UNIVERSIDADE ABERTA

Modelação Estatística aplicada a Admissão dos Estudantes na Universidade de São Tomé e Príncipe

Wadiley Sousa do Nascimento

Mestrado em Estatística, Matemática e Computação

Área de especialização em Estatística Computacional

2024

UNIVERSIDADE ABERTA

Modelação Estatística aplicada a Admissão dos Estudantes na Universidade de São Tomé e Príncipe

Wadiley Sousa do Nascimento

Mestrado em Estatística, Matemática e Computação

Área de especialização em Estatística Computacional

Dissertação orientada pela Prof. Doutora Elisabete Teresa M. Almeida Carolino
e Co-orientada pela Prof. Doutora Maria do Rosário Ramos

2024

“A mente que se abre a uma nova ideia,
jamais voltará ao seu tamanho original.”

Albert Einstein

Resumo

O presente estudo investiga alguns fatores que influenciam a admissão de estudantes na Universidade de São Tomé e Príncipe (USTP) por meio da aplicação de técnicas avançadas de análise estatística, de forma a proporcionar *insights* que possam conduzir a melhorias nas práticas de selecção e a formulação de políticas mais inclusivas.

Dada a heterogeneidade das disciplinas que compõem os cursos do ensino secundário, esta análise contemplou duas situações distintas: uma para os alunos de Letras (História como nuclear) e outra para os alunos de Ciências (Matemática como nuclear). Os resultados revelaram que variáveis como a idade, género, notas em disciplinas específicas (Biologia, Física/Química, Língua Portuguesa, Matemática, História, Direito, Psicologia/Sociologia), e a escola de origem dos candidatos possuem efeitos estatisticamente significativos tanto na média final do ensino secundário quanto na probabilidade de admissão em diferentes departamentos da universidade. A análise indicou que mais de 75% da variação na média do ensino secundário pode ser explicada pelas variáveis incluídas no modelo de Regressão Linear Múltipla. Com base nos dados do período em estudo, verificou-se que no grupo dos alunos admitidos em cursos de Ciências a média final diminui em média 0,211 pontos com o aumento da idade e que, inversamente, a média final dos alunos dos cursos de Letras aumenta 0,009 pontos, em média, com o aumento da idade.

A Regressão Logística Multinomial mostrou que os modelos que incorporaram as variáveis idade, média final do 12º ano, notas em Matemática, Língua Portuguesa, História, Direito, Física/Química, Sociologia/Psicologia, género, residência e as escolas onde os estudantes concluíram o ensino secundário contribuem significativamente para a discriminação dos cursos afetos aos departamentos da USTP. Igualmente, mostrou que a probabilidade de inscrição em determinados cursos diminui ou aumenta significativamente com base nas variáveis analisadas.

O estudo realizado mostrou que a modelação estatística é uma ferramenta eficaz para compreender e melhorar o processo de admissão na USTP, fornecendo *insights* valiosos para a tomada de decisões e contribuindo para a formação de um corpo discente mais diversificado e talentoso. Além disso, os modelos revelaram que mais 60% dos estudantes matricularam-se nos cursos afetos aos departamentos onde as probabilidades de admissão eram mais altas.

Palavras-chave: Modelação Estatística, Regressão Linear Múltipla, Regressão Logística Multinomial, Acesso ao Ensino Superior, Universidade de São Tomé e Príncipe.

Abstract

The present study investigates the factors influencing student admission at the University of São Tomé and Príncipe (USTP) through the application of advanced statistical analysis techniques, providing insights that could lead to improvements in selection practices and the formulation of more inclusive policies.

Given the heterogeneity of subjects that constitute secondary education courses, this analysis considered two distinct scenarios: one for Humanities students (with History as the core subject) and another for Science students (with Mathematics as the core subject). The results revealed that variables such as age, gender, grades in specific subjects (Biology, Physics/Chemistry, Portuguese Language, Mathematics, History, Law, Psychology/Sociology), and the candidates' originating schools have statistically significant effects on both the final secondary education average and the probability of admission to different university departments. The analysis indicated that more than 75% of the variation in the secondary education average can be explained by the variables included in the Multiple Linear Regression model. Based on the data from the study period, it was found that in the group of students admitted to Science courses, the final average decreases by an average of 0.211 points with increasing age, whereas, conversely, the final average of Humanities students increases by an average of 0.009 points with increasing age.

The Multinomial Logistic Regression showed that models incorporating variables such as age, final average of the 12th grade, grades in Mathematics, Portuguese Language, History, Law, Physics/Chemistry, Sociology/Psychology, gender, residence, and the schools where students completed secondary education significantly contribute to distinguishing courses related to USTP departments. It also demonstrated that the likelihood of enrollment in certain courses significantly decreases or increases based on the analyzed variables.

The study demonstrated that statistical modeling is an effective tool for understanding and improving the admission process at USTP, providing valuable insights for decision-making and contributing to the formation of a more diverse and talented student body. Additionally, the models revealed that more than 60% of students enrolled in courses related to departments where the probabilities of admission were highest.

Keywords: Statistical Modeling, Multiple Linear Regression, Multinomial Logistic Regression, Higher Education Access, University of São Tomé and Príncipe.

Agradecimentos

Para que esse trabalho tivesse lugar começo o agradecer ao Pai Celestial, o Todo-Poderoso, pela vida e saúde concedida para levar a frente a árdua tarefa de estudante/trabalhador, na qual sem o Seu apoio tudo estaria deveras perdido.

Agradeço aos meus pais, Hilário Jesus do Nascimento e Manuela Francisca Sousa Ponte do Espírito, à minha filha Emily Nascimento e aos meus irmãos, cujo apoios foram inestimáveis ao longo da minha jornada académica.

A seguir agradeço a minha Orientadora, Professora Doutora Elisabete Teresa M. Almeida Carolino, e à minha Co-orientadora, Prof. Doutora Maria do Rosário Ramos, pelas dedicação e sapiência com que conduziram o desenvolvimento deste trabalho.

Os meus agradecimentos são extensivos as colegas de trabalho pela atenção e pela força dadas através do companheirismo sempre que necessário. A todos os dirigentes das Unidades Orgânicas que compõem a Universidade de São Tomé e Príncipe, pela disponibilidade e autorização, tornando possível a recolha dos dados.

A todos, muito obrigado.

ÍNDICE

Resumo II
Abstract III
Agradecimentos IV
Lista de Tabelas VII
Lista dos Quadros IX
Lista de Figuras IX
Lista de Abreviaturas X
INTRODUÇÃO 1
Justificativas 1
Objectivo Geral 2
Objectivos Específicos 2
Estrutura do trabalho 2
CAPÍTULO I – A UNIVERSIDADE DE SÃO TOMÉ E PRÍNCIPE (USTP) 3
1.1 História da Universidade de São Tomé e Príncipe (USTP) 3
1.2 Admissão dos Estudantes na USTP 5
CAPÍTULO II – MÉTODOS ESTATÍSTICOS 6
2.1 Testes Paramétricos e Não Paramétricos 6
2.1.1. Pressupostos de aplicabilidade para a comparação de dois ou mais grupos independentes 7
2.1.2. Testes para k Amostras Independentes 9
2.2 Medidas de Associação – Correlação 16
2.3 Modelos Lineares Generalizados 18
2.3.1 Estimação de parâmetros dos MLG 19
2.4 Regressão Linear 22
2.4.1 Modelo de Regressão Linear 22
2.4.2 Estimação e Significância do Modelo 24
2.4.3 Qualidade do Ajustamento do Modelo 30
2.4.4 Validação dos pressupostos Modelo de Regressão Linear Múltipla 31
2.4.5 Diagnóstico de Outliers e Observações Influentes 43

2.4.6 Modelos de regressão com variáveis binárias (Dummy)....47
2.4.7 Métodos e Critérios de Seleção de Modelos 49
2.4.8 Previsão de novas observações....55
2.5 Regressão Logística....56
2.5.1 Modelo de Regressão Logística Binária....56
2.5.2 Pressuposto Modelo de Regressão Logística Binária....58
2.5.3 Estimação e Significância do Modelo 59
2.5.4 Qualidade do Ajustamento do Modelo....62
2.5.5 Métodos e Critérios de Seleção de Modelos 63
2.5.6 Diagnóstico de Outliers e Classificação por recurso a Regressão Logística....65
2.6 Regressão Logística Multinomial....72
2.6.1 Classificação por recurso a Regressão Logística Multinomial....73
CAPÍTULO III – APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS....75
3.1 Apresentação dos Dados 75
3.1.1 Características das variáveis qualitativas 76
3.1.2 Características das variáveis quantitativas 79
3.1.3 Testes de Comparação – Análise Inferencial 80
3.2 Aplicação de Regressão Linear....86
3.2.1 Estimação da Média Final do Ensino secundário dos estudantes dos Cursos de Ciências (Matemática - Base de Dados I) 87
3.2.2 Estimação da Média Final do Ensino secundário dos estudantes dos Cursos de Letras (História – Base de Dados II) 92
3.3 Aplicação de Regressão Logística....97
3.3.1 Estimação dos Modelos para a afetação dos estudantes dos Cursos de Ciências (Matemática) nos departamentos da USTP 97
3.3.2 Estimação dos Modelos para a afetação dos estudantes dos Cursos de Letras (História) nos departamentos da USTP)....102
CONCLUÕES....107
RECOMENDAÇÕES, LIMITAÇÕES E PESQUISAS FUTURAS....109
REFERÊNCIAS BIBLIOGRÁFICAS....110
APÊNDICES....112

Lista de Tabelas

Tabela 1 – Tabela de ANOVA a um Factor (One-Way ANOVA) 11
Tabela 2 – Tabela de ANOVA Factorial (Two-Way ANOVA) 12
Tabela 3 – Classificação do Coeficiente de Correlação de Pearson e Spearman 17
Tabela 4 – Tabela de ANOVA de Regressão 29
Tabela 5 – Região cítica ou de Decisão do teste DW 37
Tabela 6 – Capacidades Preditivas 70
Tabela 7 – Distribuição por Distrito/Residência 77
Tabela 8 – Distribuição por Escolas ou Centros de Formação 78
Tabela 9 – Estatísticas das variáveis métricas da Base de Dados I 79
Tabela 10 – Estatísticas das variáveis métricas da Base de Dados II 80
Tabela 11 – Tabela da ANOVA a um Factor (Base de Dados I) 82
Tabela 12 – Testes de Comparação Múltipla (Base de Dados I). 83
Tabela 13 – Tabela da ANOVA a um Factor (Base de Dados II) 84
Tabela 14 – Testes de Comparação Múltipla (Base de Dados II) 84
Tabela 15 – Testes de Homogeneidade de Variância (Base de Dados I). 85
Tabela 16 – Composição da Base de dados por Departamentos (Base de Dados II). 85
Tabela 17 – Teste de significância do Modelo de Regressão. 88
Tabela 18 – Coeficientes e significâncias do modelo de regressão. 89
Tabela 19 – Variáveis excluídas do modelo de regressão. 89
Tabela 20 – Teste de Normalidade do resíduo 91
Tabela 21 – Teste de Heteroscedasticidade. 92
Tabela 22 – Teste de significância do Modelo de Regressão. 93
Tabela 23 – Coeficientes e significâncias do modelo de regressão. 94
Tabela 24 – Teste de Normalidade do resíduo 95
Tabela 25 – Critérios de ajuste do modelo (Base de Dados I) 97

Tabela 26 – Adequação do Ajuste do Modelo (Base de Dados I) 98
Tabela 27 – Testes de Razão de Verosimilhança (Base de Dados I). 98
Tabela 28 – Estimação dos parâmetros (Base de Dados I) 101
Tabela 29 – Classificação dos candidatos por departamentos (Base de Dados I) 102
Tabela 30 – Critérios de ajuste do modelo (Base de Dados II) 103
Tabela 31 – Adequação do Ajuste do Modelo (Base de Dados II) 103
Tabela 32 – Estimação dos parâmetros (Base de Dados II) 105
Tabela 33 – Classificação dos candidatos por departamentos (Base de Dados II) 106

Lista dos Quadros

Quadro 1 – Transformações matemáticas para linearizar as relações não lineares 39

Quadro 2 – Interpretação de Modelos envolvendo Logaritmos.....39

Quadro 3 – Classificação do poder de discriminação - Curva de ROC.....71

Quadro 4 – Resumo do teste Kruskal-Wallis (Base de Dados I).....85

Quadro 5 – Resumo do teste Kruskal-Wallis (Base de Dados II).....86

Lista de Figuras

Figura 1 – Exemplo da Curva ROC.....71

Figura 2 – Distribuição por Sexo.....77

Figura 3 – Distribuição dos Estudantes por ano de conclusão do curso.....78

Figura 4 – Comparações por Método Pairwise de Escola Proveniente (Base de Dados I) 83

Figura 5 – Comparações por Método Pairwise de Escola Proveniente (Base de Dados II) 84

Figura 6 – Comparações por Método Pairwise de Curso Inscrito (Base de Dados I) 86

Figura 7 – Diagnóstico do *outliers* (Base de Dados I).....90

Figura 8 – Diagnóstico do *outliers* pela Leverage (Base de Dados I) 91

Figura 9 – Diagnóstico do *outliers* (Base de Dados II) 94

Figura 10 – Diagnóstico do *outliers* pela Leverage (Base de Dados II).....95

Lista de Abreviaturas

H_0 – Hipótese Nula (hipótese de nulidade)

H_a – Hipótese Alternativa

AG – Água Grande

AIC – (do inglês Akaike Information Criterion) Critério de Informação de Akaike,

ANOVA – (do inglês *Analysis of Variance*) Análise de Variância

BIC – (do inglês *Bayesian Information Criterion*) Critério de Informação de Bayesiano

DCE – Departamento de Ciências da Educação

DCE&EE – Departamento de Ciências Económica e de Ciências Exactas e Engenharias

DCNVA – Departamento de Ciências da Natureza, da Vida e do Ambiente

DL&CHS – Departamento de Língua e de Ciências Humanas e Sociais

ES – Escola Secundária

ESMXinho – Escola Secundária Mé Xinhô

FCT – Faculdade de Ciências e das Tecnologias

ISCSVSM – Instituto de Ciências da Saúde Victor Sá Machado

ISEC – Instituto Superior de Educação e Comunicação

RJIES – Regime Jurídico das Instituições de Ensino Superior

USTP – Universidade de São Tomé e Príncipe

MSE – (do inglês *Mean Squared Errors*) Erro Quadrático Médio

MSR – (do inglês *Mean Squared of Regression*) Quadrados Médios de Regressão

SSE – (do inglês *Sum of Squared Errors*) Soma dos Quadrados dos Erros

SSR – (do inglês *Sum of Squared Regression*) Soma dos Quadrados de Regressão

TSS – (do inglês *Total Sum of Square*) Soma dos Quadrados Totais

VIF – (do inglês *Variance Inflaction Factor*) Factor de inflação da Variância

INTRODUÇÃO

A modelação estatística é um campo de estudo que utiliza técnicas matemáticas e estatísticas para criar representações matemáticas para estimar ou prever o comportamento de um fenómeno a partir de dados observados. No contexto do ensino superior, essa abordagem tem sido amplamente utilizada para entender os factores que influenciam a admissão de alunos e suas trajectórias académicas (Heckman et al., 2006).

A Universidade de São Tomé e Príncipe (USTP) desempenha um papel crucial no desenvolvimento do capital humano do país, sendo necessário investigar como diferentes factores socioculturais ou regionais e académicos impactam o processo de selecção de estudantes.

No país, a crescente procura por patamares superiores ao nível académico enfrenta desafios relacionados à qualidade e equidade no acesso. A utilização de modelos estatísticos pode trazer à tona variáveis críticas que, de outra forma, poderiam passar despercebidas.

A análise de dados recolhidos nos anos lectivos de 2021/2022, 2022/2023 e 2023/2024 possibilitará a modelação estatística para identificar padrões de admissão e os factores que influenciam a performance futura dos alunos. Esta abordagem poderá contribuir para a formulação de políticas que promovam a inclusão e o sucesso académico.

Com vista a evidenciar a aplicabilidade e consistência dos modelos, utilizaremos o software estatístico IBM SPSS Statistics (versão 28) e o Microsoft Excel (Office 365).

Justificativas

A escolha de investigar a admissão de estudantes na Universidade de São Tomé e Príncipe por meio de modelação estatística justifica-se pela necessidade de compreender as dinâmicas que impactam a selecção e a retenção de alunos. A educação é um factor crucial para o desenvolvimento social e económico de qualquer nação, portanto é essencial garantir que as instituições atendam às necessidades dos estudantes. Além disso, dados recentes apontam que a taxa de sucesso dos alunos no ensino superior pode ser influenciada por diversos factores, como o histórico escolar e o contexto socioeconómico, os quais exigem uma análise aprofundada (Meyer et al., 2020).

Este estudo, portanto, não apenas preenche uma lacuna no conhecimento sobre a admissão estudantil em São Tomé e Príncipe, mas também serve como base para futuras pesquisas e intervenções práticas no campo da educação ao nível do ensino superior.

Objectivo Geral

Desenvolver modelos estatísticos com vista a identificar os factores que influenciam a admissão de estudantes na Universidade de São Tomé e Príncipe, proporcionando *insights* que possam conduzir a melhorias nas práticas de selecção e a formulação de políticas mais inclusivas.

Objectivos Específicos

1. Aprofundar os modelos de regressão linear e categorial, para a seleção de modelos;
2. Identificar e caracterizar as variáveis socioeconómicas ou regionais e académicas que influenciam a admissão dos estudantes na USTP, com base em dados do ensino secundário;
3. Estabelecer relações entre o desempenho académico anterior dos candidatos e suas chances de admissão, utilizando técnicas de regressão para interpretar os dados.

Estrutura do trabalho

O presente trabalho de pesquisa está organizado em cinco secções principais, cada uma contribuindo para a compreensão do tema em análise. Na Introdução, abordamos os aspectos gerais e a relevância do tema, ressaltando alguns factores que influenciam a admissão na USTP.

No Capítulo I, fornecemos uma visão abrangente da USTP, incluindo sua história, missão, objectivos, estruturas departamentais e o processo de admissão

A terceira parte, correspondente ao Capítulo II, abordamos a metodologia empregada, detalhando os métodos estatísticos utilizados, visando assegurar a robustez e a validade dos resultados.

O Capítulo III apresenta a aplicação prática dos modelos de regressão, na qual foram testadas algumas hipóteses para a compreensão e validação de características significativas dos candidatos, acompanhada pela análise empírica dos dados e sua interpretação.

Finalmente, na quinta parte, são apresentadas as considerações finais, que sintetizam os resultados da pesquisa. Além disso, delineamos algumas limitações encontradas ao longo do trabalho, reconhecendo os potenciais vieses e restrições que podem ter impactado os resultados e propondo direcções para futuras investigações.

CAPÍTULO I – A UNIVERSIDADE DE SÃO TOMÉ E PRÍNCIPE (USTP)

1.1 História da Universidade de São Tomé e Príncipe (USTP)

A Universidade de São Tomé e Príncipe (USTP), é um ente público que presta funções delegadas pelo Estado aos cidadãos, criada através do Decreto-Lei N.º 9/2014, que goza de autonomia estatutária, cultural, científica, pedagógica, administrativa, financeira, patrimonial e disciplinar (Decreto-Lei n.º 09/2018). Ela resulta da transformação do Instituto Superior Politécnico de São Tomé e Príncipe (actual Faculdade de Ciências e das Tecnologias - FCT), da Escola de Formação de Professores e Educadores (actual Instituto Superior de Educação e Comunicação - ISEC) e do Instituto de Ciências da Saúde Victor Sá Machado (ISCSVSM), estes dois últimos, até então concentrados especialmente em formações vocacionais e técnicas avançadas com orientação profissional.

Enquanto instituição de ensino superior, a USTP orienta-se na oferta de formações científicas robustas, reunindo esforços e competências das suas unidades de ensino e investigação. À luz deste panorama, busca-se fomentar a produção, disseminação e promoção de conhecimento, ciência, tecnologia e cultura, articulando o ensino e a investigação, tendo como paradigma potenciar o desenvolvimento humano e a melhoria dos padrões de qualidade do ensino e da investigação. A Universidade de São Tomé e Príncipe aspira transformar-se numa universidade de referência nacional.

Os princípios orientadores da USTP são: (Decreto-Lei n.º 09/2018)

a) A liberdade;
b) A excelência;
c) A qualidade;
d) O empreendedorismo
e) A sustentabilidade;
f) A internacionalidade.

Referente a investigação, a USTP tem um Centro de Investigação, com unidades de investigação associadas, e tem estado a realizar investigações em parcerias com outras universidades internacionais.

Concernente à Internacionalização, a USTP considera que esta corresponde a uma ferramenta estratégica e, enquanto tal, deve estar presente de forma transversal no funcionamento da instituição. É uma ferramenta estratégica porque a sua definição condiciona e caracteriza o modelo de gestão de qualquer organização. A sua transversalidade resulta de facto de não podermos ambicionar ser uma organização global.

