadas em vários manuais de estatística que constam das referências bibliográficas deste trabalho ver, por exemplo (Guimarães e Cabral, 1998). Também para a realização deste teste, podem ser usados os softwares já citados acima.

4.1.2.2 Diagnóstico da homoscedasticidade

O termo homoscedasticidade é usado na regressão para denotar a variância constante dos resíduos em observações diferentes. Quando a homoscedasticidade é violada, diz-se que o modelo apresenta heteroscedasticidade. Neste caso, os estimadores dos mínimos quadrados, apesar de se manterem lineares e não enviesados, deixam de ser estimadores de variância mínima, o que dá lugar a subestimação ou sobrestimação das verdadeiras variâncias dos estimadores ($S_{\widehat{\beta}_i}$), resultando num coeficiente de determinação ajustado (R_a^2) enganador. Esta perda de eficiência tem como consequência a não validação da inferência estatística baseada nos testes t e F, já que nestas condições as estatísticas destes testes tornam-se duvidosas.

O diagnóstico da homoscedasticidade dos resíduos pode ser efetuado com recurso a técnicas gráficas e testes estatísticos. No MRLM para diagnosticar a homoscedasticidade usa-se o gráfico dos resíduos estudantizados com o valor previsto da variável dependente na forma estandardizada (de seguida apresenta-se um exemplo deste gráfico, Figura 4.3) ou na forma não estandardizada (já que apresentam todos a mesma configuração e neste sentido bastará usar apenas um deles).

E no MRLS, para além dos gráficos usados no MRLM, também se pode usar o gráfico dos resíduos não estandardizados com a variável independente na qual se supõe existir desigualdade de variâncias.

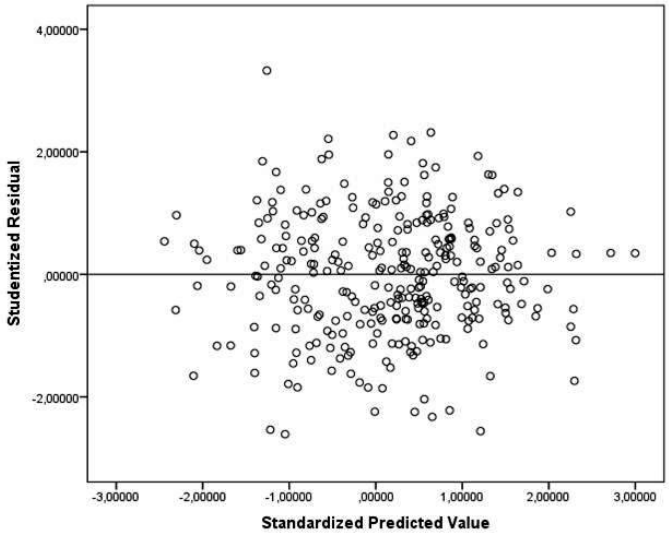

Figura 4.3: Gráfico dos resíduos estudantizados versus valores preditos estandardizados

Na análise gráfica, caso as observações se distribuam de forma aleatória em torno da linha horizontal com origem em zero (como se vê na Figura 4.3), supõe-se haver homoscedasticidade, enquanto que quando os resíduos se distribuem seguindo um determinado padrão (formando, por exemplo, um triângulo), supõe-se haver heteroscedasticidade. Além da deteção da heteroscedasticidade, esse gráfico pode indicar que não existe uma relação linear entre a variável explicativa (ou resposta) e a variável explicada por meio de alguma tendência nos pontos.

Quanto aos testes, os mais usados são o de Goldfeld e Quandt e o de White. Este último segundo Caiado (2016) assenta basicamente nos seguintes passos:

• **Teste de White**

1º passo

Estima-se o modelo original pelo método dos mínimos quadrados como, por exemplo, o que se tem abaixo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, \tag{4.7}$$

e em seguida obtêm-se os resíduos, $\varepsilon_i = y_i - \widehat{y}_i,\ i = 1, 2, \cdots, n$.

2º Passo

Estima-se a regressão auxiliar do quadrado dos resíduos sobre as variáveis independentes e os seus quadrados (incluindo o termo independente)

$$\varepsilon_i^2 = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_4 x_{i4} + \alpha_5 x_{i1}^2 + \ldots + \alpha_8 x_{i4}^2 + \upsilon_i. \tag{4.8}$$

Caso se tenham muitas observações, pode-se incluir nesta regressão os produtos cruzados, $x_{i1}x_{i2}$, $x_{i2}x_{i3}$, $x_{i3}x_{i4}$, o que resultaria em:

$$\varepsilon_i^2 = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_4 x_{i4} + \alpha_5 x_{i1}^2 + \ldots + \alpha_8 x_{i4}^2 + \alpha_9 x_{i1}x_{i2} + \alpha_{10} x_{i2}x_{i2} + \alpha_{11} x_{i3}x_{i4} + v_i. \quad (4.9)$$

Em seguida, calcula-se o respetivo coeficiente de determinação (R^2).

3º Passo

Testa-se a hipótese nula $H_0 : \alpha_0 = \alpha_1 = \ldots = \alpha_k = 0$, de ausência de heteroscedasticidade, com base na estatística nR^2, que apresenta uma distribuição qui-quadrado com $k-1$ graus de liberdade, isto é, $nR^2 \sim \chi^2_{k-1}$.

Para corrigir a heteroscedasticidade pode usar-se outro tipo de regressão como, por exemplo, a dos mínimos quadrados ponderados, também conhecida como Weighted Least Squares (WLS), (Pestana e Gageiro, 2014), cuja abordagem não vai ser considerada neste trabalho.

4.1.2.3 Diagnóstico da independência

Um dos requisitos para validação de um modelo de regressão linear, é que os seus resíduos não devem estar autocorrelacionados, ou seja, devem ser independentes.

Para detetar a presença da autocorrelação entre os resíduos do modelo, podem ser usados métodos gráficos (recorrendo ao gráfico dos resíduos) ou testes estatísticos. O teste estatístico que habitualmente se utiliza é o teste de Durbin-Watson, cuja estatística é dada pela seguinte expressão (Marôco, 2014):

$$d = \frac{\sum_{i=1}^{n} \left(\varepsilon_{i+1} - \varepsilon_i\right)^2}{\sum_{i=1}^{n} \varepsilon_i^2} \approx 2\left(1 - r_{\varepsilon_{i+1};\varepsilon_i}\right), \quad (4.10)$$

onde, ε_i representa o resíduo associado à observação i e $r_{\varepsilon_{i+1};\varepsilon_i}$ representa a autocorrelação dos resíduos da amostra. Os valores da estatística d variam entre 0 e 4. Se o resultado for aproximadamente 2, pode-se concluir que não existe autocorrelação entre os resíduos. Se for muito menor que 2 existe autocorrelação positiva e para valores muito maiores de 2 existe autocorrelação negativa. O valor da estatística d permite tomar uma decisão sobre as seguintes hipóteses:

H_0 : Não existe autocorrelação dos resíduos.

H_1 : Existe autocorrelação dos resíduos (positiva ou negativa).

Uma forma mais exata de validar os resultados deste teste proposta por Durbin e Watson, consiste em comparar o valor de d com um limite inferior, d_L, e um limite superior, d_U, para testar H_0. Existem tabelas que auxiliam na tomada de decisões em função dos valores de d, d_L, e d_U e outras que dão os valores críticos d_L e d_U, para um determinado nível de significância α (ver Tabela A.1, Anexo 1).

Na presença da autocorrelação, os estimadores de mínimos quadrados continuam a ser não enviesados e consistentes, porém, deixam de ser eficientes, o que faz com que os resultados das estatísticas dos testes t e F sejam duvidosas e consequentemente, deixa de ser válida a inferência estatística sobre os parâmetros do modelo.

4.1.3 Diagnóstico de *outliers* e observações influentes

• ***Outliers***

Os *outliers* são observações com um comportamento diferente das demais, que estão associadas a resíduos com valores elevados. Num modelo de regressão os *outliers* devem ser sempre identificados, porque se não resultarem de erros de inserção de dados, permitem conhecer características únicas ou novos segmentos válidos da população, que de outro jeito não seriam conhecidos.

Por norma, os *outliers* devem ser removidos da análise se existirem provas de que não são elementos válidos da população. Caso sejam elementos válidos da população, a decisão da sua inclusão ou não no modelo depende do tipo de informação que possam dar e do contexto ou população em que se inserem. A sua exclusão pode melhorar a qualidade do modelo, porém, pode também provocar o risco de limitar a generalização do estudo que estiver a ser feito à população em causa.

O diagnóstico de *outliers* pode ser feito através das estatísticas dos resíduos ou por meio de gráficos. Com um nível de significância de $\alpha = 0.05$, valores dos resíduos estandardizados, estudantizados e estudantizados excluídos superiores a 1.96, indicam a existência de *outliers*. Quando se efetua a análise gráfica por meio do gráfico dos resíduos estudantizados excluídos versus valores preditos estandardizados, as observações da amostra consideradas como *outliers*, usualmente tendem a posicionar-se em zonas do gráfico distante de onde estão concentradas a maioria das observações, como se pode notar no gráfico representado pela Figura 4.4.

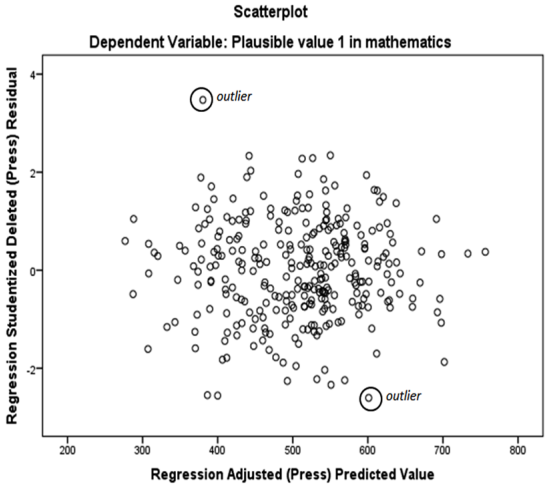

Figura 4.4: Gráfico de resíduos estudantizados excluídos versus valor predito ajustado

Os efeitos dos *outliers* podem ser considerados moderados ou severos. Segundo Pestana e Gageiro (2005), os *outliers* moderados são aqueles que se encontram entre 1,5 e 3 amplitudes inter-quartis para baixo do 1º quartil ou para cima do 3º quartil, e os severos são aqueles que se situam para além destes limites.

Entretanto, segundo Marôco (2014) usando os valores do *Leverege* (medida que vamos abordar já a seguir nas observações influentes) valores do *Leverage*, inferiores a 0.2, são aceitáveis para que os *outliers* não sejam considerados severos.

• **Observações influentes**

Uma observação influente é aquela que por alguma razão causa grandes mudanças em alguns ou em todos os parâmetros do modelo, quando é omitida do conjunto de dados. Uma das medidas usadas para testar a influência das observações no modelo é o *Leverage* (alavancagem), cujo valor depende da quantidade de observações da amostra, n, e do número de variáveis independentes, k.

Assim, usando esta medida, segundo Pestana e Gageiro (2014) pode-se considerar que no modelo existem observações influentes se:

- Para amostras com $n \leq 30$:
$$LEV > \frac{3(k+1)}{n}, \tag{4.11}$$

- E para amostras com $n > 30$:
$$LEV > \frac{2(k+1)}{n}, \tag{4.12}$$

em que n é a dimensão da amostra e k o número de variáveis independentes.

Contudo, devido às implicações que as observações influentes podem causar no modelo foram desenvolvidas outras técnicas que ajudam a identificá-las, tais como os $DFFIT$ (Difference in Fit), $DFBETA$ (Difference in Beta), e a *Distância de Cook*, que a seguir passamos a descrever.

$DFFIT$: Esta medida possibilita medir a influência que a observação i tem sobre o seu próprio valor ajustado. Os DFFIT são calculados através da seguinte expressão:

$$DFFIT_{(i)} = \frac{\widehat{y}_i - \widehat{y}_{i(i)}}{\sqrt{QME_{(i)} \times h_{ii}}},\ i = 1, \cdots, n, \tag{4.13}$$

onde $\widehat{y}_i$ é a previsão para a observação (i) com i incluído na regressão, $\widehat{y}_{i(i)}$ é a previsão para a observação (i) com i não incluído na regressão, o $QME_{(i)}$ representa o erro quadrático médio calculado sem a observação (i), e h_{ii} representa a alavancagem (*Leverage*). Esta técnica mede o quanto a inclusão da observação (i) aumenta ou diminui o seu valor predito.