No Plano Estratégico do quadriénio 2019-2023, a USTP está delineada como uma instituição de ensino superior abrangente, que engloba uma ampla gama de áreas científicas e de formação, incluindo ciências sociais e humanas, ciências da vida, ciências exatas, entre outras. Além disso, é uma universidade centrada na pesquisa, reconhecendo a produção de conhecimento científico como um pilar essencial para o cumprimento da sua missão. Esta abordagem é justificada não apenas pela inerente vocação da instituição universitária, mas também pelo papel crucial do conhecimento inovador na fundamentação do ensino e interacção com a sociedade. (Costa, P. S., 2019)

A USTP também se apresenta como um espaço de educação integral, comprometido com a formação holística de todos os seus membros e adoptando uma perspectiva educativa que não é unidimensional, mas sim sensível ao desenvolvimento do sujeito humano em diversas dimensões cognitivas, morais, éticas, relacionais e físicas. Além disso, a universidade enfatiza a interacção com a sociedade, mantendo um diálogo aberto com os diversos actores económicos, culturais e sociais, sem comprometer sua identidade institucional, e valorizando a capacidade de resposta aos desafios propostos por tais actores. (Costa, P. S., 2019)

A USTP prossegue ainda:

(i) a fomentação de actividades de investigação;

(ii) a prestação de serviços diversificados à comunidade, numa perspectiva de valorização recíproca;

(iii) o intercâmbio científico, técnico e cultural com instituições de investigação e de ensino superior, nacionais e estrangeiras;

(iv) a cooperação internacional e aproximação entre os povos;

A gestão da USTP caracteriza-se por ser unificada, participativa e colegiada, conforme estabelecido nos artigos 93.º, 95.º, 103.º e 107.º do Regime Jurídico das Instituições de Ensino Superior (RJIES). A estrutura de governança inclui o Conselho da Universidade, o Conselho de Gestão e Estratégia, o Conselho para a Qualidade, o Conselho Científico e o Conselho Pedagógico.

Os cursos são organizados de acordo com um sistema de créditos curriculares aplicáveis a cada ciclo de estudos. A estrutura curricular baseia-se na definição do número de horas de contato e de trabalho autônomo necessárias para que um estudante possa concluir cada unidade curricular, semestre ou curso, conforme estabelecido pelo Decreto-Lei n.º 25/2020.

A USTP tem a competência para conceder graus, diplomas e títulos académicos e honoríficos, incluindo os graus de licenciado, mestre e doutor, conforme o artigo 13.º do Decreto-Lei n.º 09/2018. Além disso, a universidade oferece certificados de cursos não conferentes de grau, conforme previsto na legislação vigente. Atualmente, a USTP concede os graus de licenciado e mestre.

A USTP oferece ensino e organiza a pesquisa em grandes áreas científicas, de acordo com seus departamentos, incluindo:

- Ciências da Natureza, da Vida e do Ambiente;
- Ciências Exactas, Tecnologias e Engenharias;
- Ciências Económicas e Empresariais;
- Línguas e Cultura;
- Ciências Sociais, Humanas e Artes;
- Ciências de Educação;

A Universidade de São Tomé e Príncipe (USTP) oferece mais de 27 cursos, incluindo programas de mestrado em colaboração com diversas universidades estrangeiras. Entre os cursos de graduação mais frequentemente oferecidos estão: Licenciatura em Matemática, História, Língua Portuguesa, Gestão de Empresas, Economia, Ciências da Educação, Educação Infantil, Educação Básica, Ciências da Comunicação, Biologia e Enfermagem.

1.2 Admissão dos Estudantes na USTP

A admissão de estudantes na Universidade de São Tomé e Príncipe (USTP) fundamenta-se no critério essencial de apresentação de um comprovativo de conclusão do ensino secundário ou de um equivalente legal.

O processo de candidatura inicia-se com a divulgação dos cursos disponíveis para o ano lectivo em questão, seguido pelo período de inscrição dos candidatos, que podem ser tanto nacionais quanto internacionais. A selecção dos candidatos é realizada com base nos resultados dos exames de admissão ou pela média ponderada das disciplinas chave para cada curso, conforme previamente estabelecido por uma comissão designada e aprovado pelo conselho universitário.

Em situações onde há insuficiência de candidatos para preencher as vagas disponíveis nos cursos ou quando a realização dos exames de admissão não é viável, como ocorreu nos anos letivos de 2020/2021 e 2021/2022 devido à pandemia de COVID-19 provocada pelo vírus SARS-CoV-2, a seleção dos candidatos é efetuada exclusivamente com base na média ponderada das disciplinas chave.

CAPÍTULO II – MÉTODOS ESTATÍSTICOS

2.1 Testes Paramétricos e Não Paramétricos

Na inferência estatística, tem-se a necessidade de recorrer aos testes hipóteses ou intervalos de confiança para validar ou não, determinadas suspeitas ou afirmações sobre os parâmetros de uma população.

A essência dos testes de hipóteses está em decidir sobre as hipóteses. Porém, em qualquer uma destas duas decisões, existe o risco de estar a tomar uma decisão errada. (Reis, E. *et al*, 2019)

Os mesmos autores referem que este risco está associado aos tipos de erros que podem ser cometidos. Tem-se o *Erro do tipo I* – quando se decide rejeitar a H_0 quando ela é verdadeira e o *Erro do tipo II* – quando se decide não rejeitar a H_0 quando ela é falsa.

Para aplicação do teste de hipóteses, há que se ter em conta *"o tipo de população, o conhecimento da respetiva variância e a dimensão da amostra"* (Reis, E. *et al*, 2019) de forma a ser utilizada a estatística adequada e em função da distribuição amostral. A partir desta último (distribuição amostral) conjugada com a dimensão da amostra, o teste de hipótese subdivide-se em paramétrico – se a população for normal (seguir uma distribuição normal) e em não paramétrico – se a população não for normal (não seguir uma distribuição normal). (Fávero, L. P., & Belfiore, P., 2017).

Segundo Hill, M. Hill, A. (2020) as técnicas/testes paramétricas *"são estatísticas que lidam com parâmetros"* e *"(...) assumem o prossuposto forte que, (...), os valores das variáveis de uma amostra têm uma distribuição normal"*. Além disso, os autores destacam que *"os valores de uma variável são medidos numa escala de intervalo ou rácio"*. Fávero, L. P., & Belfiore, P. (2017) corroboram com esta perspectiva, salientando que "*observações devem ser independentes, as populações devem ter variâncias iguais para testes de comparação de duas médias populacionais emparelhadas ou k médias populacionais (*$k > 3$*)*".

Na inobservância do cumprimento do prossupostos para aplicação do teste paramétrico, pode-se recorrer ao teste não paramétrico onde, não se preocupam com os parâmetros nem assumem que os valores das variáveis seguem uma distribuição normal. Nos testes não paramétricos, as hipóteses são formuladas sobre as características da população numa das seguintes situações: i) variável em estudo ser qualitativa (numa escala nominal ou ordinal) ou ii) quando a função distribuição da variável aleatória que produz os dados não está especificado ou iii) quando

existe um número suficientemente grande (infinito) de parâmetros desconhecidos. Uma das vantagens destes testes é que *"(...) permitem analisar variáveis com valores numa escala ordinal ou numa escala nominal."* (Hill, M. Hill, A., 2020). Porém, estas estatísticas não permitem extrapolar para uma população, considerações acerca de parâmetros importantes como a média, desvio-padrão, proporção, entre outros.

Os testes não paramétricos podem ser utilizados para verificar a normalidade dos dados, independência das populações ou igualdade entre as distribuições e são menos exigentes aos pressupostos. Estes testes são recomendados quando temos amostras de pequena dimensão. (Fávero, L. P., & Belfiore, P. (2017)

2.1.1. Pressupostos de aplicabilidade para a comparação de dois ou mais grupos independentes

Conforme referido anteriormente, a aplicação do teste não paramétricos não exige a pressuposição de uma distribuição amostral (como o caso da distribuição normal), o que o torna mais flexível em certos casos, com as devidas limitações e adaptações.

Porém, ao contrário deste, para que seja aplicado um teste paramétrico, devem ser verificados e confirmados, em simultâneo, alguns pressupostos que validam a aplicação do mesmo, como alerta Marôco, J. (2021). Destes pressupostos, tem-se: *i)* que a variável dependente possua a distribuição normal em cada grupo (população), e *ii)* a homocedasticidade ou homogeneidade das variâncias populacionais.

- **Normalidade Univariada**

Para verificar se a variável dependente provém de uma distribuição normal, recomenda-se o teste de Kolmogorov-Smirnov. (Marôco, J., 2021)

Este teste tem como finalidade decidir se a distribuição da variável dependente numa determinada amostra provém de uma distribuição específica, aqui subentendida como a normal.

Assim, a hipótese de teste apresenta a seguinte característica:

$$H_0: F_Y = N(\mu, \sigma^2) - \text{Provem de uma Distribuição Normal}$$

$$H_a: F_Y \neq N(\mu, \sigma^2) - \text{Não Provem de uma Distribuição Normal}$$

A estatística de teste associado a este teste é dada por:

$$D = \max\{\max(|F(y_i) - F_0(y_i)|) \ ; \ \max(|F(y_{i-1}) - F_0(y_i)|)\}$$

em que $F_0(Y) \sim N(\mu, \sigma^2)$. Onde $F(y_i)$ representa a distribuição da variável dependente em estudo e $F_0(y_i)$ da distribuição específica. Rejeita-se H_0 para $valor\text{-}p$[1] $< \alpha$, onde α é o nível de significância. Neste trabalho de pesquisa, será considerado $\alpha = 0{,}05$, salvo se indicar o contrário.

Ora, a rigor, a estatística D não pode ser aplicada se trabalhar com as estimativas, desconhecendo as informações sobre os verdadeiros valores dos parâmetros. Em alternativa, Maroco, J. (2021) recomenda ou alerta para a correção proposta por H. Lilliefors às tabelas dos valores críticos da distribuição de Kolmogorov-Smirnov onde, de igual modo, rejeita-se H_0 para $valor\text{-}p < \alpha$.

Em alternativa ao teste de Kolmogorov-Smirnov, tem-se o teste de Shapiro-Wilk que é apropriado para amostras pequenas, geralmente menores de 30 ($n < 30$).

Gozando das mesmas hipóteses do teste anterior (teste de Kolmogorov-Smirnov), a estatística de teste é dada por:

$$W = \frac{(\sum_{i=1}^{n} a_i Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

onde os Y_i são os valores da variável dependente ordenados por ordem crescente, $\bar{Y}$ a média e a_i são constantes geradas a partir da média, variância e covariância de n ordens com a distribuição normal $N(0, 1)$. De salientar que Marôco, J. (2021) realça dizendo que a_i são tabelados (Marôco, J. 2021, apud Pearson & Hartley, 1972).

- **Homogeneidade de variâncias**

Tomando como exemplo a análise de regressão múltipla, onde trabalha-se com uma variável dependente e mais de uma variável independente, Marôco, J. (2021,) aponta a violação do pressuposto da homogeneidade de variância (homocedasticidade) como a mais grave em relação a normalidade, pois a mesma afecta tanto o erro do tipo I como o erro do tipo II.

Dentre os testes para verificar a homocedasticidade tem-se o teste de Levene que é robusto a desvios de Normalidade, o que o torna mais potente. Já o teste de Bartlett, que se preocupa em investigar o nível de significância das diferenças entre as variâncias das k populações, é aplicado quando os grupos ou k populações seguem uma distribuição normal e com dimensões

[1] $valor\text{-}p$ (do inglês $p\text{-}value$) é o índice da evidência indutiva contra a hipótese nula. (Marôco, J. 2021)

superiores a 6 (Reis, E. et al, 1999). Acrescenta-se ainda de que o teste de Bartlett é muito sensível à hipótese de normalidade. Além destes testes, tem-se o teste $F_{máx}$ de Hartley – exige que os números de repetições em diferentes tratamentos[2] ou variáveis independentes sejam iguais e o teste de C de Cochran – não exige números iguais de repetições, porém exige que os dados apresentem uma distribuição Normal.

Assim, a hipótese de teste apresenta a seguinte característica:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$$

$$H_a: \exists\, i,j: \sigma_i^2 = \sigma_j^2 \quad (i \neq j; i,j = 1, \ldots, k)$$

A estatística de teste associado ao teste de Levene é dada por:

$$W = \frac{N-k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i (\bar{Z}_i - \bar{Z})}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{Z}_{ij} - \bar{Z}\right)^2} \sigma_j^2$$

onde N é a dimensão da amostra global, n_i é a dimensão de cada uma das k amostras. A variável Z é definida como $Z_{ij} = \left|Y_{ij} - \bar{Y}_i\right|$ em que Y_{ij} é a observação de j da amostra i e $\bar{Y}_i$ é a média da amostra i. $\bar{Z}_i$ é a média de Z_{ij} na amostra i e $\bar{Z}$ é a média de Z_{ij} na amostra global.

Ora, Marôco, J. (2021) acrescenta ainda dizendo que se existirem suspeitas de que $Y \nsim N(\mu, \sigma^2)$ devemos então, considerar $Z_{ij} = \left|Y_{ij} - \tilde{Y}_i\right|$ onde e $\tilde{Y}_i$ é a mediana da amostra i. A hipótese nula é rejeitada para $valor\text{-}p < \alpha$.

2.1.2. Testes para k Amostras Independentes

As estatísticas do teste *t*-Student só devem ser utilizadas para comparar médias de duas, e apenas a duas, populações das quais foram extraídas as amostras, uma vez que, na comparação de mais de duas populações, ainda que seja duas a duas, a probabilidade de erro do tipo I será de $1 - (1-\alpha)^k \times 100\%$, onde k é o número de populações (amostras). (Marôco, J., 2021)

- **Análise de Variância (ANOVA – *Analysis of Variance*)**

[2] Tratamento é um termo muito utilizado no delineamento experimental. Um tratamento é uma condição imposta ou objeto que se deseja medir ou avaliar em um experimento. Cada tratamento corresponde a uma categoria do fator ou variável independente.

Para suprir esta limitação do teste *t*-Student, Sir Ronald Fisher propôs uma metodologia conhecida com **Análise de Variância** (Marôco, J., 2021) que consiste em comparar médias de três ou mais populações, por meio da análise de variâncias amostrais.

O teste se baseia em uma amostra extraída de cada população, com o intuito de determinar se as diferenças entre as médias amostrais sugerem diferenças significativas entre as médias populacionais, através da comparação entre a variância dentro das amostras ou grupos (variância residual) com a variância entre as amostras ou grupos (variância do factor). (Fávero, L. P., & Belfiore, P., 2017)

A aplicação da ANOVA em amostras independentes pressupõe que:

- os dados nas populações devem apresentar distribuição normal;
- as variâncias populacionais devem ser homogêneas.

Porém, de acordo com Pestana, M. H., & Gageiro, J. N. (2020), o pressuposto da normalidade não é exigido quando se trata de amostras de grande dimensão, pelo teorema do limite central, nem o pressuposto da homogeneidade, quando as dimensões dos grupos são iguais ou semelhantes (dimensão é semelhante se o quociente em a maior e menor dimensão for igual ou inferior a 1,96, para $\alpha = 0{,}05$)

A ANOVA subdivide-se em ***i)* ANOVA de um factor**, conhecida em inglês como *One-Way* ANOVA, que é a extensão do teste *t* de Student para duas médias populacionais, o que permite ao investigador a comparação de três ou mais médias populacionais, que consiste em verificar o efeito de uma variável explicativa de natureza qualitativa (fator) em uma variável dependente de natureza quantitativa. (Fávero, L. P., & Belfiore, P., 2017)

A hipótese nula do teste afirma que as médias populacionais são iguais; se existir pelo menos um grupo com média diferente dos demais, a hipótese nula é rejeitada.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k \qquad vs \qquad H_a: \exists_{i,j}:\ \mu_i \neq \mu_j \quad (i \neq j;\quad i,j = 1,2,\dots,k)$$

Segundo Marôco, J. (2021) estás hipóteses são equivalentes a testar se o factor sobre o estudo teve, ou não, um efeito significativo sobre a variável dependente, o que pode testar como:

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \qquad vs \qquad H_a: \exists_i:\ \alpha_i \neq 0 \qquad (i = 1,2,\dots,k)$$

ou seja, que os efeitos do tratamento (níveis do factor) são iguais e nulos entre todas as amostras, H_0 ou, que existe pelo menos um efeito não nulo, H_1.

O modelo teórico da ANOVA a um factor, partir das observações amostrais é dado por:

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

Onde y_{ij} são as observações, $\bar{y}$ representa a média geral amostral (estimativa de μ), $(\bar{y}_i - \bar{y})$ representa o efeito do tratamento (estimativas de α_i) e $(y_{ij} - \bar{y}_i)$ representa os resíduos (estimativas de ε_{ij}).

As hipóteses analisam-se pelo teste F de Snedecor, que é o quociente entre a variação entre os grupos (comprara a média de cada grupo com a global) e a variação dentro do grupo (comprara o desvio de cada observação à sua média). Os cálculos são resumidos na seguinte tabela, como fazem maioria dos *softwares*: (Marôco, J., 2021)

Tabela 1 – Tabela de ANOVA a um Factor *(One-Way ANOVA)*

Fonte de Variação	Soma dos Quadrados	Graus de Liberdades	Quadrados Médios	*F*	Probabilidade
Factor (entre as amostras)	$SSF = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	$k - 1$	$MSF = \frac{SSF}{k-1}$	$\frac{MSF}{MSE}$	$valor\text{-}p$
Residual (dentro das amostras)	$SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2$	$N - k$	$MSE = \frac{SSE}{n-k}$		
Total	$SST = \sum_{i=1}^{k} (N_i - 1)s^2$	$N - 1$			

onde $\sum_{i=1}^{k} n_i = N$, s_i^2 é o estimador da variância da amostra i e $s^2 = MSE$ é o estimador da variância total. A hipótese nula é rejeitada para $valor\text{-}p < \alpha$.

Ora, quando a variável dependente tem distribuição normal, mas heterogénea Welch (1951) e Brown & Forsythe (1974a, 1974b) propuseram deduções alternativas ao teste F conhecidas como teste F de Welch - F_W e teste F de Brown & Forsythe - F_{BF} (ver Marôco, J., 2021) e Pestana, M. H., & Gageiro, J. N., 2020), onde Lix *et al.*, (1996 apud Marôco, J., 2021) chegou a perceber que o teste F de Welch - F_W apresenta a melhor *performance*.

E em ***ii)*** **ANOVA Factorial** que é a extensão ANOVA de um factor, assumindo os mesmos pressupostos, porém considerando pelo menos dois factores. (Fávero, L. P., & Belfiore, P., 2017)

O autor acrescenta ainda que o objetivo da ANOVA fatorial é determinar se as médias para cada nível do fator são iguais (efeito isolado dos fatores na variável dependente) e verificar a interação entre os fatores (efeito conjunto dos fatores na variável dependente).

Para fins didáticos, descreveremos **ANOVA Factorial de dois factores** (*Two-Way* ANOVA) que, a partir das observações amostrais, é dado por:

$$y_{ijr} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijr}$$

Onde y_{ijr} são as observações, μ é a estimativa que representa a média geral amostral ($\bar{y}$), α_i é a estimativa que representa o efeito do primeiro factor $(\bar{y}_i - \bar{y})$, β_j é a estimativa que representa o efeito do segundo factor $(\bar{y}_j - \bar{y})$, $\alpha\beta_{ij}$ é a estimativa que representa as iterações entres os factores $(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})$ e ε_{ijr} é a estimativa que representa os resíduos $(y_{ijr} - \bar{y}_i)$.

As hipóteses para testar se as médias de para cada nível de cada um dos factores (F_1 e F_2) são ou não iguais, são dadas por: (Marôco, J., 2021)

1. Para conhecer os efeitos isolados de cada factor

$$H_0^{F_1}: \mu_1 = \mu_2 = \cdots = \mu_a \quad vs \quad H_a^{F_1}: \exists_{i,j}: \mu_i \neq \mu_j \quad (i \neq j; \quad i,j = 1,2,\ldots,a)$$

$$H_0^{F_2}: \mu_1 = \mu_2 = \cdots = \mu_b \quad vs \quad H_a^{F_2}: \exists_{i,j}: \mu_i \neq \mu_j \quad (i \neq j; \quad i,j = 1,2,\ldots,b)$$

2. Para conhecer os efeitos dos dois factores em conjunto (interação)

$$\begin{cases} H_0^{F_1F_2}: \gamma = 0 \text{ (Não existe interação)} \\ H_a^{F_1F_2}: \gamma \neq 0 \text{ (Existe interação)} \end{cases} \qquad (i = 1,2,\ldots,a\,; \;\; j = 1,2,\ldots,b)$$

A hipótese nula é rejeitada para $valor\text{-}p < \alpha$.

Os cálculos são resumidos na seguinte tabela, como fazem maioria dos *softwares*: (Marôco, J., 2021)

Tabela 2 – Tabela de ANOVA Factorial (*Two-Way ANOVA)*

Fonte de Variação	Soma dos Quadrados	Graus de Liberdades	Quadrados Médios	*F*	Probabilidade
Factor 1	$SSF_1 = b \cdot r \cdot (a-1)s_{\bar{y}_i}^2$	$a-1$	$MSF_1 = \frac{SSF_1}{a-1}$	$\frac{MSF_1}{MSE}$	$valor\text{-}p$

Factor 2	$SSF_2 = a \cdot r \cdot (b-1)s_j^2$	$b-1$	$\frac{MSF_2}{} = \frac{SSF_2}{b-1}$	$\frac{MSF_2}{MSE}$	$valor\text{-}p$
Interação	$SSF_1F_2 = r \cdot \sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta_{ij})^2 \cdot s_{ij}^2$	$(a-1)(b-1)$	$\frac{SSF_1F_2}{(a-1)(b-1)}$	$\frac{MSF_1F_2}{MSE}$	$valor\text{-}p$
Erro	$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-1)s_{ij}^2$	$(r-1)ab$	$\frac{SSE}{(r-1)ab}$		
Total	$SST = \sum_{i=1}^{k}(N_i-1)s^2$	$N-1$			

Caso fosse **ANOVA Factorial de três factores** (*Three-Way* ANOVA), o modelo teórico seria dado por:

$$y_{ijkr} = \mu + \alpha_i + \beta_j + \delta_k + \alpha\beta_{ij} + \alpha\delta_{ik} + \beta\delta_{jk} + \alpha\beta\delta_{ijk} + \varepsilon_{ijkr}$$

Quando os testes *F* de Snedecor ou *F* de Welsh são significativos e o fator é de natureza nominal, podem usar-se para comparação de médias os testes a posteriori ou *Post-hoc*. Porém, os testes *F* podem rejeitar H_0 apesar dos *Post-hoc* não detetarem diferenças entre os grupos. (Pestana, M. H., & Gageiro, J. N., 2020)

Nos testes *Post-hoc* comparam-se as respostas entre todos os pares de grupos, com região crítica bilateral, com a exceção do teste Dunnett, que pode ser também um teste unilateral onde se compara o grupo de controle com o grupo experimental.

Reis, E., *et. al*. (2018) acrescenta que, estes testes diferem no modo como analisam as diferenças de médias e ainda no método de controlo do nível de significância, onde os mais utilizados são o teste HSD de Tukey e o teste de Scheffé.

As preferências pelo último justificam-se por várias razões, com destaque para: a sua maior simplicidade de cálculo; o facto de permitir a utilização de amostras com diferentes dimensões e ainda por ser um método robusto no respeitante aos pressupostos de normalidade e igualdade de variâncias das populações.

No entanto, quando os grupos amostrais têm idêntica dimensão, o método HSD de Tukey é mais preciso pois gera intervalos de confiança com menor amplitude. Por sua vez o método de Scheffé tende a ser mais conservativo, ou seja, nas mesmas condições, tem uma maior probabilidade de não rejeitar a hipótese nula quando ela é verdadeira.

Os testes *Post-hoc* procuram testar qual ou quais são os pares de médias significativamente diferentes, ou seja, que apresentam $valor\text{-}p < \alpha$. Assim, as hipóteses de teste são:

$$H_0: \mu_i = \mu_j \qquad vs \qquad H_a: \mu_i \neq \mu_j \quad (i, j = 1,2, \dots, k)$$

No caso do teste HSD de Tukey e o teste de Scheffé, que apresentam cálculos simples, têm H_0 rejeitada quando: (Reis, E., *et. al*., 2018)

- No teste HSD (*Honestly Significant Difference*) de Tukey

$$\left|\bar{X}_i - \bar{X}_j\right| \geq S_{T_{(1-\alpha)}} \sqrt{\frac{MSE}{2} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

onde S_T é a distribuição da "Studentized Range" com $(k,\ n-k)$ graus de liberdade.

- No teste de Scheffé

$$\left|\bar{X}_i - \bar{X}_j\right| \geq \sqrt{(k-1) \times F_{(1-\alpha)} \times MSE \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

onde $F_{(1-\alpha)}$ é a distribuição F de Snedecor com $(k-1,\ n-k)$ graus de liberdade.

- **Teste de Kruskall-Wallis**

À semelhança do teste *t*-Student, que pode ser generalizado para realizar análise para mais de dois grupos através da ANOVA, o teste de Mann-Whitney pode ser generalizado para mais de dois grupos através do **Kruskall-Wallis**.

O teste de Kruskal-Wallis é uma alternativa à ANOVA de um factor quando as hipóteses de normalidade dos dados e igualdade das variâncias forem violadas, quando o tamanho da amostra for pequeno, ou ainda quando a variável for medida ao nível pelo menos ordinal e tem por objetivo verificar se k amostras independentes $(k > 2)$ são provenientes da mesma população ou se provem de populações com mesma distribuição. (Marôco, J. (2021), Pestana, M. H., & Gageiro, J. N. (2020), Fávero, L. P., & Belfiore, P. (2017)).

Marôco, J. (2021) e Pestana, M. H., & Gageiro, J. N. (2020), consideram o teste de Kruskal-Wallis como uma alternativa não paramétrica à ANOVA.

As hipóteses do teste são:

$$\begin{cases} H_0: \theta_1 = \theta_2 = \cdots = \theta_k \\ \quad \text{(As medianas são iguais)} \\ \\ H_a: \exists_{i,j}:\ \theta_i \neq \theta_j \qquad\qquad (i \neq j = 1,2,\ldots,k) \\ \quad \text{(Existe pelo menos um par de medianas diferentes)} \end{cases}$$

Porém, Marôco, J. (2021) realça que definir as hipóteses por recurso as medianas é algo abusivo visto que, se as distribuições populacionais forem iguais, então as suas medianas são iguais, contudo o recíproco pode não ser válido.

A estatística de teste é dada por:

$$T = \frac{12}{n \times (n+1)} \times \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3 \cdot n \times (n+1)$$

Caso haja empates

$$T_E = \frac{T}{1 - \frac{\sum_{i=1}^{g} (t_i^3 - t_i)}{n^3 - n}}$$

onde $n = \sum_{j=1}^{k} n_j$ é a soma do total das observações de cada grupo, $R_j = \sum_{i=1}^{n_j} r_{ij}$ representa a soma das ordens de cada uma das j $(j = 1,2,\ldots,k)$ amostras, g representa o número de grupos de empates e t representa o número de observações em cada grupo de empates.

A estatística de teste Kruskal-Wallis segue uma distribuição Qui-quadrado com $(k-1)$ grau de liberdade $\left(T \sim \chi^2_{(1-\alpha,\ k-1)}\right)$ e rejeita-se H_0 para $valor\text{-}p < \alpha$.

A semelhança da ANOVA, em caso de rejeição do H_0 no teste de Kruskal-Wallis, recorre-se ao Teste LSD (*Least Significant Difference*) de Fisher aplicada às ordens ou teste Q de Cochran (para mais de 10 comparações), rejeitando H_0 para $valor\text{-}p < \alpha$.

2.2 Medidas de Associação – Correlação

Na maioria dos estudos estatísticos, é comum verificar se existe alguma associação entre as variáveis e quantificar a intensidade e direção entre as mesmas.

Para estudar a associação entre as variáveis, pode recorrer-se à Análise de Correlação que é um dos métodos estatísticos muito utilizado para medir o grau de associação entre as variáveis bem como a sua direção, ou seja, indicando como variam conjuntamente. A Correlação pode ser bivariada – quando se tem duas variáveis, ou multivariadas – quando se tem mais de duas variáveis. (Marôco, J., 2021)

Nesta análise, não constitui uma necessidade a distinção entre a variável explicativa e a variável resposta, ou seja, a variação conjunta entre X_1 e X_2 é igual entre X_2 e X_1.

As medidas de associações mais frequentes são:

- **Coeficiente de Correlação de Pearson**

Este coeficiente mede a intensidade e a direção da associação de tipo linear entre duas variáveis quantitativas e trata-se de um coeficiente paramétrico, ou seja, as duas variáveis devem seguir uma distribuição normal.

Este coeficiente é calculado a partir da razão entre a variância comum (Covariância – Cov ou $S_{X_1X_2}$) de duas varáveis pelo produto dos desvios-padrão de cada uma dessas variáveis.

$$r = R_{X_1X_2} = \frac{S_{X_1X_2}}{S_{X_1}S_{X_2}} = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)\,(X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}\sqrt{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}}$$

- **Coeficiente de Correlação de Spearman**

O coeficiente de Correlação de Spearman, também conhecido como Ró de Spearman é uma medida de associação não paramétrica entre duas variáveis medidas na escala ordinal ou métrica. Este coeficiente usa as ordens (r_1 e r_2) em vez dos valores de cada observação.

O Ró de Spearman é insensível tanto a assimetrias na distribuição como à presença de *outliers*, não carecendo da normalidade dos dados, o que o caracteriza como não paramétrico.

$$\rho_s = R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad \text{onde} \qquad d_i^2 = \left(r_{1_i} - r_{2_i}\right)^2$$

Porém, na existência de observações iguais ou empates, Marôco, J. (2021) orienta para a seguinte fórmula.

$$\rho_s = R_s = \frac{(n^3 - n) - 6\sum_{i=1}^{n} d_i^2 - \frac{T_{X_1} + T_{X_2}}{2}}{\sqrt{(n^3 - n)^2 - \left(T_{X_1} + T_{X_2}\right)(n^3 - n) + T_{X_1} T_{X_2}}} \quad \text{onde} \quad T_{X_i} = \sum_{j=1}^{g_i} \left(t_j^3 - t_j\right)$$

onde T_{X_i} é o factor de correcção de empates da variável X_i, g_i é o número de grupos de observações empatadas na variável X_i e t_j é o número de grupos de observações empatadas em cada grupo de empates da variável X_i.

Os critérios de classificações ou variação destes coeficientes de correlação (ecoeficiente de correlação de Pearson e Spearman) podem ser agrupados de acordo com a seguinte tabela, (Pestana, M. H., & Gageiro, J. N., 2020):

Tabela 3 – Classificação do Coeficiente de Correlação de *Pearson e Spearman*

	Nula	**Fraca**	**Moderada**	**Forte**	**Perfeita**
Correlação (em valor absoluto)	0	$< 0{,}40$	$\geq 0{,}4$ e $< 0{,}7$	$\geq 0{,}7$ e < 1	1

De salientar que os valores dos coeficientes podem variar no sentido positivo ou no sentido negativo, de acordo com o sinal do coeficiente.

2.3 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados, surge de forma consolidada, do ponto de vista teórico, a partir dos trabalhos de Nelder e Wdderburn (1967, apud Fávero, L. P., & Belfiore, P., 2017) fornecendo de tal forma, uma abordagem unificada para modelação de diversos tipos de variáveis resposta.

O autor acrescenta ainda que, os **Modelos Lineares Generalizados** "*representam um grupo de modelos de regressão lineares e exponenciais não lineares, em que a variável dependente possui, por exemplo, distribuição normal, Bernoulli, binomial, Poisson ou Poisson-Gama*".

Segundo Myers, R. H., Montgomery, D. C., Vining, G. G. & Robinson, T. J. (2012) todo modelo linear generalizado contém três componentes, designadamente *i)* **uma distribuição de variável de resposta** (*Componente Aleatório* do modelo, às vezes chamada de estrutura de erro), *ii)* **um preditor linear** que envolve as variáveis preditoras ou co-variáveis (*Componente Sistemática* do modelo) e *iii)* uma **função de ligação** que conecta o preditor linear à média natural da variável resposta.

Assim, um Modelo Linear Generalizado pode ser definido da seguinte forma: (Fávero, L. P., & Belfiore, P., 2017)

$$\boldsymbol{\eta_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj}}$$

onde $\boldsymbol{\eta}$ é a função de ligação canónica, $\boldsymbol{\beta_0}$ representa a constante, $\boldsymbol{\beta_i}$ $(\boldsymbol{i = 1, 2, 3, \ldots, p})$ são os coeficientes de cada variável explicativa e correspondem aos parâmetros a serem estimados, $\boldsymbol{X_i}$ são as variáveis explicativas (métricas ou *dummies*) e os subscritos $\boldsymbol{j}$ representam cada uma das observações da amostra em análise ($\boldsymbol{j = 1, 2, 3, \ldots, n}$, em que $\boldsymbol{n}$ é o tamanho da amostra).

São casos particulares dos Modelos Lineares Generalizados, em função da característica da variável dependente, a sua distribuição e a respectiva função de ligação canónica, os seguintes modelos:

i) Modelos de Regressão Lineares

$$\hat{Y}_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj}$$

em que $\hat{Y}$ é o valor esperado da variável dependente Y, de natureza quantitativa e com distribuição normal.

ii) Modelos Regressão com Transformação de Box-Cox

$$\frac{\hat{Y}_j^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj}$$

em que λ é o parâmetro da transformação de Box-Cox que maximiza a aderência à normalidade da distribuição da nova variável gerada a partir da variável Y original.

iii) Modelos de Regressão Logística Binária e Multinomial

$$\ln\left(\frac{p_j}{1 - p_j}\right) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj}$$

em que p_j é a probabilidade de ocorrência do evento de interesse definido por $Y = 1$, dado que a variável dependente Y é *dummy*.

iv) Modelos de Regressão Logística Multinomial

$$\ln\left(\frac{p_{j_m}}{1 - p_{j_m}}\right) = \beta_{0m} + \beta_{1m} X_{1j} + \beta_{2m} X_{2j} + \beta_{3m} X_{3j} + \cdots + \beta_{pm} X_{pj}$$

em que p_{j_m} $(m = 1, 2, 3, \ldots, M - 1)$é a probabilidade de ocorrência de cada uma das M categorias da variável dependente Y.

v) Modelos de Regressão *Poisson* para Dados de Contagem.

$$\ln(\lambda_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj}$$

em que λ é o valor esperado da quantidade de ocorrências do fenômeno representado pela variável dependente Y, que apresenta dados de contagem com distribuição *Poisson*.

vi) Modelos de Regressão Binomial Negativo para Dados de Contagem.

$$\ln(\mu_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj}$$

em que μ é o valor esperado da quantidade de ocorrências do fenômeno representado pela variável dependente Y, que apresenta dados de contagem com distribuição *Poisson-Gama*.

2.3.1 Estimação de parâmetros dos MLG

Segundo Myers, R. H., Montgomery, D. C., Vining, G. G. & Robinson, T. J. (2012), o modelo linear generalizado nos permite ajustar modelos de regressão para dados de resposta que seguem uma distribuição muito geral chamada família exponencial, que inclui as distribuições normais, binomial, *Poisson*, geométrico, binomial negativo, exponencial, gama e normal inversa.

A família exponencial é uma classe geral de distribuições que inclui muitas distribuições conhecidas como casos especiais. Sua função de densidade de probabilidade para uma resposta observada y que pode ser expressa na forma:

$$f(y;\ \theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

onde $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são algumas funções (específicas). O parâmetro θ é o parâmetro de localização da distribuição. A função $a(\phi)$ é geralmente da forma $a(\phi) = \theta \cdot \omega$, onde ω é uma constante conhecida.

A estimação dos parâmetros de modelos lineares generalizados é frequentemente feita usando o método da Máxima Verossimilhança. Contudo, e po inclui equações não lineares, recorre-se aos métodos iterativos de Newton-Raphson ou Escore de Fisher para a obtenção de solução analítica. (Agresti, A., 2012).

Assim, para Myers, R. H., Montgomery, D. C., Vining, G. G. & Robinson, T. J. (2012), as estimativas são os valores de parâmetros que maximizam o log-verosimilhança, dada por:

$$\mathcal{L} = \ln L\,(\beta; y) = \sum_{j=1}^{n}\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

Como $\eta_j = g(\mu_j) = \mathrm{x}_j'\beta$, então

$$\frac{\partial \mathcal{L}}{\partial \beta} = \frac{\partial \mathcal{L}}{\partial \theta_j}\,\frac{\partial \theta_j}{\partial \beta} = \sum_{j=1}^{n}\left[\frac{1}{a(\phi)}\left(y_j - \frac{\partial b(\theta)}{\partial \theta_j}\right)\mathrm{x}_j\right] = \sum_{j=1}^{n}\left[\frac{1}{a(\phi)}(y_j - \mu_j)\mathrm{x}_j\right]$$

Portanto, as estimativas de máxima verossimilhança dos parâmetros são obtidas a partir do seguinte sistema de equações para β:

$$\sum_{j=1}^{n}\left[\frac{1}{a(\phi)}(y_j - \mu_j)\mathrm{x}_j\right] = 0$$

E como em muitos casos $a(\phi)$ é uma constante, a equação teria o seguinte formato

$$\sum_{j=1}^{n}\left[(y_j - \mu_j)\mathrm{x}_j\right]$$

Na forma matricial, seria

$$X'(y - \mu) = 0$$

O autor acrescenta ainda que, se as suposições do modelo, incluindo a escolha da função de ligação, estiverem corretas, então, assintoticamente o valor esperado é dado por:

$$E(b) = \beta$$

Onde b é a solução para as equações do sistema anterior, na forma matricial.

A matriz de variância-covariância assintótica de b é dada por:

$$\mathrm{Var}(b) = \mathrm{I}^{-1}(b) = [\mathrm{X}'\mathrm{V}\mathrm{X}]^{-1} \cdot [a(\phi)]^2$$

Onde $V = diag\left\{\sigma_j^2\right\}$ e σ_j^2, que é função de μ, depende da distribuição em questão.

Caso não usamos a função de ligação, ou melhor, se $\eta_j = \theta_j$, a derivada da função log-verosimilhança, será dada por:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \frac{\partial \mathcal{L}}{\partial \theta_j}\frac{\partial \theta_j}{\partial \eta_j}\frac{\partial \eta_j}{\partial \beta} = \sum_{j=1}^{n}\left[\frac{1}{a(\phi)}\left(y_j - \frac{\partial b(\theta)}{\partial \theta_j}\right)\mathrm{x}_j\right] = \sum_{j=1}^{n}\left[\frac{y_j - \mu_j}{a(\phi)}\frac{\partial \theta_j}{\partial \eta_j}\,\mathrm{x}_j\right]$$

Na forma matricial, seria

$$X'\Delta(y - \mu) = 0$$

Onde $\Delta = diag\left\{\frac{\partial \theta_j}{\partial \eta_j}\right\}$

A matriz de variância-covariância assintótica de b é dada por:

$$\mathrm{Var}(b) = \mathrm{I}^{-1}(b) = [\mathrm{X}'\Delta\mathrm{V}\Delta\mathrm{X}]^{-1} \cdot [a(\phi)]^2$$

2.4 Regressão Linear

Muitos problemas em engenharia e ciências têm desenvolvimento ou objectivo de explorar as relações existentes entre as variáveis.

A Análise de Regressão é um método estatístico essencial e amplamente utilizado para investigar e quantificar a relação entre variáveis, sendo particularmente empregada para determinar o impacto ou a capacidade explicativa de uma ou mais variáveis independentes sobre o comportamento de uma ou mais variáveis dependentes.

O termo regressão surge em homenagem aos trabalhos iniciados por Francis Galton e desenvolvidos ou continuados por Karl Pearson na tentativa de se estimar uma função linear que procurava investigar a relação entre a altura dos filhos e a altura dos pais, de modo a se estabelecer uma eventual lei universal de regressão nos finais do século XIX. (Demétrio, C. G. B., & Zocchi, S. S., 2006)

Nesse estudo, onde foram utilizados os métodos de mínimos quadrados, Galton chegou a perceber que se a altura dos pais fossem acima da média, a altura dos filhos seria também acima da média, porém não tanto quanto a altura dos Pais. (Montgomery, D. C., Runger, G. C., & Calado, V., 2000) Galton percebeu ainda, a existência de uma linha que descreve a relação média entre as duas variáveis que designou de linha de regressão, dando enfase ao surgimento do modelo de regressão, através dos trabalhos de Karl Pearson. (Fávero, L. P., & Belfiore, P., 2017). Demétrio, C. G. B., & Zocchi, S. S. (2006) realça dizendo que se o pai fosse muito alto ou muito baixo, o filho teria uma altura tendendo à média, ou seja, existe uma tendência de os dados regredirem à média, razão pela qual se chamou de regressão.

2.4.1 Modelo de Regressão Linear

Segundo Fávero (Fávero, L. P., & Belfiore, P., 2017 apud Fávero et al., 2009), a técnica de regressão linear oferece, prioritariamente, a possibilidade de que seja estudada a relação entre uma ou mais variáveis explicativas, que se apresentam na forma linear, e uma variável dependente quantitativa.

De ressaltar que o modelo que está na génese da análise de regressão ficou conhecido como modelo de regressão linear simples uma vez que define uma relação linear entre a variável dependente (altura dos filhos adultos) e uma variável explicativa (altura média dos pais). Se em vez de uma forem incorporadas várias variáveis explicativas, o modelo passaria a designar-

se de modelo de regressão linear múltipla (Hair, J. F. Jr et al., 2006), ou ainda regressão linear multivariados se houver mais de uma variável dependente.

Demétrio, C. G. B., & Zocchi, S. S. (2006) apresentam a análise de regressão como uma técnica estatística que se preocupa em estudar a dependência de uma (várias) variável(is) dependente(s) (endógena(s), explicada(s) ou de resposta(s)) em relação às variáveis explanatórias (independentes, exógenas, explicativas ou preditoras), com vistas a estimar e/ou prever o valor médio (da população) da primeira em termos dos valores conhecidos ou fixados (em amostras) das segundas.

De acrescentar que outros autores como Hair, J. F. Jr et al. (2006), Marôco, J., (2021), Pestana, M. H., & Gageiro, J. N. (2020) e Fávero, L. P., & Belfiore, P. (2017), apresentam a regressão como uma técnica estatística que, dentre os objectivos, tem como foco a i) estimação dos parâmetros – ajustar o modelo aos dados, ii) inferência – realizar inferência sobre os parâmetros estimados, iii) selecção de variáveis – seleccionar as variáveis que afectam significativamente a variação da variável endógena e iv) previsão – conhecer o comportamentos das variáveis explanatórias que não estavam nos dados sobre a variável endógena.

O modelo de regressão linear (um modelo geral) pode ser definido da seguinte maneira:

$$\boldsymbol{Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj} + \varepsilon_j} \qquad \boldsymbol{(j = 1, 2, 3, \dots, n)}$$

Onde *j* representa o número de observações, β_i $(i = 1, 2, 3, \dots, p)$ são chamados de coeficientes ou parâmetros do modelo de regressão, ε_j são chamados de erros ou resíduos do modelo e X_i são as variáveis com as quais construímos o modelo.

Observa-se que o β_0 representa a ordenada na origem ou o coeficiente linear, ou seja, é o valor de Y_j quando $X_{ij} = 0$ $(i = 1, 2, 3, \dots, p)$, enquanto β_i representa os declives parciais (Marôco, J., 2021), ou seja, a variação do Y_j por unidade de variação do X_i.

Dado que o ε_j representa todas as outras variáveis (aleatórias e desconhecidas) capazes de influenciar o Y_j e que existe certa relação entre o valor esperado de Y_j e o modelo de regressão linear, o modelo de regressão anterior pode ser formulado da seguinte maneira (Marôco, J. (2021) e Jonhson, R. A., Wichern D. W. (2002)):

$$\boldsymbol{Y_j = E[Y_j | X_i] + \varepsilon_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj} + \varepsilon_j}$$

Observa que Jonhson, R. A., Wichern D. W. (2002) chama atenção para o termo "*linear*" na regressão pelo facto da média ser uma função linear para os parâmetros desconhecidos β_0, β_1, β_2, …, β_p.