Diz-se que um *outlier* é influente segundo o $SDFFIT$ (Difference in Standardized Fits) se:

- $|SDFFIT| > 1,96,$ para amostras de pequena dimensão ($n \leq 30$);

- $|SDFFIT| > 2\sqrt{\dfrac{k+1}{n-k-1}}$, para amostras de grande dimensão ($n > 30$).

$DFBETA$: esta medida indica o quanto o coeficiente de regressão β_j altera em unidades de desvio padrão, se a observação i for excluída da análise. É definida por, (ver Montgomery, Peck e Vining (2012, p. 217)):

$$DFBETA_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{MQE_{(i)} \times c_{jj}}},\ i = 1, \cdots, n;\ \ j = 1, \cdots, k, \tag{4.14}$$

onde $\widehat{\beta}_{j(i)}$ é o coeficiente de regressão j calculado sem o uso da observação (i) e c_{jj} é o $j-$ésimo elemento da diagonal da matriz $(\boldsymbol{X'X})^{-1}$, onde $\boldsymbol{X}$ é a matriz definida em (3.2).

Uma observação é considerada influente quando (Pestana e Gageiro, 2005) :

- $|SDFBETA| > 1,96$, para amostras de pequena dimensão ($n \leq 30$);
- $|SDFBETA| > \dfrac{2}{\sqrt{n}}$, para amostras de grande dimensão ($n > 30$),

onde a medida $SDFBETA$ (Difference in Standardized Betas), corresponde à $DFBETA$ estandardizada.

$DIST\hat{A}NCIA\ DE\ COOK$: esta medida permite avaliar a influência da observação i sobre todos os valores ajustados $\widehat{y}_i,\ \ i = 1, \cdots, n$, e é definida por:

$$D_i = \frac{\varepsilon_i h_{ii}}{(k+1)(1-h_{ii})},\ \ i = 1, \cdots, n, \tag{4.15}$$

onde h_{ii} é o i-ésimo elemento da diagonal da matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ e k é o número de parâmetros do modelo.

Pode-se notar que D_i é grande, quando ε_i é grande ou a medida h_{ii} é grande ou ainda se ambos forem grandes.

A distância de Cook pode ser testada ainda através da relação apresentada de seguida:

$$COOK > \frac{4}{n-k-1}, \tag{4.16}$$

onde n é a dimensão da amostra e k o número de variáveis independentes. Uma observação é excessivamente influente se tiver um valor da distância de Cook superior a 1 (Marôco, 2014).

4.2 Análise da colinearidade e multicolinearidade

Quando se utiliza o MRLM é normal que em algumas circunstâncias se tenha previamente uma noção do comportamento que determinadas variáveis independentes terão sobre a variável dependente. Na prática, o que acontece várias vezes é que se cometem alguns erros na escolha

das variáveis independentes, que resultam em alguns problemas relacionados com a estimação dos coeficientes do modelo e consequentemente com a previsão dos resultados.

Algumas das consequências dos erros cometidos na escolha das variáveis independentes são:

- Pode-se notar que o sinal de um determinado estimador $\beta_j,\ j = 1, \cdots, k$ é diferente do esperado;

- Verificação de grandes alterações em $\beta_j,\ j = 1, \cdots, k$ quando se adiciona ou se exclui variáveis e observações;

- Observação de um aumento insignificante do R^2 quando se adiciona no modelo uma ou mais variáveis independentes, mesmo que as mesmas sejam importantes na explicação da variável dependente.

Nestes casos, têm-se alguns indícios da ausência da ortogonalidade entre as variáveis independentes, ou seja, suspeita-se que o modelo apresenta problemas de colinearidade e/ou multicolinearidade.

A Colinearidade revela a existência de correlação forte entre duas variáveis independentes do modelo, enquanto que a Multicolinearidade revela a existência de correlações fortes entre mais de duas variáveis independentes. A Colinearidade é também tida como a existência de relação linear entre duas variáveis independentes e a Multicolinearidade a existência de relação linear entre uma variável independente e as demais.

A presença de colinearidade e/ou multicolinearidade não altera significativamente a qualidade do ajustamento, já que esta condição não afeta o problema de minimização da soma dos quadrados dos resíduos a não ser no caso extremo de uma ou mais variáveis serem combinações lineares perfeitas das outras variáveis independentes. O que a existência de colinearidade e/ou multicolinearidade nos indica, é que a informação presente nas variáveis correlacionadas ou multicorrelacionadas é redundante, por isso, uma ou algumas delas devem ser eliminadas da análise. Qual ou quais das variáveis devem ser eliminadas? A resposta depende do conhecimento empírico do fenómeno em estudo e da decisão do investigador. Porém, para auxiliar na decisão podem ser usados os métodos de seleção de variáveis independentes para o modelo, abordados no capítulo anterior ou diagnosticar quais são as variáveis que estão correlacionadas e/ou multicorrelacionadas e depois decidir qual ou quais delas devem ser retiradas do modelo.

A colineariadade e/ou multicolinearidade pode ser diagnosticada de várias formas. Uma delas é através da análise da matriz de correlações bivariadas entre variáveis definidas para o modelo. Apesar de não existir um valor de correlação limite a partir do qual seja possível prever problemas na estimação do modelo devido à existência de colinearidade, correlações bivariadas iguais ou superiores a 0.75 entre variáveis independentes conduzem geralmente a este tipo de problemas. Estes coeficientes de correlação são válidos apenas quando as variáveis são analisadas duas a duas. Quando mais do que duas variáveis forem colineares, a matriz de correlações já não pode ser usada, pois nada garante que a associação linear entre mais de duas variáveis seja refletida num dos coeficientes de correlação bivariado. Neste caso, há necessidade de se usarem outras técnicas para diagnosticar a multicolinearidade.

Uma das técnicas que permite diagnosticar a multicolinearidade, é a *Tolerância*, que é dada pela expressão:

$$T = 1 - R_j^2, \tag{4.17}$$

onde R_j^2 representa o coeficiente de determinação entre $x_{ij},\ j = 1, \cdots, k$ (tida como variável dependente) e as restantes variáveis independentes do modelo. Quando o seu valor estiver próximo de 0, a variável x_{ij} pode escrever-se como uma variável quase linear das outras variáveis independentes e consequentemente, o respetivo coeficiente de regressão é instável quer na magnitude quer no sinal.

Outra técnica de diagnóstico da multicolinearidade que não é influenciada pelo problema das correlações bivariadas é o Fator de Inflação da Variância (*Variance Inflaction Factor*), simbolicamente representado por VIF, que é calculado através da expressão:

$$VIF = \frac{1}{1 - R_j^2}, \tag{4.18}$$

De forma geral, os valores de VIF superiores a 5 (ou mesmo a 10) indicam problemas com a estimação do β_j devido à presença de multicolinearidade nas variáveis independentes, ver Marôco (2014).

Outra medida usada são os Valores próprios (*Eigenvalues*) da matriz de correlações entre as variáveis independentes. Se uma ou mais variáveis independentes forem colineares com algumas das restantes, então pelo menos existirá um valor próprio muito próximo de 0.

4.3 Conclusão

Com a abordagem da validação dos pressupostos impostos aos modelos de RLS e RLM, acabamos de apresentar os conteúdos que foram reservados para a parte teórica deste trabalho. A seguir, apresenta-se um estudo onde em forma de aplicação prática, vão ser considerados diversos aspetos apresentados nestes últimos três capítulos, o que de certo modo irá permitir cimentar muitas das questões tratadas nesta parte teórica.

Capítulo 5

Fatores que influenciam o rendimento académico dos alunos da província do Moxico em Angola

A problemática do rendimento académico dos alunos na disciplina de matemática é um dos assuntos que atualmente preocupa professores, alunos e outros agentes de educação em quase todos os países do mundo, incluindo Angola, um dos países onde a situação pode ser um pouco mais delicada do que em alguns, devido às especificidades do seu sistema de ensino.

Um estudo realizado pela Organização para a Cooperação e Desenvolvimento Económico (OCDE)[3] sobre as habilidades dos adultos que habitam em países membros desta organização, mostra que as poucas habilidades a matemática limitam severamente o acesso das pessoas a trabalhos melhor remunerados e mais gratificantes. Este estudo revela também que, a equidade e a inclusão na política pública dependem de certo modo das habilidades dos cidadãos a matemática (OCDE, 2014a, p. 6).

Os resultados deste estudo podem justificar as avultadas somas de dinheiro que os países membros desta organização investem anualmente nas escolas (que chegam aos USD 230 biliões por ano), para financiar o ensino da matemática, cujos ganhos compensam muito mais do que os investimentos (OCDE, 2014a, p. 6), contribuindo deste modo para a afirmação económica e política de muitos destes países no mundo. Os efeitos positivos dos investimentos que se fazem em muitos países no ensino da matemática, estão bem patentes nos resultados dos estudos internacionais que avaliam a proficiência dos alunos a matemática, como o PISA (Programme for International Student Assessment) e TIMSS (Trends in International Mathematics and Science Study), que mostram que alunos de países com níveis económicos elevados ou médios, apresentam em todas as edições destes concursos melhores classificações, se comparados com os alunos de países com baixos níveis económicos.

Em Angola, até ao momento não existem estatísticas oficiais que ilustrem qual é a situação real sobre esta problemática, porém, em cada ano letivo o número de alunos que apresenta resultados negativos nesta disciplina em todos os níveis de ensino é muito alto, se comparado com outros países, como é o caso de Portugal, embora se conclua também que, Portugal vive um problema semelhante, isto é, se se tiver em conta os dados que constam, por exemplo, do relatório dos resultados escolares por disciplina do 3º Ciclo de ensino, referentes ao ano letivo 2014/2015, publicados em novembro de 2017, pela Direção Geral de Estatísticas da Educação e Ciências (DGEEC, 2017).

Angola sempre almejou ter cidadãos com habilidades em matemática, que as permitam ser atores do desenvolvimento económico e tecnológico do país e que se incluam na política pública,

[3]Sigla em inglês: OECD - Organization for Economic Co-operation and Development.

promovendo a equidade e a integridade nas instituições públicas e privadas. Para o efeito, o país tem vindo a trabalhar desde o alcance da sua independência em 1975, no sentido de encontrar soluções para esta problemática, através da implementação de reformas no sistema de ensino, sustentadas pelas leis de bases do sistema educativo (Lei 13/01 de 31 de dezembro de 2001 e Lei 17/16 de 7 de outubro de 2016), criando novas propostas curriculares e promovendo ações de capacitação de docentes (Zau, 2009; MED, 2012; Liberato, 2014). Ainda assim, o problema persiste até ao momento, o que leva as autoridades educativas a questionarem sobre o que é que tem que ser feito, para melhorar o desempenho dos alunos na disciplina de matemática. Entretanto, no nosso entender, a realização de estudos que possibilitem identificar as principais causas que influenciam o rendimento académico à disciplina de matemática pode contribuir significativamente para a sua solução. Visto que, devido aos maus resultados por parte dos alunos nesta disciplina, atualmente existem mais alunos a escusarem de estudar a matemática em Angola do que os que desejam continuar a estudá-la, tal como mostram (André e Larrechea, 2016) no seu estudo sobre o baixo rendimento em matemática em Angola. Um fenómeno que está a contribuir significativamente, para o aumento do número de indivíduos no país sem literacia matemática, ou seja, sem capacidade de formular, aplicar e interpretar a matemática em contextos variados do seu quotidiano (OCDE, 2013a, p. 25).