De acordo com Fávero, L. P., & Belfiore, P. (2017), os termos de erro ocorrem em função de algumas razões que precisam ser conhecidas e consideradas pelos pesquisadores, onde destacam a: *i)* existência de variáveis agregadas e/ou não aleatórias, *ii)* incidência de falhas quando da especificação do modelo (formas funcionais não lineares e omissão de variáveis explicativas relevantes) e *iii)* ocorrência de erros quando do levantamento dos dados.

Portanto, a partir do modelo geral, podemos obter:

- ✓ **Modelo de Regressão Linear Simples** que define uma relação estatística linear entre uma variável preditiva e uma variável resposta do modelo, dado por:

$$Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j \qquad (j = 1, 2, 3, \dots, n)$$

- ✓ **Modelo de Regressão Linear Múltipla** que define uma relação estatística linear entre i-ésima variável preditiva e uma variável resposta do modelo, dado por:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj} + \varepsilon_j$$

- ✓ **Modelo de Regressão Linear Multivariada** que define uma relação estatística linear entre k-ésima variável resposta e um único conjunto de variável preditiva do modelo, dado por (Jonhson, R. A., Wichern D. W., 2002):

$$Y_{jk} = \beta_{0k} + \beta_{1k} X_{1j} + \beta_{2k} X_{2j} + \cdots + \beta_{pk} X_{pj} + \varepsilon_{jk} \qquad (k = 1, 2, 3, \dots, m)$$

2.4.2 Estimação e Significância do Modelo

Considerando que o modelo de regressão linear simples é uma particularidade do modelo de regressão linear múltipla e modelo de regressão linear multivariada não constitui alvo (objecto) desta dissertação, cingiremos apenas no modelo de regressão linear múltipla.

Montgomery, D. C., & Runger, G. C. (2009) chega a perceber que no ajuste do modelo de regressão múltipla, é mais conveniente expressar as operações matemáticas usando a notação matricial. Portanto, esse modelo é um sistema de n equações, que pode ser expresso na forma matricial como:

$$\underset{(n\times 1)}{\underset{\smile}{y}} = \underset{(n\times(p+1))}{\underset{\smile}{X}} \underset{((p+1)\times 1)}{\underset{\smile}{\beta}} + \underset{(n\times 1)}{\underset{\smile}{\varepsilon}}$$

onde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

De salientar que a primeira coluna da matriz das variáveis independente é formada por 1's de forma a obter o intercepto, ou seja, o β_0.

Um dos objetivos da análise de regressão é desenvolver uma equação que permita o investigador para prever certas caraterísticas para determinados valores das variáveis preditoras. (Jonhson, R. A., Wichern D. W., 2002) E esta previsão é tanto quanto ótimas se menor for a diferença entre os valores estimados e valores reais que, advém da consistência que dará aos coeficientes $\hat{\beta}$.

Assim, a estimação dos parâmetros do modelo, com recurso ao método dos mínimos quadrados, consiste na minimização da soma de quadrados dos erros – diferença entre os valores observados e valores estimados, dado por:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

onde o estimador para o vector $\hat{\beta}$ obtém-se da solução da equação

$$\frac{\partial}{\partial x}[(y - X\beta)'(y - X\beta)] = 0$$

Sem demonstração, Marôco, J. (2021 apud Myers, 1986) chega a seguinte equação

$$-2X'y + 2(X'X)\hat{\beta} = 0 \Leftrightarrow (X'X)\hat{\beta} = X'y$$

onde, X' é a transposta da matriz X. Multiplicando ambos os lados da equação pelo inverso da matriz $(X'X)$, conclui-se que o vector $\hat{\beta}$ (estimador de β) é dado por

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

A linearização da equação acima, pode ser obtida a partir da seguinte equação matricial

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{ip}x_{i1} & \sum_{i=1}^{n} x_{ip}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip}y_i \end{bmatrix}$$

De salientar que a matriz $X'X$ é simétrica e não singular. E o modelo ajustado, na forma matricial, é dado por:

$$\hat{y} = X\hat{\beta} + \epsilon$$

onde ϵ representa o resíduo – diferença entre o valor observado (y_i) e o valor ajustado $(\hat{y}_i)$ dado por: $\epsilon = y_i - \hat{y}_i = y - X\hat{\beta}$.

2.4.2.1. Propriedades do $\widehat{\boldsymbol{\beta}}$

Os estimadores de mínimos quadrados $\hat{\beta}$ são estimadores não enviesados dos coeficientes de regressão β, cujas propriedades $E[\hat{\beta}] = \hat{\beta}$ e $\text{Var}[\hat{\beta}] = \sigma^2(X'X)^{-1}$ são detalhadas a seguir: (Jonhson, R. A., Wichern D. W., 2002)

Sabe-se que $\hat{\beta} = (X'X)^{-1}X'y = [(X'X)^{-1}X'](X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$, então

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

Como a equação que melhor se ajusta aos dados devem ter i) a soma dos resíduos nula e ii) a soma dos resíduos ao quadrado o quanto mínima possível (Fávero, L. P., & Belfiore, P., 2017), então

$$E[\hat{\beta}] = \beta + (X'X)^{-1}X'E[\varepsilon] = \hat{\beta} + 0 = \hat{\beta}$$

Relativamente a matriz de variâncias e covariâncias, sabemos, por definição, que

$$\text{Var}[\hat{\beta}] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

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

$$\text{Var}[\hat{\beta}] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[((X'X)^{-1}X'\varepsilon)((X'X)^{-1}X'\varepsilon)']$$

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

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

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

Observa que na matriz $\sigma^2(X'X)^{-1}$, os elementos da diagonal principal são variâncias do $\hat{\beta}$ e os elementos fora da diagonal, são as covariâncias dos coeficientes de regressão.

2.4.2.2. Estimador para σ^2

Os autores como Montgomery, D. C., & Runger, G. C. (2009), Jonhson, R. A., Wichern D. W. (2002) e Demétrio, C. G. B., & Zocchi, S. S. (2006) chegam a perceber que existe um outro parâmetro desconhecido no modelo de regressão que precisa ser estimado σ^2 – variância do erro, onde sua raiz representa o **erro padrão**.

Assim, para encontrarmos o estimador para σ^2, seguiremos as seguintes demonstrações:

A partir da fórmula do resíduo, $\epsilon = y_i - \hat{y}_i = y - X\hat{\beta}$, tem-se

$$y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = (I - H)y$$

onde $H = X(X'X)^{-1}X'$, conhecida como a matriz Hat e $(I - H)$ é uma matriz quadrada, simétrica, idempotente e $tr(H) = p + 1$. (Kubrusly, J. (2014) e Jonhson, R. A., Wichern D. W. (2002))

Observa que se deseja encontrar um estimador não tendencioso para σ^2. Para tal Kubrusly, J. (2014) começa por demonstrar que ϵ é uma combinação linear dos erros, onde chega a confirmar pelos cálculos seguintes:

$$\epsilon = y_i - \hat{y}_i = y - X\hat{\beta} = X\beta + \varepsilon - X(\beta + (X'X)^{-1}X'\varepsilon) = (I - H)\varepsilon$$

Sabendo que encontrar um estimador não tendencioso para σ^2 consiste em encontrar $E[\epsilon'\epsilon]$, então pela equação anterior bem como a propriedade de matriz H, tem-se

$$E[\epsilon'\epsilon] = E\big[\big((I - H)\varepsilon\big)'\big((I - H)\varepsilon\big)\big] = E\big[\varepsilon'(I - H)'\big((I - H)\varepsilon\big)\big] = E[\varepsilon'(I - H)\varepsilon]$$

Observa que $\varepsilon'(I - H)\varepsilon$ é uma matriz de ordem 1, logo $\varepsilon'(I - H)\varepsilon = tr(\varepsilon'(I - H)\varepsilon)$ e pelas propriedades de traço de uma matriz

$$E[\epsilon'\epsilon] = E\big[tr\big((I - H)\varepsilon\varepsilon'\big)\big] = tr(E[(I - H)\varepsilon\varepsilon']) = tr\big((I - H)E[\varepsilon\varepsilon']\big)$$

$$= tr\big((I - H)\sigma^2 I\big) = tr\big((I - H)\sigma^2\big) = \sigma^2 tr(I - H)$$

$$= \sigma^2\big(n - (p + 1)\big) = \sigma^2(n - p - 1)$$

Portanto, estimador não tendencioso para a variância do erro, σ^2, é dada por: (Kubrusly, J. (2014) e Jonhson, R. A., Wichern D. W. (2002))

$$\frac{\epsilon'\epsilon}{n-p-1} = \frac{\sum_{i=1}^{n}\epsilon_i}{n-p-1} = \frac{y'(I-H)y}{n-p-1} = MSE$$

onde MSE – *Mean Squared Errors* (erro quadrático médio).

Sabendo que

$$\sum_{i=1}^{n}\epsilon_i = \sum_{i=1}^{n} y_i - \hat{y}_i \Leftrightarrow \sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SSE$$

onde SSE – *Sum of Squared Errors* (soma dos quadrados dos erros) que na forma matricial é dada por: (Montgomery, D. C., & Runger, G. C., 2009)

$$SSE = y'y - \beta'X'y$$

Acrescenta-se a SSR – *Sum of Squared Regression* (soma dos quadrados de regressão) dada por:

$$SSR = \beta'X'y$$

e TSS – *Total Sum of Square* (soma dos quadrados totais) dada por:

$$SST = SSR + SSE = y'y$$

2.4.2.3. Significância do modelo

Após a estimação dos parâmetros e consequentemente do modelo, o estudo da significância estatística dos mesmos é de fundamental importância, pois é através de testes de inferência estatísticas que permite não só, analisar a qualidade do ajustamento do modelo bem como verificar se o mesmo pode ser de facto inferida para o universo. (Pestana, M. H., & Gageiro, J. N., 2020)

Assim, Marôco, J., (2021) esclarece que o objectivo da análise inferencial é de avaliar, a partir das estimativas, se de facto, na população, alguma das variáveis independentes podem ou não influenciar a variável dependente, ou seja, se o modelo ajustado é ou não significativo. Esta hipótese é formalizada por:

$$\begin{cases} H_0: \beta = 0 \\ H_a: \beta \neq 0 \end{cases} \quad \text{ou} \quad \begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_a: \exists i: \ \beta_i \neq 0 \end{cases} \qquad (i = 1,2,\ldots,p)$$

Portanto, o autor explica que, se a fracção da variância total em Y explicada pelo modelo de regressão (estimada dividindo SSR pelos respectivos graus de liberdade), p – número de variável independente no modelo) for significativamente maior do que a proporção da variância total não explicada pelo modelo ou variância dos erros (estimada dividindo SSE pelos respectivos graus de liberdade, $n-p-1$), conclui-se que o modelo ajustado é significativo.

A estatística do teste para testar a significância do modelo é dada por:

$$F = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} = \frac{MSR}{MSE}$$

onde MSR – *Mean Squared of Regression* (quadrados médios de regressão) e F segue uma distribuição F-Snedecor com p e $(n-p-1)$ graus de liberdades.

Estes cálculos incluindo o *valor-p* podem ser agrupados numa tabela designada por ANOVA a Regressão, como se segue:

Tabela 4 – Tabela de ANOVA de Regressão

	SS	**Graus de Liberdades**	***MS***	***F***	**Probabilidade**
Regressão	$\beta' X' y$	p	$\frac{SSR}{p}$	$\frac{MSR}{MSE}$	$p\text{-}valorpue$
Erros	$y'y - \beta' X' y$	$n-p-1$	$\frac{SSE}{n-p-1}$		
Total	$y'y$	$n-1$			

A decisão sobre a significância do modelo dependerá do $valor\text{-}p$, onde, se $valor\text{-}p < \alpha$, rejeitamos a H_0, ou seja, pelo menos uma das variáveis independentes possui efeitos significativo sobre a variável dependente, ou ainda, o modelo ajustado aos dados é significativo.

Porém, carece-nos a verificação de significância de cada um dos parâmetros estimados, pois é preciso averiguar se todas ou apenas algumas variáveis independentes influenciam a variável dependente, ou seja, são significativos (diferentes de zero).

Assim, para averiguar qual(is) do(s) β é significativo, Marôco, J., (2021) e Jonhson, R. A., Wichern D. W. (2002) orienta-nos para a realização de múltiplos testes.

$$\begin{cases} H_0: \beta_i = k \\ H_a: \beta_i \neq k \end{cases} \qquad (i = 1,2,\dots,p)$$

onde k pode assumir qualquer valor. Contudo, na maioria das análises bem como dos *softwares* de analises estatísticas $k = 0$.

A estatística de teste é dada por:

$$T_{\beta_i} = \frac{\hat{\beta}_i - k}{\sqrt{MSE \cdot C_{ii}}}$$

onde C_{ii} é o i-ésimo elemento da diagonal da matriz $(X'X)^{-1}$ correspondente a β_i. Esta estatística de teste tem distribuição t-Student com $(n - p - 1)$ graus de liberdade se a hipótese nula é verdadeira.

A decisão sobre a significância do parâmetro dependerá do p-*value*, onde, se $valor\text{-}p < \alpha$, rejeitamos a H_0. Ora, se H_0 não for rejeitada, implica dizer que existe evidências de que a variável X_i não tem influência no modelo, ou seja, não é significativo, pelo que poderá ser excluída.

O **intervalo de confiança $100 \cdot (1 - \alpha)\%$ para o coeficiente de regressão β_i** é dado por: (Montgomery, D. C., & Runger, G. C., 2009).

$$\hat{\beta}_i - t_{(1-\alpha/2;\, n-p-1)}\sqrt{MSE \cdot C_{ii}} \leq \beta_i \leq \hat{\beta}_i + t_{(1-\alpha/2;\, n-p-1)}\sqrt{MSE \cdot C_{ii}}$$

O **intervalo de confiança $100 \cdot (1 - \alpha)\%$ para a resposta média (ou valor médio) de y_i** nas observações multivariadas $\mathrm{x}_j' = \left[1,\ x_{1j},\ x_{2j}, \ldots,\ x_{pj}\right]$ é dado por: (Marôco, J., 2021).

$$\hat{y}_i - t_{(1-\alpha/2;\, n-p-1)}\sqrt{\sigma^2 \cdot \mathrm{x}_j'(X'X)^{-1}\mathrm{x}_j} \leq y_i \leq \hat{y}_i + t_{(1-\alpha/2;\, n-p-1)}\sqrt{\sigma^2 \cdot \mathrm{x}_j'(X'X)^{-1}\mathrm{x}_j}$$

2.4.3 Qualidade do Ajustamento do Modelo

Para testar a qualidade de ajuste do modelo, Montgomery, D. C., & Runger, G. C. (2009), Marôco, J. (2021) e Fávero, L. P., & Belfiore, P. (2017) apresentam o coeficiente de determinação R^2 como a medida do grau de ajuste do modelo.

O coeficiente de determinação R^2 mede a proporção da variabilidade total que é explicada pelo modelo de regressão, ou seja, mostra quanto do comportamento da variável Y é explicado pelo comportamento de variação conjunta das variáveis X.

O R^2 pode calcular-se pela expressão:

$$R^2 = \frac{SSR}{SST}$$

O R^2 pode variar entre $0 \leq R^2 \leq 1$, porém Fávero, L. P., & Belfiore, P. (2017) considera praticamente impossível a obtenção de um $R^2 = 1$, ajustamento perfeito, uma vez que dificilmente todos os pontos situar-se-ão em cima de uma reta, uma vez que se isso acontecer, não haverá resíduos para cada uma das observações da amostra em estudo, e a variabilidade da variável *Y* estará sendo totalmente explicada pelo vetor de variáveis *X* consideradas no modelo de regressão.

Além disso, Montgomery, D. C., & Runger, G. C. (2009) considera a estatística R^2 de modo problemático como uma medida da qualidade do ajuste para um modelo de regressão múltipla, uma vez que a incorporação de mais uma variável independente tende a aumentar o seu valor, mesmo que esta possua influencia reduzida sobre a variável independente. Já Marôco, J., (2021) acrescenta a sua subjectividade no tocante a interpretação, i. e., nas ciências exactas $R^2 > 0{,}9$ são aceites como indicadores de um bom ajustamento, enquanto que nas ciências sociais basta $R^2 > 0{,}5$.

Assim, em alternativa a esta limitação Zar (Zar, 199 apud Marôco, J., 2021) propõe o coeficiente de determinação ajustado dado por:

$$R_a^2 = R^2 - \frac{p(1 - R^2)}{n - p - 1}$$

Este coeficiente de determinação ajustado segundo Marôco, J. (2021) pode ser considerado como o melhor estimador da qualidade de ajustamento uma vez que a adição de uma nova variável ao modelo apenas fará com que o R_a^2 aumente se essa variável conduzir a um melhor ajustamento do modelo aos dados.

De salientar que embora o R^2 varia ente 0 e 1, $0 \leq R^2 \leq 1$, o R_a^2 pode assumir valor negativo, i. e., quando o as variáveis explicativas, tomadas em conjunto, reduzirem a soma dos quadrados dos resíduos em um montante tão pequeno que esta redução não consiga compensar o fator $(n - 1)/(n - p - 1)$. (Fávero, L. P., & Belfiore, P., 2017)

2.4.4 Validação dos pressupostos Modelo de Regressão Linear Múltipla

Segundo Marôco, J. (2021) o modelo de regressão linear prossegue, geralmente após a estimação dos coeficientes de reggressão, com a validação dos prossupostos respeitantes aos erros ou resíduos e à (*quasi*)ortogonalidade entre as variáveis independentes.

Os erros ou resíduos, ε_j, do modelo servem tanto para estimar os parâmetros do modelo quanto para validar os prossupostos de aplicação do modelo de regressão linear, onde a previsão e a inferência acerca do modelo só é válida quando:

- ✓ $\varepsilon_j \sim N(0, \sigma^2)$ – Os erros possuam distribuição normal com média nula e variância constante;
- ✓ $Cov(\varepsilon_k,\ \varepsilon_l) = 0$ – Os erros são independentes.

Portanto, seguiremos para análise pormenorizada de cada um dos pontos anteriores, conde começaremos por:

i. Normalidade dos Resíduos

O prossuposto da normalidade dos resíduos é requerida apenas e tão somente para que sejam validados os testes de hipótese dos modelos de regressão, ou seja, o pressuposto da normalidade assegura que o $valor\text{-}p$ dos testes te t do teste F sejam válidos.

Este prossuposto pode ser testada recorrendo aos testes de Kolmogorov-Smirnov com a correção proposta por H. Lilliefors ou com o teste de Shapiro-Wilk vistos na secção 2.1.1. porém, Marôco, J. (2021) acrescenta que o SPSS Statístics apresenta, no modelo de regressão, um procedimento gráfico (resíduos padronizados) para visualizar esta condição, onde na situação ideal, os erros ou resíduos distribuir-se-ão de forma aleatória em torno de zero (0), $N(0,\ 1)$, onde o resíduo padronizado é dado por:

$$d_j = \frac{e_j}{\sqrt{MSE}}$$

Contudo, Wooldridge (2012 apud Fávero, L. P., & Belfiore, P., 2017) argumenta que a violação deste pressuposto pode ser minimizada quando da utilização de grandes amostras, devido às propriedades assimptóticas dos estimadores obtidos por mínimos quadrados.

ii. Homogeneidade da variância dos resíduos (Homocedasticidade)

A homocedasticidade se refere à suposição de que as variáveis dependentes exibem níveis iguais de variância ao longo do domínio da(s) variável(is) preditora(as), ou melhor, ocorre quando a variância dos termos de erro (ε) parece constante ao longo de um domínio de variáveis preditoras (Hair, J. F. Jr et al., 2006), i. e., $Var(\boldsymbol{\varepsilon_j}) = \sigma^2$.

Quando os termos de erro têm variância crescente ou flutuante, diz-se que os dados são heteroscedásticos.

Para Fávero, L. P., & Belfiore, P. (2017) a heteroscedasticidade se deve aos erros de especificação[3] quanto a forma funcional, presença de *outliers* ou quanto a omissão de uma variável relevante.

Segundo Wooldridge, J. M. (2011) e Pestana, M. H., & Gageiro, J. N. (2020) a heteroscedasticidade não provoca viés ou inconsistência nos estimados dos mínimos quadrados, mas deixam de ser eficientes bem como os estimadores de variâncias, $\mathrm{Var}(\hat{\beta}_j)$, são viesados sem a hipótese de homocedasticidade. Portanto, como os erros-padrão dos estimadores dos mínimos quadrados são baseados directamente nessas variâncias, elas não são mais válidas para construirmos intervalos de confiança e testes de hipóteses das estatísticas t, mesmo na presenta de amostras de grande dimensão.

O diagnóstico de heteroscedasticidade pode ser feito mediante o gráfico de resíduo padronizado versus $\hat{y}$, onde o padrão desejado (padrão de homocedasticidade) seria uma distribuição constante da variância (a heteroscedasticidade mais comum apresenta a forma de um triângulo) ou através de testes estatísticos.