Para compreender um pouco mais sobre esta problemática em outros países, foram analisados os resultados de alguns estudos realizados por pessoas singulares e instituições internacionais (OCDE, TIMSS e UNESCO) e nacionais de alguns países, com realce para Portugal (Marôco, Gonçalves, Lourenço e Mendes, 2016), Brasil (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2016)), Nigéria (Ogunsola-Bandele, 1996), África do Sul (Howie, 2003), Gana (Butakor, Ampadu e Cole, 2017), entre outros. Esta análise mostrou-nos que o fraco rendimento académico dos alunos do ensino fundamental e secundário à disciplina de matemática é um dos problemas que preocupa quase todas as nações do mundo. Em cada um dos países foram identificados um conjunto de fatores que contribuem para o mau ou bom rendimento académico dos alunos a matemática. O sexo dos alunos, a renda das famílias, a qualidade das escolas, a compreensão pelos alunos da língua em que é ensinada a matemática e o nível de preparação dos professores são, por exemplo, alguns dos fatores apontados pelos estudos, como aqueles que influenciam o desempenho dos alunos à esta disciplina importante do currículo escolar. Com base na análise efetuada e atendendo as particularidades do ensino em Angola, decidiu-se realizar este estudo, com o objetivo de identificar alguns dos fatores que influenciam o rendimento académico dos alunos do ensino secundário da província do Moxico em Angola a matemática.

O conteúdo desta investigação está subdividido em quatro secções, onde se abordam questões relacionadas com a metodologia usada no estudo, apresentação dos resultados, a discussão e as considerações finais do trabalho, que de entre outras conclusões, apresentam alguns dos fatores que influenciam o desempenho dos alunos à disciplina de matemática.

5.1 Metodologia

Esta pesquisa tem um cariz quantitativo, uma vez que, a mesma obedeceu a um processo sistemático de colheita de dados observáveis e quantificáveis (Freixo, 2012) e efetuaram-se medições controladas dos dados fornecidos pela amostra, com recurso a técnicas estatísticas (Khusainova, Shilova e Curteva, 2016).

5.1.1 População e amostra

A população em estudo é composta por um universo de 14007[4] estudantes, dos quais 5088 (36,325%) do sexo feminino, que frequentaram as 10ª, 11ª, 12ª e 13ª classes nas escolas do ensino secundário da província do Moxico, durante o ano letivo 2017 (que terminou em dezembro do mesmo ano).

Para a obtenção da amostra, começou-se por selecionar de forma aleatória três escolas, cada uma pertencente a uma região diferente (urbana, periurbana e suburbana) da cidade do Luena na província do Moxico em Angola. Depois de serem obtidas as escolas, posteriormente foram selecionadas também de forma aleatória, algumas turmas de cada uma destas escolas. Os alunos das turmas selecionadas compõem a amostra do presente estudo. Esta amostra é composta por 350 alunos, dos quais 211 (60,29%) são do sexo masculino e 139 (39,71%) do sexo feminino.

Quanto ao tamanho da amostra, este foi estimado com base nas indicações estabelecidas por Hill e Hill (2016), que aconselham que para uma análise através do modelo de regressão linear múltipla, o tamanho mínimo da amostra nunca deve ser inferior a 30; porém, para aumentar a significância da generalização do modelo esta regra aconselha a utilização de uma amostra de tamanho $n = 15k$ (onde k é o número de variáveis independentes). Caso se queira efetuar uma regressão $Stepwise$ (que é o caso da análise efetuada neste estudo), o tamanho da amostra deve ser ainda maior. Neste caso é aconselhável que no mínimo o tamanho seja $n = 30k$. Esta condição é cumprida no presente estudo, visto que, temos $n = 350$ e $k = 10$.

Por outro lado, considerando esta dimensão amostral e fixando o nível de significância em 5%, pode-se afirmar que se cometeu um erro de estimativa pouco superior a 5% (aproximadamente 5,3%), ver fórmulas para determinação do tamanho da amostra em (Levine, Berenson e Stephan, 2000).

5.1.2 Instrumento utilizado na recolha de dados

O instrumento utilizado para a recolha de dados foi o questionário apresentado no Anexo 2 (Figura A.1). O mesmo teve como unidades de análise, elementos relacionados com a vida académica, pessoal e socioeconómica dos alunos, como a idade, sexo, estado civil, as notas finais que os alunos obtiveram no ano letivo anterior às disciplinas de Matemática, Língua Portuguesa

[4]Fonte: Direção Provincial de Educação Ciência e Tecnologia do Moxico-Angola

e Física, o grau de satisfação dos alunos com o ambiente escolar, a renda familiar, situação laboral e outros aspetos não menos importantes.

A variável grau de satisfação com o ambiente escolar foi recodificada numa escala de Likert com 5 categorias, nomeadamente: $[0, 2]$ - Muito insatisfeito, $[3, 4]$ - Insatisfeito, $[5, 6]$ - Nem insatisfeito nem satisfeito, $[7, 8]$ - Satisfeito e $[9, 10]$ - Muito satisfeito.

Por outro lado, criou-se com base nos dados da variável quantitativa idade, a variável qualitativa ordinal "Faixa etária". Esta possui três categorias, nomeadamente, $[\leq 19]$, que congrega os alunos de até 19 anos, considerada como idade apropriada para se frequentar o ensino secundário; $[20 - 24]$, que congrega alunos com idades compreendidas entre 20 a 24 anos e $[\geq 25]$ que congrega alunos com 25 anos ou mais, aqueles que estão muito atrasados na frequência do ensino secundário.

5.1.3 Metodologias estatísticas

Para dar resposta às questões sobre o rendimento escolar à disciplina de matemática, recorreu-se a algumas metodologias estatísticas (Figura 5.1).

A análise exploratória dos dados foi feita com recurso a técnicas da estatística descritiva, nomeadamente foram consideradas as frequências absolutas e relativas em cada categoria das variáveis qualitativas e algumas medidas de tendência central e dispersão para o caso das variáveis quantitativas. Foram ainda traçados gráficos que ajudaram a caracterizar a amostra e foram também usados alguns métodos da estatística inferencial.

A existência de relação entre algumas das variáveis qualitativas foi verificada através do teste do Qui-quadrado de *Pearson*. Quando os pressupostos para a utilização do mesmo não foram verificados, recorreu-se ao teste Exato de *Fisher*. Recorreu-se também ao coeficiente de associação V de *Cramer*, por forma a quantificar o grau da associação entre as variáveis qualitativas. Para classificar o grau de associação, considerou-se o critério estabelecido por (Cohen, 1988), segundo o qual:

- $V < 0,3$, associação fraca;
- $0,3 \leq V < 0,5$, associação moderada;
- $V \geq 0,5$, associação forte.

Por fim, e atendendo ao objetivo principal deste estudo, ajustou-se um Modelo de Regressão Linear Múltipla aos dados, por forma a analisar quais são os fatores que influenciam as notas dos alunos à disciplina de matemática. A verificação dos pressupostos impostos ao erro do modelo (normalidade, homoscedasticidade e independência) foi devidamente analisada, assim como as situações de colinearidade e/ou multicolinearidade entre as variáveis independentes e a existência de *outliers* e observações influentes.

Para efetuar as análises, recorreu-se ao *software SPSS Statistics*, versão 24 e foi considerado um grau de significância de 5%.

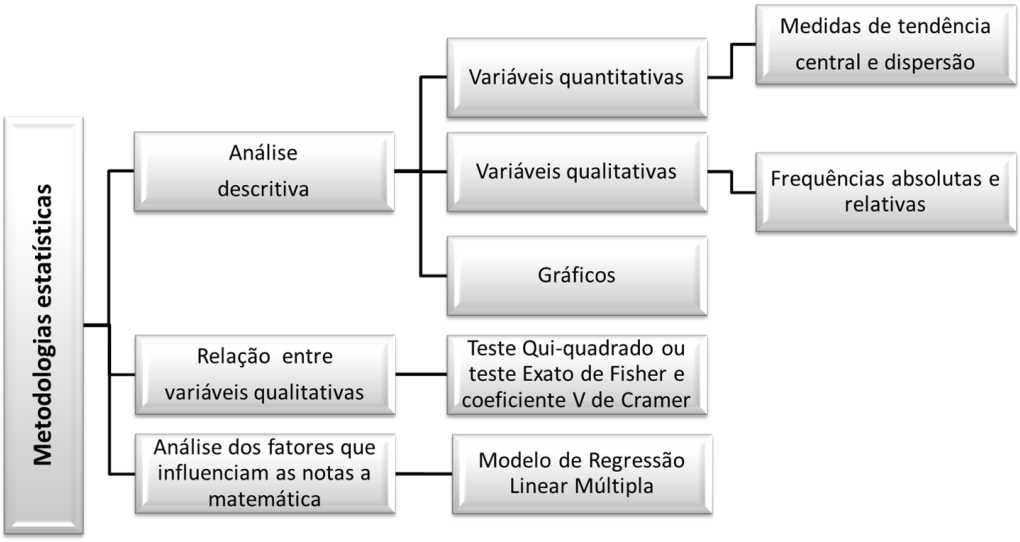

Figura 5.1: Metodologias estatísticas usadas no estudo

5.2 Resultados

Nesta secção iremos apresentar os principais resultados obtidos através da análise efetuada.

5.2.1 Análise descritiva das variáveis em estudo

As informações obtidas através do questionário aplicado, possibilitaram considerar para as análises deste estudo 12 variáveis, sendo 4 quantitativas e 8 qualitativas. A seguir apresentam-se as tabelas das estatísticas descritivas referentes a estas variáveis.

Tabela 5.1: Variáveis quantitativas

Variável	N	Mínimo	Máximo	Média	Desvio Padrão
Idade	350	16	32	21,20	3,84
Notas a Matemática	350	9,00	17,00	12,52	1,77
Notas a Língua Portuguesa	350	8,00	19,00	12,93	2,15
Notas a Física	350	7,00	18,00	12,60	2,14

Os resultados do inquérito efetuado aos 350 alunos mostram que, estes apresentam uma média de idades de aproximadamente 21 anos ($21{,}20 \pm 3{,}84$), a idade mínima é de 16 anos e a máxima 32 anos. A média das notas obtidas pelos alunos à disciplina de Matemática é de aproximadamente 13 valores ($12{,}52 \pm 1{,}77$), enquanto que a nota mínima nesta disciplina é de 9 e a máxima de 17 valores. Na disciplina de Língua Portuguesa, a média é também de aproximadamente 13

valores (12,93$\pm$2,14), enquanto que a nota mínima é de 8 valores e a máxima de 19 valores e na disciplina de Física a média é também de aproximadamente 13 valores (12,60$\pm$2,15), a nota mínima é de 7 e a máxima de 18 valores.

Tabela 5.2: Variáveis qualitativas em estudo

Variável	Categorias	Frequências	Percentagem
Sexo	Feminino	139	39,71 %
	Masculino	211	60,29 %
Estado civil	Casado/União de facto	68	19,43 %
	Solteiro	282	80,57 %
Renda familiar	Baixa	114	32,57 %
	Média	236	67,43 %
Situação laboral do aluno	Não trabalhador estudante	319	91,14 %
	Trabalhador estudante	31	8,86 %
Classificação atribuída	Mau	29	8,29 %
ao professor de matemática	Bom	321	91,71 %
Grau de satisfação c/ ambiente escolar	Muito insatisfeito	8	2,29 %
	Insatisfeito	15	4,28 %
	Nem insatisfeito nem satisfeito	63	18,00 %
	Satisfeito	68	19,43 %
	Muito satisfeito	196	56,00 %
Escola do aluno e localização	11 de Novembro-Periurbana	170	48,60 %
	338 Tchifuchi - Urbana	140	40,00 %
	4 de Abril - Suburbana	40	11,40 %
Faixa etária	≤ 19	143	40,86 %
	$20 - 24$	142	40,57 %
	≥ 25	65	18,57 %

Através da análise desta tabela conclui-se que, na sua maioria os alunos que compõem a amostra são do sexo masculino (60,29%). No que concerne ao estado civil a maioria dos alunos, 282 (80,57%), são solteiros e apenas 68 (19,43%) são casados ou vivem em união de facto. Sobre a renda familiar, 236 (67,43%) alunos têm uma renda familiar média, 114 (32,57%) uma renda familiar baixa e não existiam alunos com renda familiar alta. Quanto à situação laboral, apenas 31 (8,86%) alunos têm um emprego. Já em termos da classificação atribuída pelos alunos aos professores de matemática, 321 (91,71%) consideram os seus professores de matemática como bons e apenas 29 (8,29%) os consideram como maus.