As hipóteses de teste são:

$$\begin{cases} H_0: \text{As variâncias dos resíduos são homogénias} \\ H_a: \text{As variâncias dos resíduos não são homogénias} \end{cases}$$

Os testes estatísticos para verificar a homocedasticidade o **teste de Breusch-Pagan/Cook-Weisberg**, baseia-se no multiplicador de Lagrange (***LM***) e segue uma distribuição qui-quadrado com 1 grau de liberdade, porém Fávero, L. P., & Belfiore, P. (2017) alerta para a necessidade de verificar a normalidade dos resíduos e o **teste de White** que, segundo Marôco, J. (2021) é de aplicação mais generalizada uma vez que não tem assunções quer sobre a forma

[3] Um modelo sofre a má-especificação da forma funcional quando não explica de maneira apropriada a relação entre as variáveis explicativas e a dependente observada. (Wooldridge, J. M. 2011)
Esta 'má-explicação' do modelo pode ser devido a omissão de uma variável independente ou na escrita/definição incorrecta da variável exógenas ou mesmo endógena e pode ser testada pelo teste RESET (teste de erro de especificação de regressão), onde a estatística de teste é F com n-k-3 graus de liberdades.
A hipótese nula do teste é que o "modelo está correctamente especificado", rejeitando-a para valores de $valor\text{-}p < \alpha$.

da heteroscedasticidade (homogeneidade das variâncias dos resíduos) quer sobre a normalidade da distribuição dos erros.

Quando o modelo contém 3 variáveis independentes, o teste de White basea-se na estimativa de:

$$\hat{\varepsilon}_j{}^2 = \delta_0 + \delta_1 X_{1j} + \delta_2 X_{2j} + \delta_3 X_{3j} + \delta_4 X_{1j}^2 + \delta_5 X_{2j}^2 + \delta_6 X_{3j}^2 + \delta_7 X_{1j} X_{2j} + \delta_8 X_{1j} X_{3j} + \delta_9 X_{2j} X_{3j} + \text{erro}$$

Para Wooldridge, J. M. (2011) o teste de White é a estatística LM para testar que todos os δ_i sejam zero, excepto o intercepto. Acresce ainda que, em comparação com o teste de **Breusch-Pagan/Cook-Weisberg**, o teste de **White** envolve o maior número de variáveis preditoras o que torna uma fraqueza na forma pura do mesmo, pois usa muitos graus de liberdades para modelos com um número moderado de variáveis independentes.

A estatística de teste é dada por:

$$W = n \cdot R^2_{\hat{\varepsilon}_j{}^2}$$

onde $R^2_{\hat{\varepsilon}_j{}^2}$ é o coeficiente de determinação do novo modelo. Sob H_0, W tem uma distribuição qui-quadrado com $[2p + (p-1) \cdot p/2]$ graus de liberdades e rejeita-se H_0 se $W \geq \chi^2_{\left[1-\alpha,\ \left(2p+(p-1)\cdot\frac{p}{2}\right)\right]}$.

Para mitigar a fraqueza do teste de White, Wooldridge, J. M. (2011) propõe **teste de White Especial** que busca usar os valores estimados por mínimos quadrados para verificar a existência de heterocedasticidade. Isso sugere testar a hoterocedasticidade estimando a equação:

$$\hat{\varepsilon}_j{}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \text{erro}$$

A estatística de **teste de White Especial** é dada por:

$$W_E = n \cdot R^2_{\hat{\varepsilon}_j{}^2}$$

Sob H_0, W_E tem uma distribuição qui-quadrado com p graus de liberdades e rejeita-se H_0 para $valor\text{-}p < \alpha$.

Segundo Hair, J. F. Jr et al. (2006) muitas vezes a heteroscedasticidade é o resultado da não-normalidade de uma das variáveis ou da não linearidade, e a correção de uma das violações costuma corrigir outra.

Se for diagnosticado a heteroscedasticidade, Fávero, L. P., & Belfiore, P. (2017), Wooldridge, J. M. (2011) e Pestana, M. H., & Gageiro, J. N. (2020) sugerem i) o **método de mínimos quadrados ponderados**, que é um caso particular do método de mínimos quadrados generalizados, pode ser aplicado quando se diagnostica que a variância dos termos de erro depende da variável explicativa, i. e., $\mathrm{Var}(\boldsymbol{\varepsilon_j}) = \sigma_\varepsilon^2 \cdot X_j$ ou $\mathrm{Var}(\boldsymbol{\varepsilon_j}) = \sigma_\varepsilon^2 \cdot X_j^2$ ou $\mathrm{Var}(\boldsymbol{\varepsilon_j}) = \sigma_\varepsilon^2 \cdot \sqrt{X_j}$.

Esse método consiste em transformar o modelo de tal maneira que os termos do erro passe a apresentar variância constante.

Suponhamos uma relação linear, modelo de regressão linear simples, entre ε_j e X_j de forma que $\mathrm{E}(\varepsilon_j)^2 = \mathrm{E}(c \cdot X_j)^2 = c^2 \cdot X_j^2$, onde c é uma constante.

A transformação que ocorrerá no modelo que possa atender o prossuposto da homogeneidade seria:

$$\frac{Y_j}{X_j} = \frac{\beta_0}{X_j} + \frac{\beta_1 X_j}{X_j} + \frac{\varepsilon_j}{X_j}$$

Pois, os novos termos de erro apresentam a seguinte variância:

$$\mathrm{E}\left(\frac{\varepsilon_j}{X_j}\right)^2 = \frac{1}{X_j^2} \cdot \mathrm{E}(c \cdot X_j)^2 = \frac{1}{X_j^2} \cdot c^2 \cdot X_j^2 = c^2$$

De salientar que, Wooldridge, J. M. (2011) acrescenta dizendo que na maioria dos casos, a forma exacta de heteroscedasticidade não é óbvia, ou melhor, é difícil de encontrar um X_j causador da heteroscedasticidade. Nestes casos recorre-se a transformação do ii) **mínimo quadrado generalizados factível ou estimado** (ver Wooldridge, J. M., 2011).

Porém, Fávero, L. P., & Belfiore, P. (2017) e Wooldridge, J. M. (2011) consideram menos fatídico o iii) **método de Huber-White para erros-padrão robustos**, desenvolvido por White (1980) seguido do trabalho de Huber (1967), que consiste em estimar $\mathrm{Var}(\hat{\beta}_j)$, substituindo σ_ε^2 por ${\varepsilon_j}^2$, uma vez que σ_ε^2 não é diretamente observável. Assim,

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \epsilon_i^2}{SSE_j^2}$$

onde $\boldsymbol{\epsilon_i}$ representa o i-ésimo resíduo da regressão X_j sobre todas as variáveis independentes**,** obtidos por meio da elaboração da regressão original e $\boldsymbol{SSE_j}$ é a soma dos resíduos quadrados

dessa regressão, ou seja, representam os resíduos obtidos por meio da elaboração de cada regressão auxiliar da variável preditora X_j contra todos os demais variáveis preditoras.

Fávero, L. P., & Belfiore, P. (2017) acrescenta dizendo que o uso do **erros-padrão robustos a heteroscedasticidade** tem sido prática nos trabalhos académicos, pois evita a verificação da existência da heteroscedasticidade, o que "*acaba por eliminar uma incerteza correspondente à fonte da heteroscedasticidade e que eventualmente gera uma eventual confiança em resultados mais robustos, não representa uma verdadeira solução na grande maioria das vezes.*"

iii. Independência dos Resíduos (Autocorrelação)

Para existir autocorrelação nos resíduos é necessário que os erros sejam correlacionados com os valores anteriores ou posteriores aos dados. A autocorrelação significa a violação da independência das variáveis residuais (Pestana, M. H., & Gageiro, J. N., 2020), ou seja,

$$E[\epsilon_i, \epsilon_j] = \mathrm{Cov}(\epsilon_i, \epsilon_j) = \sigma_{ij} \neq 0 \qquad (i \neq j, \quad i, j = 1,2, \dots, n)$$

Assim, a independência dos resíduos se verifica quando a magnitude de um resíduo não influencia na magnitude do resíduo seguinte.

A autocorrelação pode surgir devido a especificação incorreta do modelo da regressão, ou por causa de erros na forma do modelo ou por exclusão de variáveis independentes importantes para a análise ou devido a existência de erros sistemáticos de medida das observações das variáveis endógenas.

O autor acrescenta ainda que, a autocorrelação é mais provável nos modelos cronológicos do que nos modelos seccionais. E para estimar a covariância, pode-se supor que ϵ_i seguem um processo de médias móveis ou um processo autorregressivo de primeira ordem, dado por $[AR_{(1)}]$:

$$\epsilon_i = \rho\epsilon_{i-1} + u_i$$

Portanto, a sua verificação pode ser pela análise do gráfico dos resíduos comparados com os valores preditos, onde este deve apresentar pontos dispersos aleatoriamente ou pelo teste de Durbin-Watson.

A hipótese do teste é dada por:

$$\begin{cases} H_0\colon\ \rho_{e_{j+1};e_j} = 0 \\ H_a\colon \exists j:\ \ \rho_{e_{j+1};e_j} \neq 0 \qquad\qquad (i \neq j, \qquad i,j = 1,2,\dots,n) \end{cases}$$

E a estatística de teste proposto por Durbin-Watson é dada por: (Marôco, J., 2021)

$$DW = \frac{\sum_{j=1}^{n-1}\left(e_{j+1} - e_j\right)^2}{\sum_{j=1}^{n} e_j^2} = 2\left(1 - r_{e_{j+1};e_j}\right)$$

onde $0 \leq d \leq 4$, uma vez que $-1 \leq r_{e_{j+1};e_j} \leq 1$. Para $d = 2$ indica ausência de autocorrelação, $d > 2$ indica autocorrelação negativa e $d < 2$ indica autocorrelação positiva.

Os critérios de decisão a considerar na estatística DW é apresentado na tabela seguinte, que consiste em compara o valor de DW com os valores críticos ou tabelados ($DW_I -$ limite inferior e $DW_S -$ limite superior): (Marôco, J., 2021)

Tabela 5 – Região cítica ou de Decisão do teste DW

	Região de Rejeição e de Não Rejeição de H_0				
	$[0; DW_I[$	$[DW_I; DW_S[$	$[DW_S; 4 - DW_S[$	$[4 - DW_S; 4 - DW_I[$	$[4 - DW_I; 4[$
Decisão	Rejeitar H_0	Nada se pode concluir	Não Rejeitar H_0	Nada se pode concluir	Rejeitar H_0
	Autocorrelação Positiva		Autocorrelação Nula		Autocorrelação Negativa

Embora o teste DW seja empregue frequentemente, Gujarati, D. N., & Porter, D. C. (2008) alerta para o cuidado a ter com as hipóteses que fundamentam a sua aplicação, designadamente i) o modelo de regressão inclui o termo de intercepto; ii) as variáveis explanatórias são não estocásticas, ou fixas, em amostras repetidas; iii) Os termos de erro são gerados pelo processo autorregressivo de primeira ordem, dado por: $\epsilon_i = \rho\epsilon_{i-1} + u_i$; iv) pressupõe-se que o termo de erro ϵ_i seja distribuído normalmente; v) O modelo de regressão não inclui os valores desfasados da variável dependente como uma das variáveis explanatórias; e vi) não faltam observações nos dados, pois a estatística DW não faria concessão para essas observações faltantes.

O autor acrescenta ainda que, na presença de uma autocorrelação positiva os erros-padrão dos coeficientes de regressão tendem a ser subestimados provocando o aumento das probabilidades do erro do tipo I nos testes de significância e na presença de uma autocorrelação negativa os erros-padrão dos coeficientes de regressão tendem a ser sobrestimados provocando o aumento das probabilidades do erro do tipo II nos testes de significância.

Na existência de autocorrelação forte, o autor recomenda a estimação dos coeficientes pelo método dos mínimos quadrados generalizados.

Se verificar a autocorrelação, depois de aplicar um ou mais testes diagnósticos, Gujarati, D. N., & Porter, D. C. (2008) orienta no sentido de i) tentar verificar se é um caso de autocorrelação pura e não o resultado da má-especificação do modelo; ii) se for autocorrelação pura, pode-se usar a transformação adequada do modelo original de modo que, no modelo transformado não tenha o problema de autocorrelação (pura); iii) Em amostras grandes, pode-se usar o método de *Newey-West* para obter os erros padrão dos estimadores de MQO que estão corrigidos para a autocorrelação; ou iv) continuar a usar o método dos mínimos ordinário organizado.

Ora, para evitar algumas das "armadilhas" ou limitação do teste de Durbin-Watson, os estatísticos Breusch e Godfrey desenvolveram um teste de autocorrelação que é genérico no sentido de que não permite i) variáveis preditoras não estocásticos; ii) esquemas autorregressivos de ordem superior, como $AR_{(1)}$, $AR_{(2)}$, $AR_{(3)}$, etc.; e (3) médias móveis simples ou de ordem mais elevada de termos de erro de ruído branco. (Gujarati, D. N., & Porter, D. C., 2008)

Assim, por meio da estimação por mínimos quadrados ordinários do modelo de regressão linear múltipla, onde os termos de erro sofrem um processo autorregressivo de ordem p, pode-se obter a seguinte equação: (Fávero, L. P., & Belfiore, P., 2017)

$$\hat{\varepsilon}_i = \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj} + \hat{\rho}_1 \hat{\varepsilon}_{i-1} + \hat{\rho}_2 \hat{\varepsilon}_{i-2} + \cdots + \hat{\rho}_k \hat{\varepsilon}_{i-k} + u_t$$

A estatística do teste de Breusch-Godfrey, baseado no teste LM (multiplicador de Lagrange), é dada por:

$$BG = (n - p) \cdot R^2_{\hat{\varepsilon}_i}$$

Onde n é o tamanho da amostra, p é a dimensão do processo autorregressivo e $R^2_{\hat{\varepsilon}_i}$ é o coeficiente de determinação do novo modelo ou equação anterior.

Sob H_0, ausência de correlação serial de qualquer ordem, o teste BG tem uma distribuição qui-quadrado com p graus de liberdades e se $BG > \chi^2_{[1-\alpha,\ p]}$, rejeita-se H_0, ou seja, pelo menos um parâmetro p é estatisticamente diferente de zero. (Gujarati, D. N., & Porter, D. C., 2008)

O autor acrescenta ainda que, a principal desvantagem do teste de Breusch-Godfrey é não permitir que se defina, a priori, o número de defasagens p na equação/modelo, fazendo com que se tenha de testar diversas possibilidades de p.

iv. Linearidade dos Resíduos

O modelo de regressão linear só é válido se na análise do resíduo indicar uma relação do tipo linear. Essa relação linear se refere à forma pela qual os parâmetros e os erros entram na equação e não necessariamente entre a relação entre as variáveis envolvidas.

O seu diagnóstico é facilmente identificado por meio de gráficos de resíduos, onde qualquer padrão curvilíneo nos resíduos indica uma relação não linear.

Para corrigir a linearidade Marôco, J. (2021) sugere algunas transformações matemáticas capazes de linearizar as relações não lineares como se segue no quadro seguinte:

Quadro 1 – Transformações matemáticas para linearizar as relações não lineares

Funções	$\boldsymbol{Y = \beta_0 X^{\sum_{i=1}^{p} \beta_i}}$	$\boldsymbol{Y = \beta_0 e^{X \cdot \sum_{i=1}^{p} \beta_i}}$	$\boldsymbol{Y = \frac{X}{\beta_0 X + \sum_{i=1}^{p} \beta_i}}$
Transformações	$Y' = \ln Y$		$Y' = \frac{1}{Y}$; $X' = \frac{1}{X}$
Forma linear	$Y' = \ln \beta_0 + \ln X \cdot \sum_{i=1}^{p} \beta_i$	$Y' = \ln \beta_0 + X \cdot \sum_{i=1}^{p} \beta_i$	$Y' = \beta_0 + X' \cdot \sum_{i=1}^{p} \beta_i$

Pestana, M. H., & Gageiro, J. N. (2020) acrescenta que no caso de funções não lineares com apenas uma variável exógena, as transformações possíveis são a potência, a exponencial, a inversa, a exponencial inversa e a logarítmica. E na presença de transformações, a interpretação das variáveis são alteradas.

As transformações logarítmicas são as mais frequentes, por ser útil em várias circunstâncias. Como por exemplo, a utilização deste tipo de variáveis permite que se faça interpretações percentuais e, por esse motivo, comparáveis em realidades distintas, e permite ainda reduzir a variação da variável, limitando o efeito dos *outliers*, que em último caso poderia corrigir a heteroscedasticidade.

Nos modelos que envolvem funções logarítmicas, os parâmetros têm as seguintes interpretações de acordo com a sua forma ou definição (Marôco, J., (2021), Pestana, M. H., & Gageiro, J. N. (2020), Gujarati, D. N., & Porter, D. C. (2008) e Fávero, L. P., & Belfiore, P. (2017)):

Quadro 2 – Interpretação de Modelos envolvendo Logaritmos

Tipo de Modelo	Variável Endógena	Variável Exógena	Interpretação

Linear-Linear	Y	X	$\Delta X = 1 \Longrightarrow \Delta Y = \beta$
Log-Linear	$\ln Y$	X	$\Delta X = 1 \Longrightarrow \Delta Y = (\beta \times 100)\%$
Linear-Log	Y	$\ln X$	$\Delta X = 1\% \Longrightarrow \Delta Y = \beta/100$
Log-Log	$\ln Y$	$\ln X$	$\Delta X = 1\% \Longrightarrow \Delta Y = \beta\%$

Seja o modelo de regressão linear dado pela seguinte equação:

$$\ln(Y_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_p X_{pj} + \varepsilon_j$$

A interpretação para o modelo seria, por exemplo, da seguinte forma: quando $\boldsymbol{X_{pj}}$ varia uma unidade, $\boldsymbol{Y}$ varia aproximadamente $(\boldsymbol{\beta \times 100})\%$, mantendo o resto constante.

v. (Multi)Colinearidade

Na análise de regressão, principalmente na regressão múltipla, é de extrema importância a verificar se as variáveis explicativas são ou estão correlacionadas.

Segundo Hair Jr. *et al* (2005), a colinearidade é a associação, medida como a correlação, entre duas variáveis independentes, enquanto a multicolinearidade refere-se a correlação entre três ou mais variáveis independentes, onde a mesma ocorre quando qualquer variável independente é altamente correlacionada com um conjunto de outras variáveis independentes.

O autor acrescenta ainda que a multicolinearidade representa o grau em que qualquer efeito de variável pode ser previsto ou explicado pelas outras variáveis na análise e, a medida que a multicolinearidade aumenta, fica mais complicada a interpretação da variável estatística, uma vez que se torna mais difícil verificar o efeito de qualquer variável, devido a suas inter-relações.

Segundo Pestana, M. H., & Gageiro, J. N. (2020), a multicolinearidade tem a sua origem mais frequente em modelos temporais do que nos modelos seccionais, como por exemplo, o recurso a variáveis desfasadas no tempo, evolução semelhante de variáveis ao longo do tempo, por estarem globalmente submetidas a grandes alterações, como é o caso dos movimentos cíclicos e de tendência ao longo prazo, já nos modelos seccionais tem-se a dimensão das unidades no comportamento das variáveis. e Fávero, L. P., & Belfiore, P. (2017) acrescenta uma outra causa

bastante comum da multicolinearidade que é a utilização de bancos de dados com um número insuficiente de observações.

A multicolinearidade entre variáveis independentes tem substancial impacto sobre a especificação do modelo, reduz o poder preditivo de qualquer variável independente na medida em que ela é associada com outras variáveis independentes, dificulta a interpretação dos coeficientes do modelo, aumenta as oscilações nas estimativas para os parâmetros do modelo quando se incluem ou excluem variáveis com elevadas correlações.

Os autores como Pestana, M. H., & Gageiro, J. N. (2020), Kubrusly, J. (2014) e Fávero, L. P., & Belfiore, P. (2017) mostram que na presença da multicolinearidade a matriz $X'X$ será quase não invertível, pois esta teria um autovalor muito próximo de zero e sua inversa, $(X'X)^{-1}$, teria um autovalor muito grande. Ainda que seja possível a inversa, a estimativa dos coeficientes β seriam pouco precisa, ou seja, pequenas mudanças nos valores observados de y alteram muito $\hat{\beta}$, bem como variância dos estimadores de mínimos quadrados assumirão valores elevados.

Os autores como Marôco, J. (2021), Pestana, M. H., & Gageiro, J. N. (2020), Kubrusly, J. (2014) e Fávero, L. P., & Belfiore, P. (2017) apresentam diversas formas para diagnosticar a multicolinearidade. A mais simples e intuitiva consiste em procurar, na matriz de correlação entre as variáveis da base de dados, aqueles que apresentam $|r| > 0{,}75$, pois conduzem geralmente ao problema da multicolinearidade.

Porém, esta técnica se aplica apenas para correlação bivariada. Para correlação entre mais de duas variáveis tem-se a ***Tolerância*** que é definida como a quantia de variabilidade da variável independente selecionada não explicada pelas outras variáveis independentes, dada por: (Hair Jr. *et al.*, 2005)

$$T = 1 - R_i^2$$

onde R_i^2 é o coeficiente de determinação entre do modelo entre a variável X_i e restante variável independente.

O valor da tolerância esta entre $0 < \boldsymbol{T} < 1$. Quanto mais próximo de zero estiver, maior a multicolinearidade e quando mais próximo de um estiver, menor a multicolinearidade. Segundo Menard (1995 apud Pestana, M. H., & Gageiro, J. N., 2020), os valores de $T < 0{,}2$ geram elevada multicolinearidade.

O ***Factor de inflação da Variância*** (do anglo-saxónico *Variance Inflaction Factor* - ***VIF***) é o inverso do valor de tolerância:

$$VIF = \frac{1}{T} = \frac{1}{1 - R_i^2}$$

Para Hair Jr. *et al.* (2005) VIF tem seu nome devido ao fato de que a sua raiz quadrada, $\sqrt{VIF}$, é o grau em que o erro padrão aumentou devido à multicolinearidade.

Na perspetiva de Montgomery & Peck (1982, apud Marôco, J., 2021), $VIF > 5$ indica problemas com a estimação de β_i devido à presença de multicolinearidades nas variáveis independentes, enquanto Myers (1990 apud Pestana, M. H., & Gageiro, J. N., 2020) considera a partir de $VIF > 10$.