Quanto ao grau de satisfação com o ambiente escolar, apenas 8 (2,29%) alunos se consideraram muito insatisfeitos, enquanto que a maioria, 196 (56%), se considerou muito satisfeita. Os dados informam ainda que 170 (48,60%) alunos estudavam numa escola localizada na zona periurbana, 140 (40%) na zona urbana e apenas 40 (11,40%) na zona suburbana. No que concerne à faixa etária, 143 (40,86 %) alunos apresentavam idades iguais ou inferiores a 19 anos, 142 (40,57 %), tinham entre 20 a 24 anos e 65 (18,57%) tinham idades iguais ou superiores a 25 anos.

5.2.2 Verificação da existência de relação entre algumas das variáveis em estudo

Tabela 5.3: Relação entre o grau de satisfação, com a escola e com o sexo do aluno

			Grau de satisfação com o ambiente escolar					p-value	V de Cramer
			Muito insatisfeito	Insatisfeito	Nem Insatisfeito nem Satisfeito	Satisfeito	Muito satisfeito		
Escola do aluno e localização	11 de Novembro - Periurbana	Contagem	3	9	39	35	84	0,003* #1	0,181
		% em Grau ...	37,5%	60,0%	61,9%	51,5%	42,9%		
	338 Tchifuchi - Urbana	Contagem	2	3	17	22	96		
		% em Grau ...	25,0%	20,0%	27,0%	32,4%	49,0%		
	4 de Abril - Suburbana	Contagem	3	3	7	11	16		
		% em Grau ...	37,5%	20,0%	11,1%	16,2%	8,2%		
Sexo do aluno	Feminino	Contagem	0	11	32	25	71	0,002* #2	0,219
		% em Grau ...	0,0%	73,3%	50,8%	36,8%	36,2%		
	Masculino	Contagem	8	4	31	43	125		
		% em Grau ...	100,0%	26,7%	49,2%	63,2%	63,8%		

#1 Teste Exato de Fisher; #2 Teste Qui-quadrado ; * $p < 0.05$

Com base nos resultados da Tabela 5.3, conclui-se que existe uma relação significativa entre a escola e o grau de satisfação com o ambiente escolar ($p - value = 0,003$), embora o grau de associação seja fraco ($V = 0,181$). Existe também uma relação significativa entre o sexo e o grau de satisfação com o ambiente escolar ($p - value = 0,002$), com associação também fraca ($V = 0,219$). Os resultados indicam que, em todas as escolas e para ambos os sexos a maioria dos alunos está satisfeita ou muito satisfeita. O número de alunos muito insatisfeitos e insatisfeitos é muito pouco representativo (ver também, Figuras A.2 e A.3 do Anexo 2).

Considerando as escolas, nota-se que entre os alunos muito satisfeitos é na escola 338 Tchifuchi que se tem maior percentagem. Já considerando os alunos muito insatisfeitos e insatisfeitos estes estão mais presentes na escola 11 de Novembro. Por outro lado, quanto ao sexo, os alunos satisfeitos e muito satisfeitos são em maior número os do sexo masculino.

Tabela 5.4: Relação entre a escola, com a classificação atribuída pelo aluno ao professor de matemática e com a faixa etária

			Escola do aluno e localização			(p-value)	V de Cramer
			11 de Novembro - Periurbana	338 Tchifuchi - Urbana	4 de Abril - Suburbana		
Classificação atribuída pelo aluno ao professor de matemática	Má	Contagem	22	5	2	0,009 * #2	0,165
		% em Escola do aluno	12,9%	3,6%	5,0%		
	Boa	Contagem	148	135	38		
		% em Escola do aluno	87,1%	96,4%	95,0%		
Faixa etária	≤19	Contagem	70	49	24	0,017 * #2	0,131
		% em Escola do aluno	41,2%	35,0%	60,0%		
	20 - 24	Contagem	62	67	13		
		% em Escola do aluno	36,5%	47,9%	32,5%		
	≥25	Contagem	38	24	3		
		% em Escola do aluno	22,4%	17,1%	7,5%		

#2 Teste Qui-quadrado ; * $p < 0.05$

Os resultados da Tabela 5.4 mostram que existe uma relação significativa entre a escola e a classificação atribuída pelos alunos aos seus professores de matemática ($p - value = 0,009$), entretanto, esta associação é classificada como fraca ($V = 0,165$). Por outro lado, conclui-se também que, existe uma relação significativa entre a escola e a faixa etária dos alunos ($p - value = 0,017$), sendo igualmente fraco o grau de associação ($V = 0,131$). Da análise da Tabela 5.4, conclui-se ainda que, em geral, os alunos classificam os seus professores de matemática como bons. A percentagem mais elevada destes alunos é verificada na escola 338 Tchifuchi. Conclui-se também que, a maioria dos alunos das escolas 4 de Abril e 11 de Novembro possuem idades menores ou iguais a 19 anos, enquanto que na escola 338 Tchifuchi a faixa etária predominante é a de alunos com idades compreendidas entre os 20 e 24 anos.

5.2.3 Análise dos fatores que influenciam as notas à disciplina de Matemática

Para esta análise, vai ser ajustado um modelo de regressão linear múltipla, considerando como variável dependente as notas à disciplina de matemática e como variáveis independentes as outras três variáveis quantitativas (notas a língua portuguesa, notas a física e idade) e um conjunto de 11 variáveis *dummies* obtidas a partir das categorias de 7 das 8 variáveis qualitativas[5] que constam da Tabela 5.2.

5.2.3.1 Modelo ajustado

Começa-se a análise introduzindo as 14 variáveis no modelo com o método de inserção de variáveis $Enter$. O modelo obtido é estatisticamente significativo, uma vez que relativamente à ANOVA se obteve um $p - value < 0,001$, porém, analisando os resultados da Tabela A.4 do Anexo 2, conclui-se que 10 das 11 variáveis *dummies* não apresentam significância estatística ($p - value > 0,05$). Prosseguindo a análise para a obtenção do melhor modelo que se ajusta aos dados, recorreu-se aos métodos de seleção de variáveis, começando pelo $Backward$, depois pelo $Forward$ e por último o $Stepwise$. Com o primeiro ($Backward$) obtiveram-se 8 submodelos. Já com os outros dois métodos obtiveram-se 5 submodelos em cada um, em que o modelo final sugerido por ambos os métodos é igual e é o que será considerado. O referido modelo é estatisticamente significativo, visto que, quanto à ANOVA, $p - value < 0,001$ (Tabela 5.5).

Neste modelo foram consideradas como significativas na predição das notas a matemática, as variáveis notas a língua portuguesa [NLP] ($p - value < 0,001$), notas a física [NF] ($p - value < 0.001$), idade [ID] ($p - value < 0,001$), renda familiar (média) [RF_M] ($p - value = 0,001$) e, no que diz respeito ao grau de satisfação com o ambiente escolar, a classe muito insatisfeito [MI] ($p - value = 0,046$). As demais variáveis foram consideradas não significativas ($p - values > 0,05$), tendo sido excluídas do modelo ajustado (ver Anexo 2, Tabela A.5). O modelo ajustado obtido possui um coeficiente de determinação ajustado de $R_a^2 = 0,866$. Com base neste conclui-se que 86,6% da variabilidade da nota a matemática é explicada pelas variáveis independentes

[5] Obs.: Não se considerou na análise de regressão a variável faixa etária, pelo facto de se ter considerado a variável idade.

selecionadas para o modelo e apenas os restantes 13,4% da variabilidade serão explicados por fatores alheios ao modelo obtido.

Tabela 5.5: Estimativas dos coeficientes de regressão do modelo ajustado

			IC a 95% para β		Estatística de colinearidade	
	Estimativas	$p-value$	Limite inferior	Limite superior	Tolerância	VIF
Constante	2,572	0,000	1,730	3,414		
Notas a Física	0,451	0,000	0,413	0,489	0,704	1,420
Notas a Língua Portuguesa	0,381	0,000	0,344	0,419	0,734	1,363
Idade	-0,041	0,000	-0,061	-0,021	0,763	1,311
Renda familiar (Média)	0,276	0,001	0,112	0,440	0,788	1,269
Grau de satisf. c/ o ambiente escolar (Muito insatisfeito)	0,465	0,046	0,007	0,924	0,991	1,009
ANOVA $(p-value < 0,001)$ R_a^2 = 0,866						

Classes de referência: Renda familiar - Baixa; Grau de satisfação com o ambiente escolar - Nem insatisfeito nem satisfeito.

Com estes resultados e recorrendo aos valores obtidos das estimativas dos coeficientes, formulou-se o seguinte **modelo de regressão linear múltipla** ajustado:

$$NMAT = 2,572 + 0,451 \times NF + 0,381 \times NLP - 0,041 \times ID + 0,276 \times RF_M + 0,465 \times MI \quad (5.1)$$

Deste modelo derivam os submodelos apresentados na tabela seguinte.

Tabela 5.6: Submodelos obtidos

Renda Familiar	Grau de satisfação	
Baixa	NINS	$^{(1)}NMAT = 2,572 + 0,451 \times NF + 0,381 \times NLP - 0,041 \times ID$
	MI	$^{(2)}NMAT = 3,023 + 0,451 \times NF + 0,381 \times NLP - 0,041 \times ID$
Média	NINS	$^{(3)}NMAT = 2,834 + 0,451 \times NF + 0,381 \times NLP - 0,041 \times ID$
	MI	$^{(4)}NMAT = 3,299 + 0,451 \times NF + 0,381 \times NLP - 0,041 \times ID$

NINS - Nem insatisfeito nem satisfeito; MI - Muito insatisfeito;
$^{(1)}\ldots{}^{(4)}$ - Numeração dos submodelos.

A equação do modelo (5.1) diz-nos que, cada valor a mais que um estudante obtiver à disciplina de Física aumenta a sua nota de Matemática em 0,451 valores e cada valor a mais que um estudante obtiver à disciplina de Língua Portuguesa a sua nota a Matemática aumenta 0,381 valores, mantendo tudo o resto igual. Quanto à idade, cada ano que se acrescenta na idade do aluno a sua nota a Matemática tende a diminuir 0,041 valores, mantendo tudo o resto igual. Por outro lado, se o aluno pertence a uma família com uma renda familiar média, a sua nota a Matemática tende a aumentar 0,276 valores, em relação a um aluno com uma renda familiar baixa. Se o aluno for alguém muito insatisfeito com o ambiente escolar a sua nota a Matemática aumenta 0,465 valores, em relação a um aluno nem insatisfeito nem satisfeito.

5.2.3.2 Verificação dos pressupostos impostos ao erro do modelo

• Análise da normalidade

Da análise feita aos resíduos, obteve-se um resultado do teste de Kolmogorov-Smirnov que possibilitou concluir que os mesmos possuem uma distribuição normal ($p - value = 0,906$).

Tabela 5.7: Resultados dos testes para a verificação da normalidade e independência dos resíduos

Kolmogorov-Smirnov		Durbin-Watson
Estatística	$p - value$	Estatística (d)
0.030	0.906	1.991

• Análise da independência

O valor obtido da estatística de Durbin-Watson corresponde a $d = 1,991$. Como se pode notar este valor é muito próximo de 2, o que faz com que não se rejeite a hipótese nula e concluir que, os resíduos são independentes (Marôco, 2014). Este resultado confirma-se usando os limites definidos por Durbin-Watson (Tabela A.1, Anexo 1), tal como se mostra a seguir.

Dado que a amostra tem a dimensão $n = 350$ e o número de variáveis independentes do modelo ajustado é $k = 5$, os valores críticos do teste de Durbin-Watson para este caso são: limite inferior $d_L = 1,80$ e limite superior $d_U = 1,85$. Deste modo se tem:

$$d_U = 1,85 < d = 1,991 < 4 - d_U = 2,15.$$

Assim, não se rejeita a hipótese nula e conclui-se que os resíduos são independentes.

• Análise da homoscedasticidade

Analisando a Figura 5.2 pode-se notar que os resíduos se distribuem aleatoriamente em torno da reta horizontal, por isso, conclui-se que os mesmos são homoscedásticos.

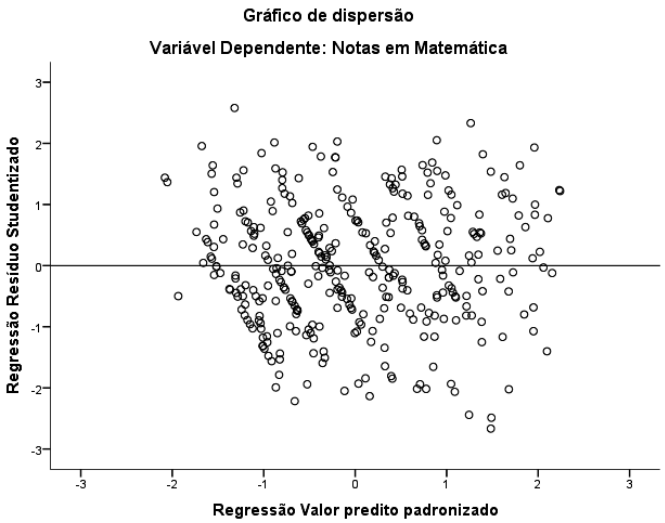

Figura 5.2: Gráfico dos resíduos estudantizados versus valores preditos ajustados

5.2.3.3 Análise da colinearidade e/ou multicolinearidade

Dos resultados obtidos do diagnóstico da colinearidade e/ou multicolinearidade (ver Tabela 5.5), notou-se que todas as variáveis independentes possuem valores do Fator de Inflação da Variância (VIF) inferiores a 5 e o valor da *Tolerância* próximo de um, o que permite concluir que, não existem problemas de colinearidade e/ou multicolinearidade entre as variáveis independentes inseridas no modelo.

5.2.3.4 Diagnóstico de *outliers* e observações influentes

• Diagnóstico de *Outliers*

Para verificar a existência de *outliers*, usou-se a análise gráfica, por meio do gráfico de resíduos estudantizados excluídos versus valores preditos estandardizados (ver Anexo 2, Figura A.4), onde se pode verificar que não existem observações visivelmente distanciadas das demais. Ainda assim, pela análise da Figura A.4 nota-se que algumas observações como, por exemplo, das observações 37, 61, 183 e 197 possuem resíduos em valores absolutos superiores a 1.96, confirmando assim a existência de alguns *outliers*. Para medir a sua severidade, recorreu-se ao valor da *Leverage*, cujo valor máximo obtido na análise é inferior a 0,2, (ver Anexo 2, Figura A.5) que é considerado um limite aceitável para que os *outliers* não sejam considerados severos, ver Marôco (2014).

• Diagnóstico de observações influentes

Para saber se existem observações que exercem influência na estimação dos parâmetros do modelo usaram-se as estatísticas dos $DFFIT\ Estandardizados$. Os resultados obtidos mostram que não existe nenhuma observação cujo $DFFIT\ Estandardizado$ é superior, em módulo, a 1,96, como se pode notar através da Figura A.6 do Anexo 2. Deste modo, conclui-se que não existem observações que influenciam a estimação dos parâmetros do modelo.

Observações:

Após a verificação dos pressupostos (análise de resíduos, colinearidade e/ou multicolinearidade e a existência de *outliers* e observações influentes), uma vez que nenhum deles foi violado, conclui-se que o Modelo de Regressão Linear Múltipla obtido pode ser usado com o objetivo de explicar as notas dos alunos à disciplina de Matemática com base nas variáveis independentes tidas como significativas. Além disso, foi obtido um valor de $R_a^2 = 0,866$, que nos dá a indicação de uma boa qualidade de generalização deste modelo.

5.3 Discussão

Depois de termos apresentado os resultados que, na nossa opinião, proporcionam um conjunto de informações importantes sobre aspetos sociodemográficos, socioeconómicos e académicos dos alunos do ensino secundário da província do Moxico em Angola, de seguida vamos sintetizar e ilustrar a relevância dos principais resultados obtidos neste estudo e efetuar uma análise crítica dos mesmos.

Entre os resultados, o estudo revela que nas escolas do ensino secundário da província do Moxico em Angola, a maioria dos alunos está satisfeita com o ambiente que as escolas proporcionam-lhes. No nosso entender, este resultado justifica-se com a oferta de novas infraestruturas escolares e melhoria das já existentes, que de algum modo está a contribuir para a melhoria do ensino na província.

Examinando os resultados da relação entre a localização da escola e a classificação atribuída pelos alunos aos professores de matemática nota-se que é na escola periurbana, em que existe uma maior percentagem de alunos que atribuem má classificação aos professores de matemática. Por outro lado, nota-se que cerca de 90% dos alunos que foram inquiridos consideram os seus professores de matemática como bons, apesar de alguns não possuírem um bom desempenho a esta disciplina. Este resultado vem mostrar que, nem sempre o desempenho dos alunos a matemática depende da qualidade dos seus professores.

Quanto à faixa etária dos alunos, notou-se que a maior percentagem de alunos com idades compreendidas entre os 20 a 24 frequentava as escolas do meio urbano, enquanto que as escolas situadas no meio suburbano e periurbano congregavam uma proporção maior de alunos de até 19 anos. Por outro lado, conclui-se ainda que na amostra, mais da metade (59,14%) dos alunos possui idades superiores a 19 anos, ou seja, estão acima da idade recomendada para a frequência do ensino secundário.

O Modelo de Regressão Linear Múltipla obtido prevê explicar 86,6% da variabilidade da nota à disciplina de matemática. Com base neste modelo, conclui-se que as notas dos alunos às disciplinas de língua portuguesa e de física, a idade, a renda familiar média e a insatisfação elevada com o ambiente escolar influenciam significativamente as notas dos alunos à disciplina de matemática. Entre as variáveis explicativas do modelo, no caso específico das notas a língua portuguesa e a física, o estudo revela que quando as mesmas são maiores, a nota do aluno a matemática tende a ser mais elevada. Este resultado vem despertar as entidades educativas, no sentido de, quando estiverem a traçar estratégias que visam melhorar o desempenho dos alunos a matemática no ensino secundário, não pensarem simplesmente nas metodologias e conteúdos desta disciplina, e na qualificação profissional dos docentes, mas também, na melhoria do nível de entendimento dos alunos da língua oficial utilizada nas escolas, através da qual os mesmos poderão aprender a interpretar e resolver problemas matemáticos, assim como, na melhoria dos conhecimentos das ciências experimentais como é o caso da física, onde se demonstra a utilidade prática de vários conteúdos matemáticos na resolução e explicação de vários problemas do quotidiano do aluno. No caso da idade, nota-se que por norma os alunos mais novos possuem melhores notas a matemática em relação aos mais velhos. Por outro lado, o estudo mostra que alunos oriundos de famílias com uma renda familiar média, têm maiores possibilidades de te-

rem melhores notas a matemática, em relação aos alunos, cujas famílias têm uma renda baixa. Já os alunos muito insatisfeitos com o ambiente escolar apresentam uma tendência de terem melhores notas a matemática em relação aos que não estão nem insatisfeitos nem satisfeitos.

Da análise efetuada aos resultados de alguns estudos (a que tivemos acesso), realizados em países de África, América e Europa, quanto à influência das notas a física nas notas à disciplina de matemática, não se encontraram resultados semelhantes ao obtido neste estudo nem os que o contrariem. Porém, o nigeriano (Ogunsola-Bandele, 1996) realizou um estudo que teve como população, alunos do ensino médio do Norte da Nigéria, onde comprovou a influência da matemática no ensino da física. Um estudo semelhante realizado em Portugal com estudantes do 9º ano de escolaridade provou também a influência da matemática no desempenho dos alunos à disciplina de física (Fernandes, 2007).

Já sobre a influência da língua portuguesa no desempenho a matemática, Costa (2007) num estudo realizado em Portugal sobre "A importância da língua portuguesa na aprendizagem da matemática" provou a existência da influência da língua portuguesa no desempenho a matemática, nos alunos do ensino básico. Num estudo sobre a língua e outros fatores de fundo que afetam o desempenho dos alunos do ensino secundário em Matemática na África do Sul, (Howie, 2003) concluiu que o desempenho dos alunos na língua inglesa era um forte preditor do seu sucesso em matemática. E o estudo efetuado por Essien (2018) onde apresenta uma revisão das pesquisas realizadas entre 2006 e 2016 sobre o papel da língua no ensino e aprendizagem da matemática em três países africanos considerados multilíngues à exemplo de Angola, nomeadamente, o Quénia, Malawi e África do Sul, revelou também a existência da influência da língua no desempenho a matemática dos alunos destes três países. Estes resultados lançam um alerta às autoridades educativas sobre a atenção que se deve dar à língua oficial utilizada nas escolas, dada a sua importância e influência na aprendizagem das ciências, com realce para a matemática.

No caso da idade, o conflito armado que assolou Angola durante décadas, afetou bastante o sistema de ensino, o que fez com que em várias regiões do país, muitos indivíduos não pudessem terminar o ensino secundário no tempo devido (Zau, 2009). A paz alcançada há dezasseis anos no país, permitiu a extensão do ensino a várias regiões e muitos indivíduos com idades acima da recomendada para a frequência do ensino secundário voltaram às escolas (MED, 2012). O presente estudo constatou com base na amostra recolhida, que estes indivíduos correspondem a perto de metade dos alunos que frequentam o ensino secundário atualmente. Sobre o desempenho a matemática destes alunos, o estudo conclui que por norma os alunos mais novos possuem melhores notas a matemática em relação aos mais velhos. Este resultado sugere às autoridades a necessidade de criarem-se condições, para que daqui em diante os alunos possam frequentar o ensino secundário na idade recomendada (entre 15-17 anos, segundo a Lei 17/16 de 7 de outubro de 2016, que aprova as bases para o sistema de educação em Angola), isto é, caso se queira ter nas escolas maior número de indivíduos com fortes probabilidades de ter um bom desempenho académico.

Sobre a renda familiar, quando se trata de uma renda média, a sua influência no desempenho académico dos alunos foi identificada em estudos anteriores realizados no país pelo Instituto Nacional de Estatística de Angola (INE). Estes estudos apontaram diferenças significativas em crianças pobres e não pobres tanto no acesso à escola, assim como no rendimento académico,

em geral (INE, 2013). Fora do país, a OCDE provou no PISA 2012 que, a alta renda das famílias tinha uma influência positiva no desempenho dos alunos a matemática (OCDE, 2013b). Resultados semelhantes foram verificados em outros estudos nacionais e internacionais, como o PISA 2015 e TIMSS 2011 em países como, Portugal (Marôco, Gonçalves, Lourenço e Mendes, 2016), Gana, Botsuana e África do Sul (Mullis, Martin, Foy e Arora, 2012) e no Brasil (INEP, 2016). Estes resultados sugerem às autoridades angolanas a necessidade de se adotarem políticas que visam aumentar a renda das famílias, para que no futuro se tenha a possibilidade de ter mais alunos com bom desempenho nas escolas, apesar de se estar consciente que, a desigualdade da renda no mundo inteiro tende a se reproduzir de geração a geração, como o afirma (Corak, 2013) num estudo que teve como foco avaliar as desigualdades socioeconómicas dos indivíduos nos países de alta renda.

Quanto aos alunos muito insatisfeitos com o ambiente escolar que apresentam uma tendência de terem melhores notas a matemática em relação aos que não estão nem insatisfeitos nem satisfeitos, este resultado sugere mais pesquisas para se apurar o que está na base do mesmo. O certo é que, segundo a OCDE os resultados do PISA 2015 que contou com a participação de 540 mil estudantes de 72 países e economias mostraram que existia apenas uma relação fraca entre a satisfação do aluno e o desempenho na escola, facto que levou esta organização a concluir que a excelência académica nem sempre resulta da satisfação na vida dos alunos (OCDE, 2017, p. 79).