Tem-se também, os **valores próprios** ($eigenvalues - \lambda_i$) que medem a quantidade de variâncias contida na matriz de correlações, dando indicação do número de dimensões distintas que existem entre as variáveis exógenas. Cada variável exógena está distribuída por várias dimensões ou valores próprios ou ainda, componentes principais, onde a primeira dimensão explica a maior proporção de variância dos dados, a segunda dimensão, ortogonal à primeira, explica a segunda maior proporção e assim sucessivamente.

Portanto, os valores indicam a proporção da dispersão relativa em cada uma das dimensões do espaço das variáveis exógenas, enquanto os vetores próprios identificam a combinação linear das variáveis exógenas que definem uma dada dimensão.

Se uma ou mais variáveis exógenas for colinear com as restantes, haverá pelo menos um valor próprio muito perto de zero. E quando há muitos valores próprios perto de zero, significa que existe elevada corelação entre as variáveis, fazendo com que pequenas variações nos dados possam conduzir a grandes variações nos coeficientes estimados. (Pestana, M. H., & Gageiro, J. N., 2020)

Dada a dificuldade em decidir o quando pequeno um dado $eigenvalues$ terá de ser (relativamente aos restantes) para que se possa concluir a existência de multicolinearidades, Marôco, J. (2021) e Pestana, M. H., & Gageiro, J. N. (2020), apresentam o ***conditional index*** e a ***variance proportion*** – *proporção da variância*, que devem ser analisadas simultaneamente.

O ***conditional index*** fornece o tamanho relativo da matriz dos valores próprios e é dado por:

$$k = \sqrt{\lambda_{\max}/\lambda_i}$$

onde $\lambda_{\max}$ é o maior valor próprio e λ_i é o valor próprio de cada dimensão definida pelo número de variáveis independentes no modelo.

Para $15 < k < 30$ indica possível problema de multicolinearidades, já $k > 30$ indica sérios problemas de multicolinearidades.

Variance proportion é a proporção da variância de cada coeficiente de regressão explicada por cada dimensão, variando entre zero e um.

No que respeita a decisão, as componentes ou dimensões que apresentam cumulativamente os $eigenvalues$ – valores próprios próximos de zero, *conditional index* maior que 30 e a *variance proportion* superior a 0,90 em pelo menos duas variáveis exógenas possuem elevada multicolinearidade.

Marôco, J. (2021), Kubrusly, J. (2014) e Pestana, M. H., & Gageiro, J. N. (2020), sugerem algumas alternativas para suprir a elevada multicolinearidade, e que levam a uma nova estimação do modelo, como: i) alterações à especificação do modelo, eliminando alguma das variáveis que se correlacionam entre si e que menos poder preditivo tenham em Y; ii) substituição de cada variável excluída por outra que meça o mesmo tipo de influência, mas que mas tenha menos problemas de multicolinearidade; iii) agregação de variáveis que apresentem elevada multicolinearidade num indica que as substitua, e que meça a sua evolução conjunta, recorrendo por exemplo, à análise fatorial. No entanto este índice deve ter suporte teórico, para evitar erros de especificação do modelo; iv) transformação das variáveis; v) inclusão de novas observações no modelo; e vi) utilização de modelos de equações estruturais quando há diferentes variáveis colineares, de modo a estimar simultaneamente as várias relações existentes entre os seus coeficientes.

2.4.5 Diagnóstico de *Outliers* e Observações Influentes

Os pressupostos do modelo ajustado precisam ser validados para que os resultados sejam confiáveis, uma vez que ao ajustar uma equação de regressão aos dados, na maioria das vezes o valor observado não corresponde exatamente ao valor predito.

Os ***outliers*** são observações distintas das demais, que estão associadas a resíduos com valores elevados, melhor, são observações extremas, não características, que apresentam resíduos que são consideravelmente superiores, em termos absolutos, aos resíduos das outras observações.

Os *outliers* são oriundos do erro na introdução dos dados ou da própria característica da variável em estudo e tem um impacto severo na estimação dos coeficientes de regressão. Pois, Pestana, M. H., & Gageiro, J. N. (2020) salienta que a sua exclusão ou não depende por um

lado do tipo de informação que possam dar e, por outro lado, do contexto em ou população em que se inserem. Acresce que, excluindo-os, embora possa melhorar a qualidade de análise, corre-se o risco de generalização dos estudos.

A identificação dos *outliers* é feita essencialmente através dos resíduos estandardizados, estudantizados e estudantizados *deleted*, pela verificação de valores superiores a $Z_{\frac{\alpha}{2}}$ (valor crítico ou tabelado da distribuição normal padrão), em pelo menos um dos resíduos.

Os **resíduos estandardizados**, com média igual a zero e desvio-padrão igual a um, consiste em eliminar os efeitos da magnitude da escala de medida sobre os erros. Esse resíduo é dado por:

$$e'_j = \frac{e_j}{\sqrt{SSE}}$$

Porém, na prática, os erros geralmente apresentam algum tipo de correlação e raramente apresentam variâncias iguais, uma vez que as variâncias dependem do local onde a observação x_{ij}, que dá origem ao resíduo j, se encontra no domínio das variáveis independentes, pelo que a variância dos erros deve ser ponderada pela medida da influência (***Leverage***) que esta observação possui no ajustamento do modelo.

A este resíduo, dá-se o nome de **resíduos estudantizados**, dado por:

$$r_j = \frac{e_j}{\sqrt{SSE(1 - h_{jj})}}$$

onde h_{jj} é o elemento da diagonal da matriz $H = X(X'X)^{-1}X'$, conhecida como matriz *hat* – chapéu, e X' representa a transporta da matriz X.

Embora os resíduos estandardizados e estudantizados tomem valores semelhantes, este último é mais apropriado para detetar casos muitos influentes. (Marôco, J., 2021)

Já os **resíduos estudantizados *deleted*** ou **resíduos PRESS** são particularidades dos resíduos estudantizados para o caso em que o desvio-padrão dos erros é calculado a partir do modelo ajustado sem a observação x_{ij}. Esse resíduo é dado por:

$$t_j = \frac{e_j}{s_{-j}\sqrt{1 - h_{jj}}}$$

onde s_{-j} é o desvio-padrão, que pode ser calculado sem ajustar um novo modelo onde foi eliminado observação x_{ij}, dado por:

$$s_{-j} = \sqrt{\frac{(n-p)SSE - \frac{e_j^2}{1-h_{jj}}}{n-p-1}}$$

As **observações influentes** são aquelas que individualmente ou em conjunto com outras observações demonstram ter maior impacto do que as restantes observações, no cálculo do valor dos vários estimadores (coeficientes, erros padrão, valores do teste t de Student, valores previstos, resíduos e teste F de Snedecor). (Pestana, M. H., & Gageiro, J. N., 2020)

Para identificar as observações influentes, pode-se recorrer a:

- ✓ ***Leverage***

Leverage é uma medida de observações influentes que varia ente zero e um, $0 \leq h_{jj} \leq 1$, relativa a distância entre uma observação em relação ao centro das observações. E é obtido a partir da diagonal da matriz $H = X(X'X)^{-1}X'$.

Considera-se um *leverage* elevado quando: (Pestana, M. H., & Gageiro, J. N., 2020)

$$\begin{cases} h_{jj} > \dfrac{3 \cdot (p+1)}{n}, & \text{para } n \leq 30 \\ h_{jj} > \dfrac{2 \cdot (p+1)}{n}, & \text{para } n > 30 \end{cases}$$

onde n é o tamanho da mostra e p é o número de variáveis exógenas.

Valores de *leverage* superiores a 0,5, indica a presença de *outliers* multivariado. (Marôco, J., 2021)

Uma observação com um elevado *leverage* só será influente se o seu resíduo não for pequeno, pois, um elevado *leverage* implica apenas que essa observação é influente.

- ✓ ***DfFitS***

DfFitS (do inglês *Difference in Fits Statistic*) mede a alteração provocada no valor ajustado retirando a observação j, ou melhor, é uma medida estandardizada que indica o número de

erros-padrão em que os valores previstos serão alterados se a observação j for removida da análise, dado por: Marôco, J., 2021)

$$DfFitS_j = \frac{\hat{y}_j - \hat{y}_{j,-j}}{\sqrt{SSE_{-j} \cdot h_{jj}}}$$

onde $\hat{y}_{j,-j}$ representa o valor estimado para o caso x_{ij} $(i = 1, 2, \dots, p)$ e SSE_{-j} é a estimativa da variância do modelo quando os coeficientes do modelo de regressão são ajustados sem o caso j.

O autor acrescenta que, uma observação é influente se $|DFFitS| > 2 \cdot \sqrt{p/n}$

- ✓ ***DfBeta***

DfBeta (do inglês *Difference in Beta*) mede a alteração no vetor estimado $\hat{\beta}$ ao se retirar a j-ésima observação da análise, ou melhor, é uma medida estandardizada que indica a influência que a observação x_{ij} tem sobre a estimação de cada um dos coeficientes de regressão, dado por:

$$DfBeta_{i\,;j} = \frac{\beta_i - \beta_{i\,,-j}}{\sqrt{SSE_{-j} \cdot h_{jj}}}$$

onde $\beta_{i\,,-j}$ representa o valor do coeficiente de regressão i que se obtém quando a observação x_{ij} $(i = 1, 2, \dots, p)$ é eliminada ada análise.

Considera-se uma observação é influente quando: (Pestana, M. H., & Gageiro, J. N., 2020)

$$\begin{cases} \left|DfBeta_{i\,;j}\right| > 1{,}96, & \text{para } n \leq 30 \\ \left|DfBeta_{i\,;j}\right| > \dfrac{2}{\sqrt{n}}, & \text{para } n > 30 \end{cases}$$

- ✓ **Distância de Cook**

Para Vaz, F. E. D. C. R. (2020), em 1986, Cook apresentou uma proposta inovadora no diagnóstico de modelos, conhecida como **Distância de Cook**, que propõe avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo ou nos dados, ao invés da avaliação pela remoção individual ou conjunta de pontos.

Marôco, J. (2021) acrescenta que esta medida combina a informação dos resíduos estudantizados e da *leverage*. Esta medida é dada por:

$$DC_j = \frac{r_j \cdot h_{jj}}{(p+1)(1-h_j)}$$

onde $(p+1)$ é o número de parâmetros no modelo, incluindo a constante.

O autor considera ainda que $DC_j > 1$ são excessivamente influentes na estimação dos coeficientes de regressão e quanto maior for DC_j maior será o resíduo.

Com recurso ao IBM SPSS Statistics, tem-se ainda as seguintes medidas (Pestana, M. H., & Gageiro, J. N., 2020):

- ✓ **Rácio de Covariância (COV)**

Esta medida mede a influência da exclusão da observação i na variância estimada dos coeficientes estimados.

O rácio de covariância assume valores elevados quando a *leverage* é elevado e assume valores pequenos quando o resíduo estudantizado *deleted* é elevado.

Uma observação é influente quando $|\mathrm{COV} - 1| > \frac{3 \cdot (p+1)}{n}$.

- ✓ **Distância de Mahalanobis (MAH)**

Esta medida mede a influência de uma observação i em relação ao valor médio de todas as observações das variáveis exógenas e não o impacto nos valores preditos.

Uma observação é influente se apresentar valores substancialmente maiores do que restantes.

2.4.6 Modelos de regressão com variáveis binárias (*Dummy*)

O modelo de regressão linear é frequentemente contruído pelas variáveis quantitativas com escalas de medidas pelo menos intervalar. Porém, nas ciências sociais, é frequente o uso de variáveis medidas na escala nominal ou ordinal (qualitativa), o que inviabiliza a aplicação dos métodos de regressão sem devida precaução.

Pois, quando na regressão usamos variável dependente medida numa escala nominal dicotómica, a regressão ideal seria logística binária, e se essa variável for policotómica, a regressão logística multinomial.

Ora, se a variável dependente for medida na escala pelo menos intervalar e uma das variáveis independente for nominal o modelo de regressão linear pode ser usado com recurso a variáveis auxiliares indicadoras, chamada variáveis *dummy*.

O número de variáveis necessárias para a investigação de um fenómeno é direta e simplesmente igual ao número de variáveis utilizadas para mensurar as respectivas características, porém, Fávero, L. P., & Belfiore, P. (2017), alerta para o cuidado ter quando existem variáveis medidas na escala qualitativa, visto que é diferente, uma vez que as variáveis *dummy* obriga a obtenção de número acrescido de observações de modo a garantir a credibilidade dos valores estimados, além de criar novas variáveis para o modelo. Pois, uma variável qualitativa com k categorias necessita apenas de $k - 1$ variáveis artificiais, evitando a multicolinearidades.

Para Pestana, M. H., & Gageiro, J. N. (2020), *dummy* ou variáveis artificiais possibilitam a inclusão sob a forma aditiva, de variáveis qualitativas exógenas no modelo de regressão, assinalando a presença (valor um - 1) ou ausência (valor zero - 0) de determinada característica. Porém, Fávero, L. P., & Belfiore, P. (2017), acrescenta que as variáveis *dummy* devem ser utilizadas quando desejarmos estudar a relação entre o comportamento de determinada variável explicativa qualitativa e o fenômeno em questão, representado pela variável dependente.

- ✓ **Incorporação e determinação do número de variáveis artificiais.**

Nos modelos de regressão com variáveis artificiais, vários autores (Marôco, J., (2021), Pestana, M. H., & Gageiro, J. N. (2020), Gujarati, D. N., & Porter, D. C. (2008) e Fávero, L. P., & Belfiore, P. (2017)), recomendam a definição da categoria de referência ($dummy = 0$), onde a escolha é do pesquisador.

Este procedimento permitirá ao pesquisador estudar as diferenças que acontecem na variável resposta ao se alterar a categoria da variável qualitativa, uma vez que o β_j desta *dummy* representará exatamente a diferença que ocorre no comportamento da variável resposta quando se passa da categoria de referência da variável qualitativa para a outra categoria, estando o comportamento da categoria de referência representado pelo intercepto β_0. (Wooldridge, J. M., 2011)

O autor acrescenta ainda que, em muitos casos, variáveis *dummy* refletem escolhas de indivíduos ou de outras unidades (em oposição a algo predeterminado, como género).

Importa salientar que um modelo pode conter variáveis *dummy* com categorias múltiplas, como por exemplo sexo e estado civil que seria, homem-solteiro, mulher-solteira, homem-casado ou mulher-casada, ou várias variáveis *dummy*.

No caso das variáveis ordinais, que podem ter três ou mais categorias, podem ser configuradas como várias variáveis *dummy*.

Importa acrescentar que a categoria para a qual nenhuma variável binária é atribuída é conhecida como **categoria-base**, de controle, de comparação, de referência ou categoria omitida. Todas as comparações são feitas em relação à categoria de referência e o valor do intercepto (β_0) representa o valor médio da categoria de referência. (Gujarati, D. N., & Porter, D. C., 2008).

2.4.7 Métodos e Critérios de Seleção de Modelos

Um número grande de variáveis explanatórias (ou covariáveis) pode levar a um modelo que explique bem os dados, mas com um aumento de complexidade na interpretação e que torna o modelo mais dependente dos dados observados.

Por outro lado, um número pequeno de variáveis explanatórias (ou covariáveis) pode levar a um modelo de interpretação fácil, porém, com um ajuste fraco aos dados.

Em geral, deseja-se um modelo intermediário que explique bem os dados e que seja parcimonioso, isto é, com o menor número possível de parâmetros.

Segundo Vaz, F. E. D. C. R. (2020, apud Oliveira, 2010), no processo de seleção dos modelos envolve, por um lado, o **Modelo completo ou saturado** que contém n parâmetros linearmente independentes, um para cada observação, cuja matriz do modelo é uma matriz identidade de ordem n. Este modelo atribui toda a variação dos dados à componente sistemática, que o permite ajustar-se perfeitamente, reproduzindo os próprios dados. O modelo saturado serve de referência para medir a discrepância de um modelo intermédio com $p + 1$ parâmetros. E por outro lado, o **Modelo Nulo** que é o mais simples por ter apenas um único parâmetro, representado por um valor comum a todas as observações, $E[Y] = \beta_0$. A matriz do modelo corresponde a um vetor coluna unitário.

Segundo Pestana, M. H., & Gageiro, J. N. (2020), se por um lado a seleção de variáveis exógenas a incluir é fundamental para que não haja erros de especificação do modelo, o que a acontecer implicaria discrepâncias entre o modelo estimado e a teoria em que se sustenta, impossibilitando a generalização dos resultados para o universo, por outro lado, o método da sua introdução no modelo, tem também grande impacto no modelo de regressão linear múltipla.

Quanto a este último aspeto, o autor acrescenta que existem essencialmente três métodos de introdução de variáveis exógenas na regressão linear múltipla, designadamente a i) a **regressão *standard***, que no SPSS corresponde ao método *Enter*, ii) a **regressão hierárquica ou sequencial**, que se obtém através da *Syntax* do comando da regressão do SPSS e iii) a **regressão estatística ou Stepwise**, que se desdobra na *Forward selection*, na *Backward deletion* e na *Stepwise regression*.

De salientar que, segundo Marôco, J., (2021), os diferentes métodos podem conduzir aos diferentes "melhores modelos", porém, aconselha-se a utilização de todos e identificar quais variáveis são eliminadas por todos em simultâneos e de seguida utilizar o modelo mais parcimoniosos para ajustar as outras variáveis.

- **Regressão *Standard* (Método Enter*)***

Neste método todas as variáveis exógenas introduzidas são forçadas a permanecer modelo, com base num suporte teórico.

Cada X_j, supõe que entra no modelo após todos os restantes X já terem entrado, medindo assim o contributo adicional para a explicação da variação de Y, que seja diferente do contributo dos restantes X.

Pestana, M. H., & Gageiro, J. N. (2020), acrescenta que a interpretação dos resultados baseado neste método, deve ter sempre presente por um lado, a variação total de Y explicada pelo conjunto dos X, representada pelo R^2, e por outro lado, a contribuição única de cada X_j, representada pelo quadrado dos coeficientes de correlação parciais.

- **Regressão Hierárquica ou Sequencial**

Neste método o investigador com base na teoria, controla a entrada de variáveis, pois os X entram na equação segundo a ordem especificada, traduzindo pela ordem de entrada o seu grau de importância na afetação de Y.

Cada X_j, é avaliado em termos do que contribui para o R^2 no momento da entrada, após ter sido eliminado o efeito das variáveis que se encontram no modelo.

Por exemplo, admita-se que as variáveis são introduzidas pela ordem X_2, X_3, X_4. Assim, a questão que se põe é saber se X_4 dá algum contributo adicional para a previsão de Y, após as diferenças entre os sujeitos em X_2 e X_3, serem eliminadas. (Pestana, M. H., & Gageiro, J. N., 2020).

A soma dos quadrados das suas correlações semiparciais é igual ao R^2.

- **Stepwise**

Segundo Pestana, M. H., & Gageiro, J. N. (2020), este método, também conhecido como **Regressão Stepwise**, é um processo controverso, no qual a ordem das variáveis de entrada é apenas baseada no critério matemático e não na teoria.

O autor acrescenta ainda que, existem três versões deste método, designadamente i) Forward selection, ii) Backward delection e iii) Regressão Stepwise.

Para os autores Pestana, M. H., & Gageiro, J. N. (2020) e Marôco, J., (2021), na **Forward selection**, a equação começa com a constante, β_0, e os X_j são adicionados em cada etapa desde que satisfaçam os critérios estatísticos de entrada. Uma vez que un X_j, entra na equação já não é abandonado.

A variável X_j, que melhor prevê Y, ou seja, aquela que apresentar a maior correlação, em termos absoluto, com a variável dependente, é a primeira escolhida para entrar.

O critério para selecionar a próxima variável exógena é a que possua o maior coeficiente de correlação semiparcial, ou melhor, é aquela que apresentar a maior correlação com Y depois de ajustados os efeitos da variável anterior já adicionada sobre Y.

Na **Backward delection**, a equação começa com todos as variáveis independentes e vai-se eliminando um X_j de cada vez, na medida que não contribua significativamente para a regressão.

Já a **regressão Stepwise** é um é um híbrido dos dois métodos anteriores, ou seja, é um compromisso entre os dois métodos anteriores, na medida em que a equação começa com a constante, incluindo-se de cada vez a variável exógena que satisfaça o critério estatístico de

entrada, mas também podem ser eliminadas variáveis exógenas previamente incluídas numa dada etapa, por já não contribuírem para a regressão.

A vantagem deste método, é que permite a remoção de uma variável cuja importância no modelo é reduzida pela adição de novas variáveis. Este procedimento termina quando nenhuma das variáveis independentes ainda de fora, consegue entrar no modelo e nenhuma das variáveis independentes presente no modelo é expulsa. (Marôco, J., 2021)

Enquanto isso, Pestana, M. H., & Gageiro, J. N. (2020) assegura que este método essencialmente usada com três objetivos: i) identificar as variáveis exógenas que estejam correlacionadas entre si; ii) explorar os X_j de maior utilidade na previsão de Y_j, eliminando variáveis supérfluas com vista a melhorar a investigação futura, pelo que é uma técnica exploratória e não confirmatória, sendo nesta etapa apenas um processo de construção de um modelo e não um procedimento para o testar, iii) caso a validação cruzada, permita generalizar os resultados da alínea anterior para o universo, então, a etapa 2) passa a figurar como teoria a ser aplicada em investigações futuras, recorrendo-se agora aqui aos métodos Enter ou Hierárquico.

O autor alerta ainda que, no método ***Stepwise*** as decisões sobre a inclusão e omissão das variáveis depende fortemente dos critérios matemáticos escolhidos para tal, onde pequenas diferenças nessas estatísticas podem levar a outra seleção de variáveis, pelo que se torna indispensável proceder à validação cruzada.