Entre os preditores tidos como não significativos na explicação da nota dos alunos a matemática, quanto ao sexo, o resultado é um pouco controverso, pois que, até então em Angola, em geral e na província do Moxico em particular, tinha-se a perceção de que os alunos do sexo masculino tinham um melhor desempenho a matemática em relação aos alunos do sexo feminino. Porém, este estudo não mostra exatamente isso, uma vez que a nota a matemática não varia mediante o sexo. Estudos realizados nos países africanos como a África do Sul, Botsuana e Gana mostraram diferenças significativas entre alunos dos dois sexos no desempenho a matemática, a favor dos alunos do sexo masculino (Butakor, Ampadu e Cole, 2017). Entretanto, segundo a OCDE e a UNESCO, em alguns países (incluindo os da África subsariana região em que se situa Angola) onde no passado se registavam diferenças acentuadas entre alunos do sexo feminino e masculino no desempenho a matemática, na sua maioria têm vindo a mostrar que as raparigas estão a superar as lacunas a esta disciplina, e consequentemente o seu desempenho aproxima-se ao dos rapazes (OCDE, 2014b; UNESCO, 2015). Contudo, a falta de estudos do género no país, não nos permite afirmar se Angola é ou não um destes países, onde as raparigas estão a superar as lacunas a matemática.

Quanto à escola e a sua localização, a sua exclusão do modelo indica que não se verificaram diferenças significativas entre as notas a matemática dos alunos das escolas do ensino secundário da província do Moxico em Angola localizadas em diferentes zonas (urbana, suburbana e periurbana). Este resultado contrasta com o que se verifica, por exemplo, em Portugal, onde foram verificadas diferenças significativas no desempenho dos alunos a matemática nas escolas portuguesas situadas em zonas urbanas e suburbanas, com os alunos das zonas urbanas a apresentarem melhores resultados em relação aos das zonas suburbanas (Portela e Faria, 2015). No Brasil, os resultados do estudo sobre "Os determinantes do desempenho escolar" revelaram a existência de diferenças significativas no desempenho a matemática dos alunos do ensino médio das escolas públicas das redes municipais e federais (Menezes-Filho, 2006), estas diferenças

foram verificadas também nos resultados do PISA 2012 e 2015 publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2014, 2016). Deste modo, o resultado obtido com este estudo, vem dizer à comunidade educativa angolana que, na província do Moxico em Angola, a nota a matemática não é influenciada pela zona onde está localizada a escola secundária. Uma realidade que, com base nos resultados dos estudos já citados, é distinta, por exemplo, nos distritos de Portugal e regiões do Brasil.

Dos estudos a que se teve acesso, não se encontrou nenhum que tenha considerado o estado civil, a situação laboral e a classificação atribuída pelos alunos aos professores de matemática, com o objetivo de explicar as notas a matemática. Atendo à realidade angolana decidiu-se incluir estas variáveis na análise e os resultados mostraram que as mesmas não influenciam significativamente as notas a matemática dos alunos do ensino secundário da província do Moxico em Angola.

5.4 Conclusão

Este estudo fornece um conjunto de resultados que doravante, na nossa opinião, vão poder ajudar a caracterizar em termos sociodemográfico, socioeconómico e académico, os alunos do ensino secundário da província do Moxico em Angola. Por outro lado, identifica um conjunto de fatores que influenciam as notas dos alunos à disciplina de matemática e apresenta um modelo estatístico que poderá auxiliar as autoridades educativas e outras entidades a explicarem o desempenho dos alunos a matemática. Estes resultados levam-nos a acreditar que o mesmo cumpre com os objetivos pelos quais foi realizado.

Com base nos resultados apresentados e discutidos nas secções anteriores, conclui-se o seguinte, relativamente aos alunos do ensino secundário da província do Moxico em Angola:

- O sexo dos alunos e as zonas onde se situam as escolas que frequentam, não influenciam significativamente as notas a matemática dos alunos.

- As notas a língua portuguesa e a física, a idade, a renda familiar e a insatisfação elevada com o ambiente escolar, são alguns dos fatores que influenciam significativamente as notas a matemática dos alunos.

- Baseando-se na estrutura do modelo de regressão linear múltipla obtido, conclui-se que, quando se regista um aumento nas notas do aluno a língua portuguesa e a física, a sua nota a matemática também aumenta, enquanto que, quando a idade do aluno aumenta a nota a matemática tende a diminuir.

- Os alunos cujas famílias possuem uma renda média, tendem a ter melhores notas a matemática em relação aos que pertencem à famílias com uma renda baixa e os que apresentam um grau de insatisfação elevado com o ambiente escolar tendem a ter melhores notas a matemática em relação aos que estejam nem insatisfeito e nem satisfeito com o ambiente escolar.

Por outro lado, caso se queira que num futuro próximo se possa melhorar o desempenho dos alunos do ensino secundário a matemática e por consequência disso os indivíduos ganhem capacidade de formular, aplicar e interpretar a matemática em contextos variados que implicam raciocinar matematicamente e usar conceitos, processos, factos e ferramentas matemáticas para descrever, explicar e prever fenómenos que ocorrem à sua volta, na nossa opinião, deve-se ter em conta o seguinte:

1. No que concerne à influência do desempenho da língua portuguesa no desempenho a matemática, é necessário que se criem mecanismos que visem melhorar o ensino da língua portuguesa no ensino primário, visto que, muitos estudantes chegam ao ensino secundário com poucas habilidades na leitura e expressão da língua portuguesa. Este facto dificulta de algum modo, o ensino e aprendizagem da matemática neste nível de ensino, já que em muitas ocasiões é necessário recorrer às habilidades linguísticas, para a compreensão dos conteúdos e resolução de problemas matemáticos.

2. Melhorar o ensino das ciências experimentais, com realce para a Física, que apesar de ser ministrada apenas a partir do 7º ano de escolaridade, permite em muitas situações mostrar a utilidade prática dos conteúdos aprendidos em matemática desde o ensino primário. Para isso, é preciso que existam nas escolas professores capazes de interligarem os conteúdos das duas disciplinas (matemática e física).

3. As autoridades governamentais devem trabalhar para que as famílias possam ter uma renda familiar mais elevada, visto que, esta exerce certa influência no desempenho escolar dos alunos.

4. Devem ser criadas condições no país, para que os indivíduos possam frequentar o ensino secundário com as idades consideradas normais. Já que, os alunos mais velhos tendem a ter menor rendimento escolar do que os mais novos.

Durante a realização deste estudo, tivemos alguns constrangimentos que de certo modo impuseram algumas limitações no mesmo. Uma destas limitações está relacionada com a obtenção da amostra, apesar de, na nossa opinião, termos uma amostra representativa da população, o nosso desejo inicial era de alargar a recolha de dados a escolas situadas em outras localidades, porém, não foi possível, devido às largas distâncias existentes entre as localidades onde estão situadas estas escolas em relação à sede da província (que vão até cerca de quinhentos Kms). Também é importante referir a inexperiência de estudos do género e o pouco tempo que se teve para mobilizar e informar os alunos dos objetivos do estudo por forma a incentivar a sua participação no mesmo, pois que é raro estes alunos participarem neste tipo de estudos. Por este facto, após a realização do inquérito, na altura do tratamento de dados invalidaram-se cerca de cem questionários, por não possuírem mais da metade das informações solicitadas.

Portanto, apesar destas limitações, este estudo, o primeiro do género realizado na província do Moxico em Angola, na nossa opinião, apresenta resultados importantes, que caso sejam considerados pelas autoridades que gerem o sistema de ensino, poderão ajudar a explicar parte dos problemas relacionados com o desempenho académico a matemática, dos alunos que frequentam o ensino secundário na província do Moxico em Angola e a ter uma melhor perceção das características sociodemográficas e socioeconómicas dos mesmos.

Capítulo 6

Considerações finais do trabalho

O presente trabalho teve como objetivo, realizar uma abordagem sobre os aspetos teóricos essenciais do modelo de regressão linear e mostrar como esta metodologia pode ser aplicada na resolução dos problemas do dia a dia, através de um estudo que tinha como finalidade identificar os fatores que influenciam as notas dos alunos do ensino secundário da província do Moxico em Angola, à disciplina de matemática.

Os conteúdos que dão resposta a estes objetivos, foram subdivididos em seis capítulos. Após a introdução, a qual corresponde ao primeiro capítulo, apresentam-se os conteúdos ligados aos modelos de regressão linear simples e múltipla respetivamente, no segundo e terceiro Capítulos. Os pressupostos impostos ao modelo de regressão ocupam o quarto capítulo. O quinto capítulo apresenta um estudo prático que nos permitiu identificar os fatores que influenciam as notas a matemática dos alunos do ensino secundário na província do Moxico em Angola. Este estudo mostra também como se pode aplicar a estatística, em geral e a regressão linear em particular na busca de soluções para a resolução de problemas do quotidiano.

Com base na abordagem feita e nos resultados que o trabalho apresenta, pode-se concluir que os objetivos pelos quais foi concebido e realizado foram cumpridos.

Em cada um dos capítulos mencionados acima, apresentaram-se algumas conclusões sobre as matérias abordadas. A seguir será apenas apresentada uma síntese das principais conclusões descritas nos capítulos anteriores deste trabalho.

Na nossa opinião, a abordagem teórica apresentada sobre o modelo de regressão linear constitui uma ferramenta útil para o ensino e a aprendizagem desta metodologia estatística por parte de estudantes do ensino secundário e universitário e outras pessoas que queiram aprender ou aprofundar os seus conhecimentos neste domínio.

Sobre o estudo prático que envolveu alunos do ensino secundário da província do Moxico em Angola, o modelo de regressão linear obtido, identifica as notas a língua portuguesa e a física, a idade, a renda familiar (média) e a insatisfação elevada com o ambiente escolar como os fatores que influenciam as notas dos alunos a matemática.

Este modelo, na nossa opinião, poderá ajudar as autoridades educativas da província do Moxico em Angola e não só, a tomar medidas por forma a melhorar o desempenho dos alunos nesta disciplina.

No que concerne às limitações, este trabalho centrou a sua abordagem apenas nos aspetos essenciais dos modelos de regressão linear simples e múltipla. Os tópicos avançados da teoria de regressão como, por exemplo, a estimação dos parâmetros pelo método dos mínimos quadrados ponderados (WLS), que se aplica quando existe heteroscedasticidade no modelo, e dos mínimos quadrados em dois passos (2SLS), que é aplicado quando a variável residual se correlaciona com uma ou mais variáveis independentes, e outros tópicos e tipos de modelos de regressão não foram abordados, devido ao caráter e objetivos que foram estabelecidos.

6.1 Sugestões para trabalhos futuros

Os resultados obtidos neste trabalho no geral e no estudo prático em particular são, na nossa opinião, animadores. Por este facto, acredita-se que este trabalho venha a incentivar a realização de outros, a que se sugere o seguinte:

- Que se possa reformular o questionário, adicionando-lhe mais unidades de análises e melhorar a especificação das questões que foram consideradas neste estudo. Por outro lado, aconselha-se que antes da aplicação dos inquéritos se promovam encontros de esclarecimento com todos os intervenientes no processo, por formas a consciencializá-los sobre a importância do trabalho a realizar e dos efeitos que os resultados possam vir a causar nas suas vidas, caso sejam tidos em conta pelas entidades do estado;

- Ao analisar os dados, para além do modelo de regressão linear, sugere-se que se utilizem outras metodologias estatísticas, com o objetivo de explicar outras informações contidas nos dados;

- E por fim, recomenda-se que se generalize este tipo de estudos a outros níveis de ensino e províncias do país, por formas a obter informações sobre a realidade de outro tipo de alunos.

Bibliografia

[1] ANDRÉ, B. Z. e LARRECHEA. E. M. (2016). *Baixo rendimento na aprendizagem da Matemática: um estudo de caso dos estudantes do II Ciclo do ensino secundário em Lubango-Angola*. Lubango. Disponível em: https://www.iusur.edu.uy/ publicaciones/index.php/RESUR/article/view/24/66.

[2] BUTAKOR, P. K., AMPADU, E. e COLE, Y. (2017). Ghanaian Students in TIMSS 2011: Relationship between Contextual Factors and Mathematics Performance,*African Journal of Research in Mathematics, Science and Technology Education*, 21:3, 316-326, DOI: 10.1080/18117295.2017.1379281

[3] CAIADO, J. (2016). *Métodos de Previsão em Gestão* . 2^{nd} ed. Lisboa: Edições Sílabo.

[4] COHEN, J. (1988). *Statistical power and analysis for the behavioral sciences*. 2^{nd} ed. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

[5] CORAK, M. (2013), Income inequality, equality of opportunity, and intergenerational mobility, *The Journal of Economic Perspectives*,Vol. 27/3, pp. 79-102.

[6] COSTA, A. M. (2007). *Importância da língua portuguesa na aprendizagem da matemática*. Tese (Dissertação de Mestrado em Estudos da Criança - Área de Especialização em Ensino e Aprendizagem da Matemática). UMinho, Braga.

[7] Critical values for the Durbin-Watson Test: 5% Significance Level. Stanford. Disponível em: https://web.stanford.edu/ clint/bench/dwcrit.htm.

[8] D'HAINAUT, L. (1992). *Conceitos e Métodos da Estatística*. Vol. nº 2. Lisboa: Fundação Calouste Gulbenkian.

[9] DGEEC-Direção Geral de Estatísticas de Educação e Ciências. (2017). *Resultados Escolares por Disciplina - 3º ciclo do ensino público ano letivo 2014/2015*. Lisboa: DGEEC. Disponível em: <http://www.dgeec.mec.pt/np4/61.

[10] ESSIEN, A. A. (2018). The Role of Language in the Teaching and Learning of Early Grade Mathematics: An 11-year Account of Research in Kenya, Malawi and South Africa. *African Journal of Research in Mathematics, Science and Technology Education*. DOI: 10.1080/181 295.2018.1434453.

[11] FERNANDES, C. A. F. (2007). *A matemática na disciplina de ciências físcio-químicas, um estudo sobre as atitudes de alunos do 9º ano de escolaridade*. Tese (Mestrado em educação, área de especialização em supervisão pedagógica no ensino de Física e Química). UMinho, Braga, Portugal.

[12] FERNANDES, E. M. da G. P. (1999). *Estatística Aplicada*. Braga: Universidade do Minho.

[13] FERREIRA, M. C. C. dos S. (2013). *Modelos de Regressão: Uma Aplicação em Medicina Dentária*. Tese para obtenção do Título de Mestre. Universidade Aberta, Lisboa,Portugal.

[14] FERREIRA, M. J. e TAVARES, I. (2002). *VI Notas sobre a História da Estatística - Dossiers Didáticos*, INE. Disponível em: http://alea-estp.ine.pt/Html/statofic/html/ dossier/doc/dossier6.pdf.

[15] FREIXO, M. J. V. (2012). *Metodologia científica*. 4. ed. Lisboa: Instituto Piaget.

[16] GUIMARÃES, R. C. e CABRAL, J. A. S. (1998). *Estatísticas*. Edição Revista, Lisboa: McGraw-Hill.

[17] HILL, M. M. e HILL, A. (2016). *Investigação por questionário*. 2. ed. Lisboa: Edições Sílabo.

[18] HOWIE, S. J. (2003). Language and other background factors affecting secondary pupils' performance in Mathematics in South Africa, *African Journal of Research in Mathematics, Science and Technology Education*, 7:1, 1-20.

[19] Instituto Nacional de Estatística de Angola (INE). (2013). *Inquérito Integrado sobre o Bem-Estar da População, IBEP - Relatório Analítico, Perfil da Pobreza*. Vol. III. Luanda: INE.

[20] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). (2014). *Relatório Nacional PISA 2012 - Resultados brasileiros*. Brasília: Fundação Santillana, INEP, MEC. Disponível em: http://download.inep.gov.br/acoes-internacionais/pisa/resultados/2014/relatorio-nacional-pisa-2012-resultados-brasileiros.pdf.

[21] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). (2016). Ministério da Educação. *Brasil no PISA 2015: análises e reflexões sobre o desempenho dos estudantes brasileiros*. Brasília: Fundação Santillana, INEP, MEC. Disponível em: https://www.oecd.org/pisa/PISA-2015-Brazil-PRT.pdf.

[22] JOHNSTON, J. e DINARDO, J. (1997). *Econometric Methods*. 4th Ed., Nova Iorque: McGraw-Hill Education.

[23] KHUSAINOVA, R.M., SHILOVA, Z.V. e CURTEVA, O.V. (2016). Selection of Appropriate Statistical Methods for Research Results. Processing. *IEJME-Mathematics Education*, 11(1), 303-315.

[24] *Lei n.° 17/16, de 7 de Outubro de 2016*. Lei de Bases do Sistema de Educação e Ensino em Angola. Disponível em: https://www.lexlink.eu/codigosimples/geral/ 639922/lei-de-bases-de-educacao-e-ensino-lei-n-1716-de-7-de-outubro/26797/por-tema.

[25] LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. (2000). *Estatística: teoria e aplicações usando o Microsoft Excel em português*. Rio de Janeiro: LTC.

[26] LIBERATO, E. (2014). Avanços e retrocessos da educação em Angola. *Rev. Bras. Educ.*, vol.19, nº. 59. DOI:10.1590/S1413-24782014000900010.

[27] MARÔCO, J. (2014). *Análise estatística - com o SPSS Statistics*. 6ª ed. Lisboa: ReportNumber.

[28] MARÔCO, J., GONÇALVES, C., LOURENÇO, V. e MENDES, R. (2016). *PISA 2015-Potugal: Literacia cientícfica, literacia de leitura & literacia matemática*. Lisboa: IAVE, I.P.

[29] Ministério da Educação da República de Angola (MED). (2012). *Balanço da implementação da 2ª reforma educativa em Angola*. MED - Governo de Angola. Disponível em: http://www.med.gov.ao/VerPublicacaoaspx?id=705.

[30] MENEZES-FILHO, N. (2006). *Os determinantes do desempenho escolar do Brasil*, São Paulo: Inst. Futuro Brasil, Ibmec-SP, FEA-USP, 2006. Disponível em: http://www.todospelaeducacao.org.br/arquivos/biblioteca/f4e8070a-8390-479c-a532-803bbf14993a.pdf.

[31] MONTGOMERY, D. C., PECK, E. A., e VINING, G. G. (2012). *Introduction to linear regression analysis*. 5^{th} ed. New York: Wiley.

[32] MULLIS, I. V. S., MARTIN, M. O., FOY, P. e ARORA, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Disponível em: https://timssandpirls.bc.edu/timss2011/downloads/T11IRMathematics-FullBook.pdf.

[33] MURTEIRA, B., RIBEIRO, C. S., SILVA, J. A. PIMENTA, C. e PIMENTA, F. (2015). *Introdução à Estatística*, 3ª Edição, Lisboa: Escolar Editora.

[34] OCDE (2013a). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing. DOI: 10.1787/9789264190511.

[35] OCDE (2013b). *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II)* , PISA, OECD Publishing. DOI: 10.1787/9789264201132.

[36] OCDE (2014a). *PISA 2012 Results in Focus What 15-year-olds know and what they can do with what they know*, PISA, OECD Publishing. Disponível em: https://www.oecd.org/pisa/.../pisa-2012-results-overview.pdf.

[37] OCDE (2014b). *PISA 2012 Results: What Students Know and Can Do - Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)* , PISA, OECD Publishing. DOI: 10.1787/9789264208780.

[38] OCDE (2017). *PISA 2015 Results (Volume III): Students' Well-Being*, PISA, OECD Publishing. DOI: 10.1787/9789264273856.

[39] OGUNSOLA-BANDELE, M. F. (1996). Mathematics in physics - Which way forward: the influence of mathematics on students' attitudes to the teaching of physics. *ERIC*. Disponível em: https://eric.ed.gov/?id=ED400199.

[40] OLIVEIRA, M. M. de., SANTOS, L. D. e FORTUNA, N. (2011). *Econometria*. Lisboa: Escolar Editora.

[41] PEDROSA, A. C. e GAMA, S. M. A. (2016). *Introdução computacional à probabilidade e estatística com Excel*. 3. Ed. Porto: Porto Editora.

[42] PESTANA, D. e VELOSA, S. (2002). *Introdução à probabilidade e à estatística*. Lisboa: Fundação Calouste Gulbenkian.

[43] PESTANA, M. H. e GAGEIRO, J. N. (2005). *Descobrindo a regressão com a complementaridade do SPSS*. 1ª Edição. Lisboa: Edições Sílabo.

[44] PESTANA, M. H. e GAGEIRO, J. N. (2014). *Análise de dados para ciências sociais - A complementaridade do SPSS*. 6ª Edição. Lisboa: Edições Sílabo.

[45] PORTELA, C. e FARIA, S. (2015). *Alguns resultados da análise do desempenho dos alunos portugueses no teste de Matemática PISA 2009*. In: Seminário sobre Investigação em Educação e os Resultados do PISA, 2015. Lisboa. Investigação em Educação e os Resultados do PISA. Lisboa: CNC, 5 Dez. 2015. pp. 97-102 .

[46] REIS, E. (2017). *Estatística Descritiva*. 7ª Edição. Lisboa: Edições Sílabo.

[47] RODRIGUES, S. C. A. (2012). *Modelo de Regressão Linear e suas Aplicações*. Relatório de estágio para obtenção do título de Mestre, UBI, Covilhã, Portugal.

[48] SCHIVANI, J. e SOUSA, G. (2015). *Do método dos mínimos quadrados de Legendre (1805) à regressão linear de Galton (1875)* . In: XI Seminário Nacional de História da Matemática, 2015, v. 11, Natal. Disponível em: https://slidex.tips/download/juliana-schivani -ufrn-giselle-sousa-ufrn.

[49] SPIEGEL, M. (1984). *Estatística*. 2ª Edição. São Paulo: McGraw-Hill do Brasil, Ltda.

[50] UNESCO (2015). *Relatório de monitoramento Global de EPT 2015, Educação para Todos 2000-2015: Progressos e desafios*. Paris: UNESCO.

[51] VALLE, P. O. e REBELO, E. (2002). O uso de regressores *Dummy* na especificação de modelos com parâmetros variáveis. *Revista de Estatística*. Vol. nº 3, 3º Quadrimestre, Lisboa: INE-Portugal.

[52] ZAU, F. (2009). *Educação em Angola: novos trilhos para o desenvolvimento*. Luanda: Movilivros.

Apêndice A

Anexos

A.1 - Anexos Capítulo 4

Tabela A.1: Tabelas de valores críticos de Durbin-Watson

	Região de rejeição e não rejeição de H_0: Não existe autocorrelação				
Valor de d	$[0; d_L[$	$[d_L; d_U[$	$[d_U; 4-d_U[$	$[4-d_U; 4-d_L[$	$[4-d_L; 4[$
Decisão	Rejeitar H_0	Nada se pode concluir	Não Rejeitar H_0	Nada se pode concluir	Rejeitar H_0

Valores críticos de d_L e d_U para $\alpha = 0.05$														
p / n	1		2		3		4		5		10		15	
	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	2.32	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	1.80	1.73	1.81	1.72	1.82	1.67	1.87	1.61	1.93
250			1.78	1.80	1.78	1.81	1.77	1.82	1.76	1.83	1.72	1.87	1.68	1.91
300			1.80	1.82	1.80	1.82	1.79	1.83	1.78	1.84	1.75	1.87	1.71	1.91
350			1.82	1.83	1.81	1.84	1.81	1.84	1.80	1.85	1.77	1.88	1.74	1.91
400			1.83	1.84	1.83	1.85	1.82	1.85	1.82	1.87	1.79	1.88	1.76	1.91
450			1.84	1.85	1.84	1.85	1.83	1.86	1.83	1.86	1.80	1.89	1.78	1.91
500			1.85	1.86	1.85	1.86	1.84	1.87	1.84	1.87	1.82	1.89	1.80	1.91

Fontes: Marôco, (2014) e https://web.stanford.edu/ clint/bench/dwcrit.htm (valores sombreados)

A.2 - Anexos Capítulo 5

Tabela A.2: Estimativas dos coeficientes de regressão do modelo ajustado (obtidas através do método $Enter$)

			IC a 95% para β	
	Estimativas	$p-value$	Limite inferior	Limite superior
Constante	2,789	0,000	1,859	3,719
Sexo do aluno (Masculino)	0,004	0,960	-0,143	0,150
Estado civil (Solteiro)	-0,091	0,336	-0.277	0,095
Notas a Língua Portuguesa	0,381	0,000	0,343	0,420
Notas a Física	0,451	0,000	0,411	0,490
Idade	-0,043	0,000	-0,064	-0,022
Renda familiar (Média)	0,273	0,002	0,102	0,444
Situação laboral(estudante trabalhador)	0,153	0,235	-0,100	0,407
Classificação atrib. p/ aluno ao prof. Mat. (Bom)	0,111	0,405	-0,150	0,371
Grau de satisfação c/ ambiente escolar				
• Muito insatisfeito	0,310	0,215	-0,181	0,801
• Insatisfeito	-0,337	0,074	-0,708	0,033
• Satisfeito	-0,164	0,158	-0,392	0,064
• Muito satisfeito	-0,169	0,081	-0,360	0,021
Escola e localização				
• Escola localizada em zona urbana	-0,071	0,557	-0,309	0,167
• Escola localizada em zona periurbana	-0,060	0,620	-0,295	0,176
ANOVA $(p-value < 0,001)$				
R_a^2 = 0,866				

Classes de referência: Sexo do aluno - Feminino; Estado civil - Casado; Renda familiar - Baixa; Situação laboral - Não trabalhador estudante; Como os alunos classificam os professores de matemática - Maus; Grau de satisfação com o ambiente escolar - Nem insatisfeito nem satisfeito; Escola e localização - Escola localizada em zona suburbana.