- **Critérios para Seleção do Modelo**

A seleção de modelos pode estar baseada em diferentes critérios, como, por exemplo, na escolha de subconjuntos de tamanho pré-determinado, ou, então, na comparação de estatísticas com valores tabelados de referência, pois a seleção do modelo centra-se fundamentalmente na qualidade de ajuste e a sua complexidade.

Existem vários critérios para seleção de modelos, onde Gujarati, D. N., & Porter, D. C. (2008) destacam cinco critérios usados na seleção do modelo nomeadamente o i) teste F, ii) Coeficiente de determinação ajustado; iii) Estatística de Mallows; iv) Critério de Informação de Akaike (AIC); e o v) Critério de Informação de Bayes (BIC).

- ✓ **Coeficiente de determinação ajustado R_a^2 e o teste F**

O **teste F** permite verificar se o modelo que está sendo estimado de fato existe, uma vez que, se todos os β_j forem estatisticamente iguais a zero, o comportamento de alteração de cada uma das variáveis explicativas não influenciará em absolutamente nada o comportamento de variação da variável dependente. Fávero, L. P., & Belfiore, P. (2017) Razão pela qual, é imprescindível que modelo seja significativo.

Assim, para o **teste F** o bom modelo é quele que é significativo.

Demétrio, C. G. B., & Zocchi, S. S. (2006), acrescenta que comparação entre dois modelos M_{p+1} e M_{q+1} com $p < q$ parâmetros, a estatística do teste F poderia ser feita usando:

$$F = \frac{SSR_{p+1} - SSR_{q+1}}{(q-p)MSE}$$

onde MSE é obtido no modelo que contém o maior número de termos que podem ser considerados, também chamado de ***modelo maximal***.

Porém, esse teste não permite seleccionar o melhor modelo dentre os modelos significativos, ou seja, conhecer ou seleccionar aquele que apresenta o maior poder explicativo de determinado modelo de regressão, ou o percentual de variabilidade da variável endógenas que é explicado pelo comportamento de variação das variáveis exógenas.

Portanto, os autores como Kubrusly, J. (2014), Marôco, J., (2021), Hair, J. F. Jr et al. (2006) e Pestana, M. H., & Gageiro, J. N. (2020), o melhor modelo é aquele que apresenta o maior coeficiente de determinação ajustado R_a^2.

Entretanto, Gujarati, D. N., & Porter, D. C. (2008) adverte para que o tamanho da amostra n e a variável dependente sejam os mesmos.

- ✓ **Estatística de Mallows**

A estatística de Mallows baseia-se na razão entre a SSR de um modelo com p variáveis preditoras ($p \leq k$), sabendo que MSE é um estimador de σ^2 de um modelo com $k = p + 1$ variáveis preditoras incluindo intercepto, conhecido por:

$$C_p = \frac{SSE_p}{MSE} - (n - 2p)$$

Sabemos que $E[SSE_p]$ é um estimador não tendencioso do σ^2. Se o modelo com p variáveis preditoras for adequado na medida em que não sofre da falta de ajustamento, então $E[SSE_p] = (n-p)\sigma^2$, logo

$$E[C_p] \approx \frac{(n-p)\sigma^2}{\sigma^2} - (n-2p) \approx p$$

Ao selecionar um modelo de acordo com o critério C_p, deve-se procurar aquele que tenha um valor baixo de C_p, quase igual a p. Em outras palavras, seguindo o princípio da parcimônia, será selecionado um modelo com p variáveis preditoras $(p < k)$, que se ajuste bem aos dados. (Gujarati, D. N., & Porter, D. C., 2008)

- ✓ **Critério de Informação de Akaike (AIC)**

O Critério de Informação de Akaike (AIC - Akaike Information Criterion), proposto por Akaike (1974) se diferencia dos métodos de seleção anteriores por se tratar de um processo de minimização que não envolve testes estatísticos, onde a ideia central consiste em selecionar um modelo que esteja bem ajustado e tenha um número reduzido de parâmetros.

Segundo Ferreira, M. C. C. D. S. (2013), não envolve testes estatísticos e pode ser expresso em função do desvio do modelo, e é baseado na função de verossimilhança.

Para Kubrusly, J. (2014), o AIC é dado por:

$$AIC = n\ln(SSE) - n\ln(n) + 2k$$

onde k representa o número de parâmetros no modelo ajustados e n representa o número de observações. Para o AIC, um modelo será melhor se apresentar o menor AIC.

O autor acrescenta ainda que, quanto menor for SSE menor será o valor de AIC. Além disso quanto menor for k, isto é, quanto menos variáveis preditivas tiver o modelo, menor será o valor de AIC.

- ✓ **Critério de Informação de Bayesiano (BIC)**

De acordo com Ferreira, M. C. C. D. S. (2013), o Critério de Informação Bayesiano (BIC), também conhecido como Critério de Schwarz, foi proposto por Schwarz (1978), e é um critério de avaliação de modelos definido em termos da probabilidade a posteriori.

Demétrio, C. G. B., & Zocchi, S. S. (2006), acrescenta dizendo que o critério BIC penaliza mais fortemente modelos com um maior número de parâmetros do que o AIC tendendo, dessa forma, a selecionar modelos com um menor número de parâmetros.

O BIC é dado por: (Kubrusly, J., 2014)

$$BIC = n\ln(SSE) - n\ln(n) + k\ln(n)$$

Para o BIC, um modelo será melhor se apresentar o menor BIC.

Assim como a medida AIC, quanto menor for SSE menor será o valor de BIC e quanto menor for k menor será o valor de BIC

2.4.8 Previsão de novas observações

Outra das aplicações do modelo de regressão é a previsão pontual (e não média) de novos valores da variável dependente para um dado conjunto de observações das variáveis independentes.

No caso, da previsão, os novos valores da variável dependente que não fazem parte da amostra inicial onde foi deduzido o modelo de regressão. Assim, a variância do valor médio de y_i não serve para inferir acerca do valor futuro de um novo valor da variável dependente.

Portanto o **intervalo de confiança $100 \cdot (1-\alpha)\%$ para a $\hat{y}_i$ (ou futuras observações)** nas observações multivariadas $\mathrm{x}_j' = [1,\ x_{1j},\ x_{2j}, \dots,\ x_{pj}]$ é dado por: (Marôco, J., 2021).

$$\hat{y}_i - t_{(1-\alpha/2;\, n-p-1)}\sqrt{MSE \cdot (1+h_{jj})} \leq \hat{y}_i \leq \hat{y}_i + t_{(1-\alpha/2;\, n-p-1)}\sqrt{MSE \cdot (1+h_{jj})}$$

onde o desvio-padrão de uma estimativa pontual de y_j é dado por:

$$S_{\hat{y}_i} = \sqrt{MSE \cdot (1 + \mathrm{x}_j'(X'X)^{-1}\mathrm{x}_j)} = \sqrt{MSE \cdot (1+h_{jj})}$$

2.5 Regressão Logística

Os modelos de **regressão** mais conhecidos são aqueles em que a variável endógena é quantitativa. Porém, quando esta é não-métrica, a precisão dos modelos e dos estimadores são questionáveis, ou seja, apresentam fraca consistência e confiabilidade, embora sendo muitas as situações em que se pretende realizar uma análise de regressão, mas a variável dependente é qualitativa com valores de classe discreta e mutuamente exclusivos. (Marôco, J., 2021)

Para solucionar este problema Hair, J. F. Jr et al. (2006) apresenta a regressão logística como a técnica estatística apropriada. Estas técnicas que se apropriam da variável categórica estão associados ao grupo de modelos de regressão chamado de **Regressão Categorial**, que carrega consigo os mesmos propósitos da regressão linear nomeadamente a inferência e estimação.

Este tipo de regressão, na perspectiva de (Marôco, J., 2021), apresenta semelhanças com a Análise Discriminante, porém mais abrangente visto que i) os preditores podem ser variáveis qualitativas e/ou quantitativas, ii) não carece de relações lineares entre as variáveis, iii) não pressupõe a normalidade (distribuição normal) e iv) é menos sensível aos *outliers*.

A Regressão Categorial toma designações diferentes segundo as características da variável dependente em análise. Se esta for dicotómica (assume apenas um de dois valores como 1 - Sucesso, 0 - Insucesso) a regressão categorial é chamada de **Regressão Logística Binária**. Se a variável dependente tiver mais de duas categorias e medida em escala nominal, a regressão categorial é designada de Regressão Logística Multinomial. Se a variável dependente for medida em escala ordinal, a regressão categorial é designada por Regressão Ordinal.

De salientar que a regressão logística é uma técnica estatística muito utilizada na modelação estatística devido aos feitos de Cox (1970), embora conhecida desde os anos 50 (Paula, G. A., 2004) esta técnica tem a sua aplicação na medicina, quando se trata da razão de chances ou mesmo do risco associado a ocorrência de um evento, na área Financeira, temos a sua aplicação nos modelos como *behaviour scoring* e *credit scoring* e noutras áreas onde se tem variáveis resposta com características binárias.

2.5.1 Modelo de Regressão Logística Binária

Hair, J. F. Jr et al. (2006) apresenta a regressão logística como uma forma especializada de regressão que é formulada para prever e explicar uma variável categórica binária (dois grupos)

e não uma medida dependente métrica. Acrescenta ainda dizendo que a forma da sua variável estatística é semelhante à da regressão múltipla.

Embora haja essa semelhança, Hair, J. F. Jr et al. (2006) alerta para sua importância no sentido de que *i)* ela foi especificamente elaborada para prever a probabilidade de um evento ocorrer, *ii)* a natureza binária da variável dependente (0 ou 1) tem propriedades que violam as suposições da regressão múltipla e *iii)* nenhuma violação pode ser remediada por meio de transformações das variáveis dependente ou independentes.

A natureza binária da Regressão Logística permite associar a variável resposta a distribuição de Bernoulli cuja probabilidade de sucesso representa a característica de interesse, valor 1, dado por π. Porém, por se tratar de uma sequência de eventos com distribuição de Bernoulli, a variável resposta estará associada a distribuição Binomial.

O processo de estimação dos coeficientes do modelo logístico é semelhante ao do modelo de regressão linear, porém, obtido pelo método de máxima verossimilhança, por se tratar de variável dependente dicotómica.

Assim, a função de regressão Logística Binária para estimar a probabilidade de sucesso de j eventos com p variáveis independentes é dado por:

$$\hat{\pi}_j = \frac{e^{\beta_0+\beta_1 X_{1j}+\beta_2 X_{2j}+\cdots+\beta_p X_{pj}}}{1+e^{\beta_0+\beta_1 X_{1j}+\beta_2 X_{2j}+\cdots+\beta_p X_{pj}}}$$

ou na forma matricial

$$\hat{\pi} = \frac{e^{\mathrm{X}\beta}}{1+e^{\mathrm{X}\beta}}$$

onde $\hat{\boldsymbol{\pi}}$ é o vetor das probabilidades estimadas, **X** é a matriz das variáveis independentes cuja primeira coluna é um vetor de 1's e $\boldsymbol{\beta}$ é o vetor dos coeficientes de regressão.

Por se tratar de um modelo de regressão que não carece de linearidade, convencionalmente o referido modelo é ajustado mediante a linearização através da transformação $Logit(\pi)$. $Logit$ é a função de ligação (*link function*) nos modelos lineares generalizados que permitem linearizar a variável dependente permitindo a modelação em função de um modelo linear. (Marôco, J., 2021)

Além dessa vantagem que as funções de ligação apresentam, Vaz, F. E. D. C. R. (2020) acrescenta as mesmas, o pequeno número de restrições como: i) inclusão todas as variáveis para que se obtenha maior estabilidade, ii) valor esperado do erro igual a zero, iii) inexistência

de correlação entre os erros e as variáveis independentes e iv) ausência de multicolinearidade entre as variáveis independentes.

O $logit$ é dado pelo logaritmo natural da chance ou rácio de verosimilhança ou $Odds$, que se traduz na razão entre a probabilidade do sucesso face a probabilidade do insucesso, dado por:

$$Logit(\hat{\pi}) = Ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$

Marôco, J., (2021) chega a perceber que os $Odds$ e π são formas equivalentes de descrever o objectivo primário da regressão logística, cujo **modelo de regressão logística** é dado pela seguinte fórmula:

$$\boldsymbol{Logit}(\hat{\boldsymbol{\pi}}_{\boldsymbol{j}}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_{1j}} + \boldsymbol{\beta_2 X_{2j}} + \cdots + \boldsymbol{\beta_p X_{pj}}$$

ou na forma matricial

$$Logit(\hat{\pi}_j) = \mathrm{X}\beta$$

2.5.2 Pressuposto Modelo de Regressão Logística Binária

O modelo de **Regressão Logística** além de ser uma técnica recomendada para situações em que a variável dependente é de natureza binária, ela apresenta algumas vantagens em relação ao modelo de Regressão Linear, como i) facilidade para lidar com variáveis independentes categóricas, ii) fornece resultados em termos de probabilidade, iii) facilidade de classificação de indivíduos em categorias, bem como iv) o alto grau de confiabilidade.

Porém, Marôco, J., (2021) alerta para alguns prossupostos que devem ser satisfeitos como:

i. **Linearidade e aditividade:** a escala de $Logit(\pi)$ é aditiva e linear (mas a de π não);
ii. **Proporcionalidade:** a contribuição para cada X_i é proporcional ao seu valor com um fator β_i;
iii. **Constância de efeito**: a contribuição de uma variável independente é constante e independente da contribuição das outras variáveis independentes;
iv. **Os erros são independentes e apresentam distribuição binomial**; e
v. **Os preditores não são multicolineares** (à semelhança da regressão linear múltipla).

O autor acrescenta ainda que a validação dos prossupostos do modelo pode fazer-se graficamente pela análise dos resíduos e a multicolinearidade pode ser diagnosticada calculando a Tolerância a partir do R^2 dado por $T = 1 - R^2$.

Nakamura, K. G. (2013, apud Schaefer, 1986) afirma que, existe multicolinearidade no modelo de regressão logística quando há uma dependência linear exata ou aproximada entre as covariáveis do modelo e na sua presença tanto a estimação dos parâmetros quanto a da matriz de covariância dos estimadores dos parâmetros nos modelos ficam mais instáveis do que nos modelos de regressão linear.

A autora acrescenta ainda dizendo que as estimativas dos β, as estimativas das variâncias dos estimadores de seus componentes e de seu erro quadrático médio estarão distantes dos valores reais, fazendo com que a inferência com base nesta estatística fica seriamente comprometida, podendo ficar muito maior ou muito menor do que o seu verdadeiro valor, resultando em uma menor ou maior chance de a hipótese nula ser não rejeitada.

E a mesma finaliza, salientando que a covariável pode ser retirada ou incluída no modelo sem ser necessariamente significante.

2.5.3 Estimação e Significância do Modelo

Na regressão logística binária, a variável dependente segue uma distribuição de Bernoulli, tal que $Y_j \sim B(1, \pi)$, onde se tem:

$$P(Y = y_j) = \pi^{y_j}(1 - \pi)^{1-y_j}$$

Para estimação deste modelo, uma vez que a variável dependente é binária, Hair, J. F. Jr et al. (2006) alerta-nos para o uso do método de máxima verossimilhança (*likelihood function*).

Assim, aplicando a função verosimilhança e de seguida substituir pela função de regressão logística, tem-se

$$L = \prod_{j=1}^{n}[\pi^{y_j}(1 - \pi)^{1-y_j}] = \prod_{j=1}^{n}\left[\left(\frac{e^{X\beta}}{1 + e^{X\beta}}\right)^{y_j}\left(\frac{1}{1 + e^{X\beta}}\right)^{1-y_j}\right]$$

Para Fávero, L. P., & Belfiore, P. (2017) é mais conveniente se trabalhar com o logaritmo da função de verossimilhança (*log likelihood function*), onde se obtém o seguinte

$$LL = Ln(L) = \sum_{j=1}^{n}\left[y_j \ln\left(\frac{e^{\mathrm{X}\beta}}{1+e^{\mathrm{X}\beta}}\right) + (1-y_j)\ln\left(\frac{1}{1+e^{\mathrm{X}\beta}}\right)\right]$$

As estimativas de máxima verossimilhança para os β é obtido após a derivação da equação anterior em relação a cada um dos parâmetros e igualar a zero. Contudo, Marôco, J., (2021) afirma que a função que maximiza *LL* também maximiza *L*, porém o sistema não apresenta solução analítica, pelo que o β é obtido recorrendo ao algoritmo de Newton-Raphson, com o uso de *software*, para encontrar um mínimo do desvio entre o valor observado e o valor estimado.

Hosmer Jr, DW, Lemeshow, S., & Sturdivant, RX (2013) demonstra que o desvio ou erro-padrão, $\widehat{SE}(\hat{\beta}_i) = \sqrt{\hat{\sigma}_i(\hat{\beta}_i)}$, é obtido a partir da matriz de informação observada, citado por Marôco, J., (2021) como matriz da informação de Fisher, dada por $I(\hat{\beta}) = X'\hat{V}X$, estimado por recurso a matriz Hessiana (matriz das segundas derivadas da função verosimilhança em ordem aos parâmetros do modelo), onde X é, a matriz com a informação das variáveis independentes, dada por

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

e V é uma matriz de ordem n, cujo *i*-ésimo elemento da diagonal principal é a estimativa da variância, $\hat{\sigma}_i(\hat{\beta}_i)$

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \ddots & \vdots \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Após estimar os coeficientes de regressão, a significância da variável é o primeiro aspeto a observar antes de seguir com a análise. Na regressão logística, a comparação dos valores observados com os preditos é baseada na *função logaritmo da verosimilhança*.

Assim, as hipóteses estatísticas para testar a significância do modelo são dadas por:

$$\begin{cases} H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_a\colon \exists i:\ \beta_i \neq 0 \end{cases} \qquad (i = 1,2,\ldots,p)$$

onde H_0 representa um modelo que não é estatisticamente significativo o que impossibilita predizer a probabilidade de sucesso a partir das variáveis independentes incluídas no modelo.

A estatística do teste busca comparar a verosimilhança de um modelo nulo ou reduzido, $Logit(\hat{\pi}_j) = \beta_0$, com o modelo completo, $Logit(\hat{\pi}_j) = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_p X_{pj}$ e é dada por:

$$G^2 = -2LL_0 - (-2LL_C) = -2Ln\left[\frac{L_0}{L_C}\right] \sim \chi_p^2$$

onde G^2 segue uma distribuição assimptótica Qui-quadrado com p graus de liberdade, também conhecido como Teste do Rácio de Verosimilhança (Marôco, J., 2021), onde Hair, J. F. Jr et al. (2006) interpreta os seus resultados como o valor mínimo para $-2LL$ é 0, que traduz num ajuste do modelo aos dados, ou seja, quanto menor o valor $-2LL$, melhor o ajuste do modelo.

Observa-se que, concluir que o modelo completo é significativo, permite apenas afirmar que pelo menos uma variável independente incluída no modelo influencia significativamente a variável dependente do modelo ajustado.

Para verificar se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente, recorre-se ao teste de Wald que tem como objetivo testar a significância de cada coeficiente dentro do modelo obtido, ou seja, se o coeficiente é diferente de zero, comparando o valor obtido da estimação de máxima verosimilhança e o seu erro padrão.

A hipótese do teste é dada por:

$$\begin{cases} H_0\colon\ \beta_i = 0 \\ H_a\colon\ \beta_i \neq 0 \end{cases} \qquad i = 1, 2, 3, \ldots, p$$

A estatística do teste segue uma distribuição t-Student, que se aproxima assintoticamente a distribuição $N(0,1)$ quando a dimensão da amostra é grande, que é dada por:

$$T_{Wald_i} = \frac{\hat{\beta}_i}{\widehat{SE}(\hat{\beta}_i)}$$

O ator acrescenta que o IBM SPSS Statistics utiliza o quadrado do T_{Wald_i} que tem distribuição Qui-quadrado assimptótica com 1 grau de liberdade, dada por:

$$\chi^2_{Wald_i} = \left(\frac{\hat{\beta}_i}{\widehat{SE}(\hat{\beta}_i)}\right)^2$$

Relativamente ao critério de decisão, rejeita-se H_0 para cada um dos testes aos β_i quando $valor\text{-}p < \alpha$.

2.5.4 Qualidade do Ajustamento do Modelo

Para testar a qualidade de ajuste do modelo, Hair, J. F. Jr et al. (2006) e Marôco, J., (2021) existem três maneiras, cujas hipóteses estatísticas associadas são:

$$\begin{cases} H_0\text{: O modelo se ajusta ao dados} \\ H_a\text{: O modelo não se ajusta ao dados} \end{cases}$$

A primeira forma de testar a qualidade de ajuste do modelo é recorrendo a estatística de teste *clássica*, onde os dados são agrupados em J células numa tabela resultante dos preditores, conhecido como qui-quadrado de Pearson (Marôco, J., 2021 apud Hosmer & Lomeshow, 2000), dado por:

$$X_P^2 = \sum_{j=1}^{J} \frac{\left(O_j - E_j\right)^2}{E_j}$$

onde O_j e E_j representam o número de sucessos observados e esperados na célula j, respectivamente.

A segunda é dada pela estatística $Deviance$ que segue uma distribuição assimptótica Qui-quadrado com $n - p - 1$ graus de liberdade, dado por:

$$D = -2Ln\left[\frac{L_C}{L_S}\right]$$

onde L_C é a verosimilhança do modelo ajustado (completo ou não) e L_S é a verosimilhança do modelo saturado. O autor acrescenta ainda que, quando a variável dependente é dicotómica e toma valores 0 ou 1, a verosimilhança do modelo saturado é 1, pelo que neste caso, a estatística D assume a expressão:

$$D = -2LL_C$$

A terceira maneira consiste em avaliar a dimensão do efeito do modelo, apresentado pelas diversas medidas do tipo R^2, conhecido como $pseudo - R^2$, variando entre 0 e 1. Hair, J. F. Jr et al. (2006) acrescenta que essas medidas $pseudo - R^2$ são interpretadas de uma maneira parecida com o coeficiente de determinação em regressão múltipla, onde $pseudo - R^2 = 1$ indica ajuste perfeito.