Tabela A.3: Variáveis excluídas do modelo (pelos métodos $Stepwise$ e $Forward$)

Variáveis	Estimativas	$p-value$
Sexo do aluno (masculino)	0,000	0,981
Estado civil (solteiro)	-0,453	0,222
Situação laboral (estudante trabalhador)	0,028	0,161
Classificação atrib. p/ aluno ao prof. Mat. (Bom)	-0,041	0,583
Grau de Satisfação c/ o ambiente escolar		
• Insatisfeito	-0,023	0,251
• Satisfeito	-0,004	0,828
• Muito satisfeito	-0,018	0,376
Escola e localização		
• Escola localizada em zona urbana	-0,011	0,575
• Escola localizada em zona periurbana	0,004	0,838

Classes de referência: Sexo - Feminino; Estado civil - Casado; Situação laboral - Não trabalhador estudante; Classificação atribuída pelo aluno ao professor de Matemática - Mau; Grau de satisfação com o ambiente escolar - Nem insatisfeito nem satisfeito; Escola e localização - Escola localizada em zona suburbana.

CARIMBO DA ESCOLA PARA VALIDAÇÃO DA FICHA DE INQUÉRITO

UNIVERSIDADE DA BEIRA INTERIOR

FICHA DE INQUÉRITO PARA RECOLHA DE DADOS DE ESTUDANTES DE TRÊS ESCOLAS DO 2º CICLO DO ENSINO SECUNDÁRIO E FORMAÇÃO DE PROFESSORES DA CIDADE DO LUENA, PROVÍNCIA DO MOXICO EM ANGOLA, PARA ELABORAÇÃO DA BASE DE DADOS, QUE PODERÁ SER USADA NA PARTE PRÁTICA DA CADEIRA DE PROJECTO DE ENSINO II E DA DISSERTAÇÃO DO CURSO DO 2º CICLO (MESTRADO) EM MATEMÁTICA PARA PROFESSORES NA UNIVERSIDADE DA BEIRA INTERIOR EM PORTUGAL.

EM PRIMEIRO LUGAR AGRADECEMOS DESDE JÁ PELA SUA COLABORAÇÃO NESTE TRABALHO.
RESPONDA O QUESTIONÁRIO ASSINALANDO EM CADA PERGUNTA APENAS UM CAMPO COM X

DADOS PESSOAIS E SITUAÇÃO ACADÉMICA DO ESTUDANTE

SEXO: MASCULINO ☐ FEMININO ☐ | ESTADO CIVIL: SOLTEIRO ☐ CASADO ☐ VIVE MARITALMENTE COM ALGUÉM ☐ | IDADE: (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (M)

RESULTADO OBTIDO NO ANO LECTIVO 2016: APROVADO ☐ REPROVADO ☐ |

MÉDIAS FINAIS OBTIDAS EM MATEMÁTICA, LÍNGUA PORTUGUESA E FÍSICA NO ANO LECTIVO ANTERIOR (2016):

MATEMÁTICA: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)

LÍNGUA PORTUGUESA: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)

FÍSICA: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)

COMO CLASSIFICA O TEU PROFESSOR DE MATEMÁTICA DO ANO PASSADO: BOM ☐ MAU ☐ |

QUAL É O TEU NÍVEL DE SATISFAÇÃO COM A TUA ESCOLA: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

COMO CLASSIFICA AS RELAÇÕES NO AMBIENTE ESCOLAR?

RELAÇÃO ENTRE ALUNOS E PROFESSORES: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

RELAÇÃO ENTRE ALUNOS: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

RELAÇÃO ENTRE ALUNOS E A DIRECÇÃO DA ESCOLA: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

SITUAÇÃO SOCIAL DO ESTUDANTE

CONDIÇÃO ECONÓMICA DA FAMÍLIA (RENDA): ALTA ☐ MÉDIA ☐ BAIXA ☐

TEM UM EMPREGO: SIM ☐ NÃO ☐

DISTÂNCIA DA CASA À ESCOLA EM QUILÓMETROS: (-1) (1) (2) (3) (4) (M)

ZONA EM QUE HABITA: URBANA ☐ NÃO URBANA ☐

MUITO OBRIGADO PELA COLABORAÇÃO.

Figura A.1: Questionário usado na investigação

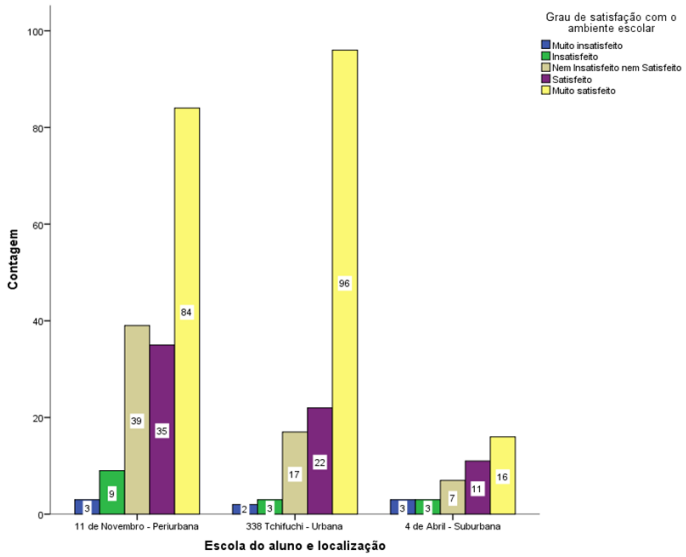

Figura A.2: Gráfico de frequência dos alunos por escola e o grau de satisfação com o ambiente escolar

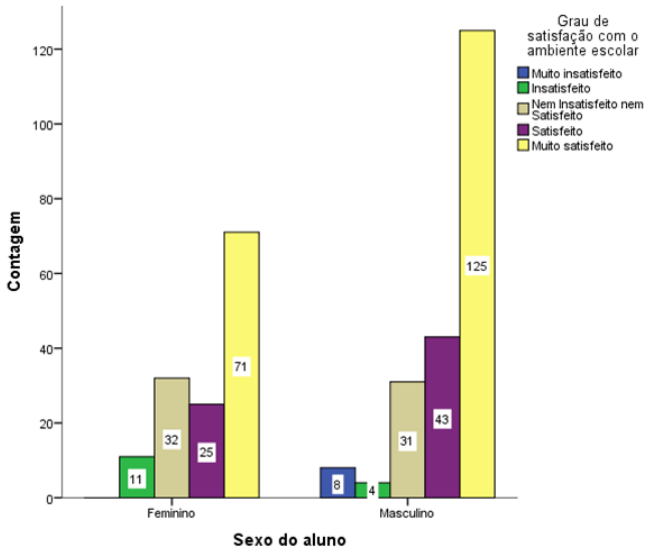

Figura A.3: Gráfico de frequências dos alunos por sexo e o grau de satisfação com o ambiente escolar

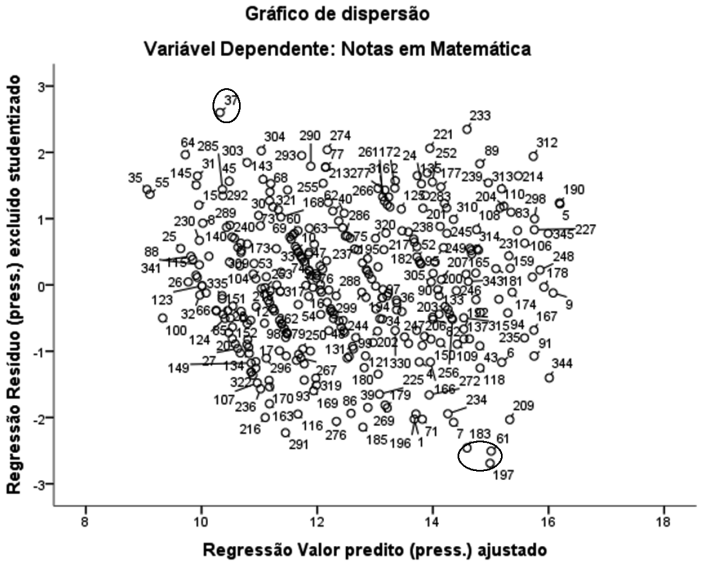

Figura A.4: Gráfico de resíduos estudantizados excluídos versus valor predito estandardizado

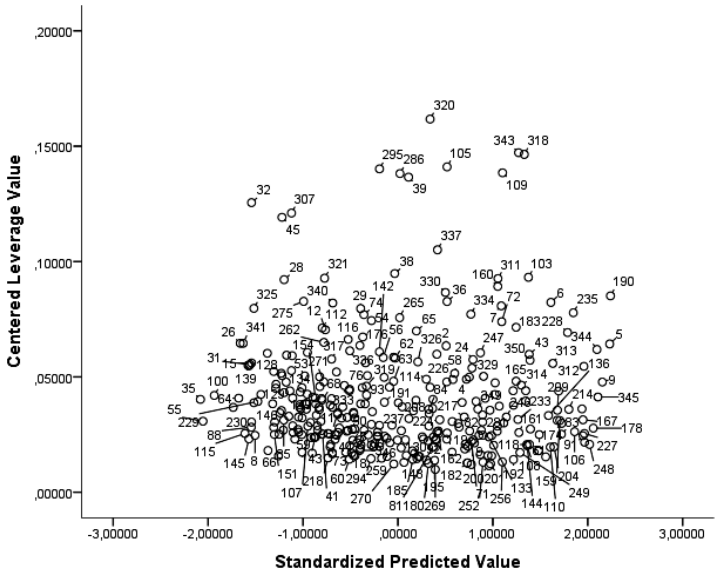

Figura A.5: Gráfico do valor centrado da Leverage versus valor predito estandardizado

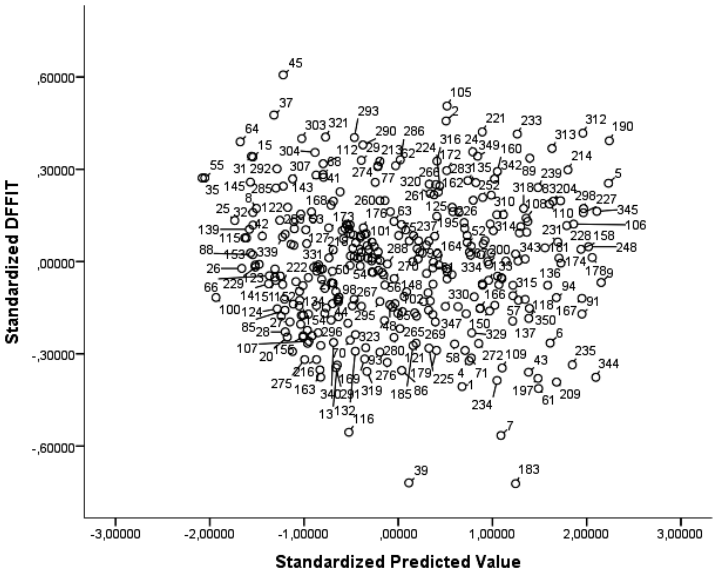

Figura A.6: Gráfico de DFFIT estandardizados versus valor predito estandardizado