Marôco, J., (2021) de forma muito sintética apresenta o R^2 de Cox e Snell dado por:

$$R_{CS}^2 = 1 - e^{\frac{2[LL_C - LL_0]}{n}}$$

onde a estatística de teste nunca atinge 1, mesmo quando o modelo está ajustado perfeitamente.

Para corrigir foi proposta a estatística seguinte, conhecido como R^2 de Negelkerke dado por:

$$R_N^2 = \frac{R_{CS}^2}{1 - e^{\frac{2LL_0}{n}}}$$

Porém os valores podem não ser interpretáveis como sendo a percentagem da variabilidade da variável dependente que é explicada pelo modelo. (Mittlbock & Schemper, 1996 apud Marôco, J., 2021)

Contudo, Mittlbock & Schemper, (1996 apud Marôco, J., 2021) chegaram a perceber que R^2 de McFadden apresenta uma melhor interpretabilidade, ou seja, R_{MF}^2 pode ser interpretado como a proporção de redução do modelo nulo, relativamente ao modelo completo, isto é, o rácio do ganho de informação estimada pelo modelo completo em comparação com o modelo nulo, com o ganho de informação potencialmente, recuperável por um modelo saturado. E a estatística é dada por:

$$R_{MF}^2 = 1 - \frac{LL_C}{LL_0}$$

2.5.5 Métodos e Critérios de Seleção de Modelos

De cordo com Marôco, J., (2021), os métodos de seleção das variáveis com o poder preditor mais usados são os Método *Forward, Backward* que consistem num conjunto de critérios para a adição ou remoção de covariáveis baseados na significância estatística, comparando modelos com e sem as variáveis em questão de maneira a identificar um pequeno número de modelos suficientemente bons de acordo com determinados critérios.

- **Método Forward**

Os diferentes tipos de método *Forward* usados na regressão logística são baseados no método de selecção *stepwise* que segundo Ferreira, M. C. C. D. S. (2013), baseia-se na seleção automática das variáveis importantes para o modelo, ou seja, é usado para selecionar as variáveis que mais influenciam o conjunto de saída podendo, assim, diminuir o número de variáveis a compor a equação de regressão.

Os tipos de método *Forward* usados são: ***i)*** **Método *Forward* Condicional** – a entrada de uma variável independente no modelo é feita em função da significância estatística "*Score"*[4] do modelo e a remoção de uma variável do modelo é feita em função da significância do teste de rácio de verosimilhanças baseada nas estatísticas condicionais dos coeficientes do modelo; ***ii)*** **Método *Forward* baseado no Rácio de Verosimilhanças** – a entrada de uma variável independente no modelo é feita em função da significância estatística "*Score"* do modelo e a remoção de uma variável do modelo é feita em função da significância do teste de rácio de verosimilhanças baseada nas estimativas parciais de máxima verosimilhança do modelo; e ***iii)*** **Método *Forward* baseado no Teste de Wald** – a entrada de uma variável independente no modelo é feita em função da significância estatística "*Score"* do modelo e a remoção de uma variável do modelo é feita em função da significância do teste de Wald.

- **Método *Backward***

Os diferentes tipos de métodos *Backward* usados na regressão logística consiste na inclusão de todos os preditores de uma só vez na equação, e depois são retirados, um a um, até que se identifiquem os preditores significativos.

Os tipos de método *Backward* usados são: ***i)*** **Método *Backward* Condicional** – são removidas as variáveis cuja a probabilidade do rácio de verosimilhanças baseada nas estatísticas condicionais dos coeficientes do modelo é superior ao $valor\text{-}p$ de remoção selecionado; ***ii)*** **Método *Backward* baseado no Rácio de Verosimilhanças** – é um método semelhante ao anterior, mas o $valor\text{-}p$ de remoção são calculados a partir do teste do rácio de verosimilhança baseado nas estimativas parciais de máxima verosimilhança do modelo; e ***iii)*** **Método *Backward* baseado no Teste de Wald** – a remoção de uma variável do modelo é feita a partir da significância do teste de Wald.

- **Critérios de Seleção do Modelo**

A semelhança dos modelos de regressão linear, a seleção de modelos pode estar baseada em diferentes critérios, como, por exemplo, na escolha de subconjuntos de tamanho pré-

[4] O teste "*Score*", também conhecido como teste de multiplicadores de Lagrange, é um teste a cada um dos coeficientes do modelo, baseado nas derivadas parciais da função verosimilhança L, que também pode ser calculado para todos os coeficientes do modelo em simultâneos, cujo critério de rejeição é dado por $valor\text{-}p \leq \alpha$. Este teste estima o modelo com restrições e avalia o declive da função *log* verosimilhança na restrição.

determinado, ou, então, na comparação de estatísticas com valores tabelados de referência, pois a seleção do modelo centra-se fundamentalmente na qualidade de ajuste e a sua complexidade.

Porém, na regressão categorial, apenas o Critério de Informação de Akaike (AIC) e o Critério de Informação de Bayes (BIC) são mais utilizados devido a natureza dos dados.

- ✓ **Critério de Informação de Akaike (AIC)**

Segundo Ferreira, M. C. C. D. S. (2013), Critério de Informação de Akaike não envolve testes estatísticos e pode ser expresso em função do desvio do modelo, e é baseado na função de verossimilhança. Porém, alerta para um determinado conjunto de dados, o AIC por si só, não tem qualquer significado e é útil quando são comparados diversos modelos.

Para Demétrio, C. G. B., & Zocchi, S. S. (2006), o AIC é dado por:

$$AIC = -2\ln(L) + 2k$$

onde k representa o número de parâmetros no modelo ajustados. Para o AIC, um modelo será melhor se apresentar o menor AIC.

- ✓ **Critério de Informação de Bayesiano (BIC)**

Segundo o Demétrio, C. G. B., & Zocchi, S. S. (2006), o Critério de Informação de Bayesiano (BIC) é dado por:

$$BIC = -2\ln(L) + 2k\ln(n)$$

onde n representa o número de observações. Para o BIC, um modelo será melhor se apresentar o menor AIC.

2.5.6 Diagnóstico de *Outliers* e Classificação por recurso a Regressão Logística

Os pressupostos do modelo ajustado precisam ser validados para que os resultados sejam confiáveis, uma vez que ao ajustar uma equação de regressão aos dados, na maioria das vezes o valor observado não corresponde exatamente ao valor predito. A esta diferença chama-se de resíduos ou variação residual. E ao conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos chama-se de Análise dos Resíduos ou Análise de Diagnóstico.

Na regressão logística, Marôco, J., (2021) apresenta análise de resíduos como uma técnica que permite identificar *outliers* e casos influentes na estimação do modelo através das seguintes medidas: resíduos, *Leverage*, Distancia de Cook e *DfBetas*.

Assim, pode-se dizer que resíduo ou erro é a diferença entre o número de sucessos observados e número de sucessos estimados, dada por:

$$e_j = y_j - \hat{y}_j = y_j - n_j\hat{\pi}_j$$

onde n_j é o número de Observações da célula j e $\hat{\pi}$ é a probabilidade do sucesso estimada na célula j.

- ✓ **Resíduo de Pearson e a *Leverage***

Para obter o resíduo de Pearson, basta dividir o resíduo pela estimativa do desvio padrão dos valores estimados, como se segue:

$$e'_j = \frac{e_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

A estatística de teste baseada nos resíduos de Pearson é denominada por estatística de Qui-quadrado de Pearson, dada por:

$$\chi^2 = \sum_{j=1}^{J}(e'_j)^2$$

onde J é o número de células resultantes do cruzamento das variáveis independentes.

Porém, o resíduo de Pearson não é completamente estandardizado uma vez que a variância depende das observações e suas influências nas estimativas do modelo.

Antes de prosseguir, tem-se alguns conceitos importantes no diagnóstico do modelo, como: i) Um ponto inconsistente é considerado um *outlier* quando tem *leverage* pequeno e resíduo grande; ii) Ponto de alavanca: quando tem medida de *leverage* grande, podendo ser classificado de bom, quando consistente e mau, quando inconsistente; iii) Uma observação influente é aquela cuja omissão do conjunto de dados resulta em mudanças substanciais em certos aspetos do modelo, podendo ser um *outlier*, ou não.

Demétrio, C. G. B., & Zocchi, S. S. (2006) conceitua *leverage* como a distância de uma observação em relação às demais observações.

As influências obtidas nas estimativas do modelo podem ser estimadas pela *leverage*, dada pelos elementos da diagonal da matriz *H* (Hosmer & Lemeshow, 2000 apud Marôco, J., 2021):

$$h_j = n_j\hat{\pi}_j(\mathrm{x}_j)\left[1-\hat{\pi}_j(\mathrm{x}_j)\right]\mathrm{x}'_j(\mathrm{X}'\mathrm{V}\mathrm{X})^{-1}\mathrm{x}'_j$$

onde $\mathrm{x}'_j = \left[1, x_{1j}, x_{2j}, x_{3j}, \ldots, x_{pj}\right]$ é o vector da observação j em todas as variáveis independentes e V e X são as matrizes definidas no subcapítulo 2.5.3.

Com as informações da *leverage*, o resíduo de Pearson estandardizado, com variância constante e igual a 1, é dado pela seguinte equação:

$$r_j = \frac{e'_j}{\sqrt{1-h_j}}$$

Marôco, J. (2021) destaca a importância do resíduo de Pearson estandardizado e a *leverage* no sentido de permitir identificar os *outliers* e avaliar as influências de uma observação no ajustamento do modelo.

- ✓ ***Deviance***

Além do Resíduo de Pearson e a *Leverage*, a ***Deviance*** é outra medida que permite avaliar a influência de uma observação (ou célula) j no ajustamento do modelo, dado por:

$$d_j = Sign(e_j)\sqrt{2\left[y_j \ln\left(\frac{y_j}{n_j\hat{\pi}_j}\right) + (n_j - y_j)\ln\left(\frac{n_j - y_j}{n_j(1-\hat{\pi}_j)}\right)\right]}$$

onde $Sign$ devolve o sinal $(+ \text{ ou } -)$ de e_j.

Marôco, J. (2021) acrescenta ainda dizendo que as observações influentes no modelo são aquelas em que apresentam valores elevados de d_j, que por sua vez permite afirmar que as observações com *Deviance* elevadas são mal especificadas pelo modelo. E a estatística de teste associada é dada por:

$$D = \sum_{j=1}^{n}(d_j)^2$$

Segundo Vaz, F. E. D. C. R. (2020), D tem uma distribuição assimptótica Qui-quadrado com $n-(p+1)$ graus de liberdade, sendo que p é o número de covariáveis do modelo.

- ✓ **Distância de Cook e a $DfBeta$**

Para o mesmo autor, Vaz, F. E. D. C. R. (2020), em 1986, Cook apresentou uma proposta inovadora no diagnóstico de modelos, conhecida como **Distância de Cook**, que propõe avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo ou nos dados, ao invés da avaliação pela remoção individual ou conjunta de pontos. Contudo, Marôco, J. (2021) apresenta uma medida análoga, que poderia se chamar de **Distância de Cook *"análoga"*** que permite estimar a variação dos resíduos quando a observação j é eliminada no ajustamento do modelo, ou seja, estima a influência de uma determinada observação ou célula j na estimação dos coeficientes do modelo.

Devido a essas características, esta medida carrega consigo a *Leverage* e os resíduos de Pearson estandardizados e é dada por (Pregibon, 1981 apud Marôco, J., 2021):

$$DC_j = r_j^2 \frac{h_j}{1 - h_j}$$

onde $\frac{h_j}{1-h_j}$ representa a potencial de influência e mede a distância do ponto k em relação às demais observações e DC_j superiores a 1 indicam observações influentes na estimação dos coeficientes.

Para conhecer a influência de cada observação na estimação de cada um dos coeficientes de regressão tem-se de recorrer a $\boldsymbol{DfBeta}$, dada por:

$$DfBeta_{ij} = \hat{\beta}_i - \hat{\beta}_{i(-j)}$$

onde $\hat{\beta}_i$ é a estimativa do coeficiente de regressão, ajustado com todas as observações e $\hat{\beta}_{i(-j)}$ sem a observação j.

Porém, Demétrio, C. G. B., & Zocchi, S. S. (2006) alerta dizendo que $DfBeta$ não tem interpretação simples, onde acrescenta ainda dizendo que Cook & Weisberg (1982) propuseram curvas empíricas para o estudo da mesma.

Ainda assim, Marôco, J. (2021) apresenta observações influentes como aquelas que apresentam $DfBeta > 2$ ou superior a $2 \times \sqrt{\frac{p+1}{n}}$, onde p é o número de coeficientes no modelo e n a dimensão da amostra.

A alteração provocada no valor ajustado pela retirada da observação i é dada pela $DFFitS$ (ver Demétrio, C. G. B., & Zocchi, S. S., 2006)

- ✓ **Classificação do Modelo**

Uma da vantagem da Regressão Logística é a possibilidade de estimar o modelo em termos de probabilidades, após obter o $Logit$. A probabilidade de cada uma das observações j pertencerem ao grupo de sucesso (1), em comparação ao grupo de referência, insucesso (0), é modelado através da seguinte fórmula:

$$\hat{\pi}_j = \frac{e^{\beta_0+\beta_1 X_{1j}+\beta_2 X_{2j}+\cdots+\beta_p X_{pj}}}{1+e^{\beta_0+\beta_1 X_{1j}+\beta_2 X_{2j}+\cdots+\beta_p X_{pj}}}$$

ou

$$\hat{\pi} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1j}+\beta_2 X_{2j}+\cdots+\beta_p X_{pj})}}$$

onde, para Marôco, J. (2021), o sujeito é classificado no grupo 1, se a probabilidade de $Y_j = 1$ for superior a $0{,}5$, caso contrário é classificado no grupo 0. Porém, o autor adverte que essa classificação é enviesada a favor da maior classificação. Contudo, sugere fracionar a amostra de maneira a ajustar o modelo com uma parte e outra para o classificar ou usar o procedimento *Jackknife* que consiste em ajustar o modelo se a observação do grupo que se pretende realizar a previsão e depois usar esse modelo para classificar a observação deixada de fora no ajustamento do modelo.

Ora, para certificar ou avaliar a qualidade de classificação feita, deve-se comparar a percentagem global de classificação correta obtida com o modelo e a percentagem proporcional de classificação correta por acaso, dada por:

$$CCPA = \sum_{i=1}^{k} \left(\frac{C_i}{n}\right)^2 \times 100$$

onde $CCPA$ é a classificação correta proporcional por acaso dado em termos percentais, C_i é o número de sujeitos observados em cada uma das k classes de variáveis dependente e n é o tamanho da amostra global.

A partir deste cálculo é possível afirmar que se as percentagens de casos classificados corretamente pelo modelo forem superiores em pelo menos 25% à $CCPA$, considera-se que o modelo tem boas capacidades preditivas.

Ora, para avaliar a eficiência dessa classificação, tem-se de recorrer a *sensibilidade* e *especificidade* do modelo.

De salientar que estas duas noções, foram inicialmente utilizados na medicina, com a finalidade de classificar os pacientes em função dos testes de acordo com a patologia.

Assim, a ***sensibilidade*** é entendida como a probabilidade de o teste gerar um resultado positivo, dado que o indivíduo é realmente portador da "patologia", ou seja, o sujeito tem a característica que se quer modelar e o modelo prevê corretamente essa característica.

$$Sensibilidade = P\left[\hat{Y} = 1 | Y = 1\right]$$

A ***especificidade*** probabilidade de o teste gerar um resultado negativo quando o indivíduo não é portador da "patologia", ou seja, o sujeito não tem a característica que se quer modelar e o modelo prevê que o sujeito não tem essa característica.

$$Especificidade = P\left[\hat{Y} = 0 | Y = 0\right]$$

As pontuações relativas as avaliações das capacidades preditivas, segundo Marôco, J. (2021), podem ser agrupadas de acordo com a seguinte tabela:

Tabela 6 – Capacidades Preditivas

	Medíocre	**Razoáveis**	**Boa**
Pontuações	$< 50\%$	$\geq 50\%$ e $\leq 80\%$	$> 80\%$

O autor sugere ainda, uma outra medida da capacidade do modelo para discriminar os sujeitos de acordo com as características de interesse chamada de **Curva ROC (*Receiber Operaing Characteristic*)**.

Segundo Ferreira, M. C. C. D. S. (2013) as curvas ROC foram desenvolvidas no ramo das comunicações como uma forma de demonstrar as relações entre sinal-ruído, onde Nakamura, K. G. (2013) é uma ferramenta poderosa para medir e especificar problemas no desempenho do diagnóstico por permitir estudar a variação da sensibilidade e especificidade para diversos valores de corte.

O valor de corte ou ponto de corte ou *cut off* é o valor que pode ser selecionado arbitrariamente pelo investigador entre os valores possíveis para a variável de decisão, acima da qual o paciente é classificado positivo e abaixo do qual é classificado como negativo.

A curva ROC é um gráfico de sensibilidade, ou seja, o percentual de indivíduos que foram classificados corretamente como sucesso, versus um menos a especificidade, ou seja, o percentual de indivíduos que foram classificados erroneamente como sucesso.

A seguir tem-se um exemplo da Curva de ROC extraído de Nakamura, K. G. (2013).

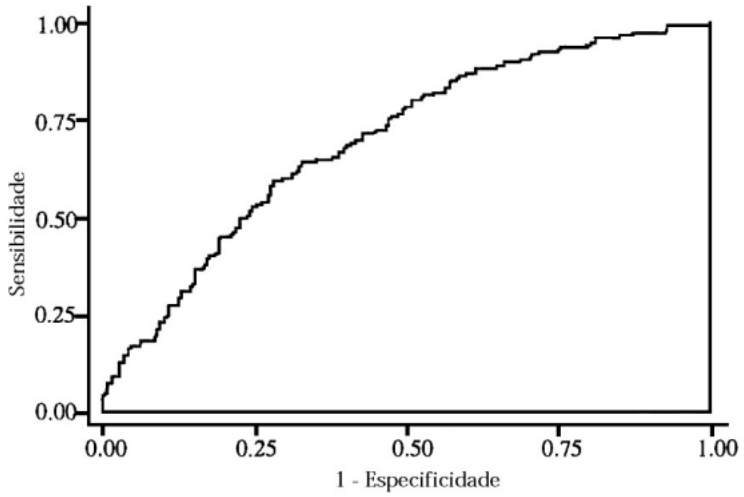


Figura 1 – Exemplo da Curva ROC

Vaz, F. E. D. C. R. (2020) acrescenta dizendo que a curva ROC é muito utilizada na medicina para avaliar o desempenho de diagnósticos médicos, em que se procura identificar a presença ou ausência de certa doença, associando a respetiva probabilidade de erro.

A área abaixo da curva ROC, que pode variar de 0 a 1, fornece uma medida de discriminação entre os indivíduos que apresentaram a característica de interesse versus aqueles que não apresentaram, ou seja, ela permite avaliar a capacidade preditiva de um modelo de regressão logística binária.

A sua classificação do poder de discriminação é a seguinte (Marôco, J., 2021 apud Hosmer & Lomeshow, 2000):

Quadro 3 – Classificação do poder de discriminação - Curva de ROC

Área sob a curva ROC	Poder discriminatório no modelo
$\mathbf{0,5}$	Sem poder discriminatório
$]\mathbf{0,5};\ \mathbf{0,7}[$	Discriminação fraca
$]\mathbf{0,7};\ \mathbf{0,8}[$	Discriminação aceitável
$]\mathbf{0,8};\ \mathbf{0,9}[$	Discriminação boa
$\geq \mathbf{0,9}$	Discriminação excecional

2.6 Regressão Logística Multinomial

No subcapítulo anterior, trabalhou-se o modelo onde a variável dependente é nominal dicotómica ou binária. Porém, nas análises de dados em ciências sociais, depara-se com variáveis do tipo nominal com mais categorias, assemelhando-se a escala de *likert*, onde estas são policotómicas, ou seja, assumem mais de dois valores como 0 - Norte, 1 - Este, 2 - Nordeste, 3 - Centro, entre outros níveis.

Diante destes dados, o modelo de regressão logística binária, vê-se limitado dando oportunidade à **Regressão Logística Multinomial**.

Embora não sendo muito utilizado, o modelo de regressão logística multinomial é uma extensão do modelo de regressão logística binária e ambas são aplicadas nos casos em que a variável dependente é medida em escala nominal.

Tomando em consideração uma variável dependente que apresenta 3 (três) classe, *i.e.* 0 - Norte, 1 - Sul, 2 – Centro, a probabilidade da vaiável dependente tomar o valor de qualquer uma das 3 classes é dada por (Marôco, J., 2021 apud Hosmer & Lomeshow, 2000):

$$P[Y=0|X] = \frac{e^{\beta_{00}+\beta_{01}X_1+\beta_{02}X_2+\cdots+\beta_{0p}X_p}}{e^{\beta_{00}+\beta_{01}X_1+\cdots+\beta_{0p}X_p} + e^{\beta_{20}+\beta_{21}X_1+\cdots+\beta_{2p}X_p} + e^{\beta_{10}+\beta_{11}X_1+\cdots+\beta_{1p}X_p}}$$

$$P[Y=1|X] = \frac{e^{\beta_{10}+\beta_{11}X_1+\beta_{12}X_2+\cdots+\beta_{1p}X_p}}{e^{\beta_{00}+\beta_{01}X_1+\cdots+\beta_{0p}X_p} + e^{\beta_{20}+\beta_{21}X_1+\cdots+\beta_{2p}X_p} + e^{\beta_{10}+\beta_{11}X_1+\cdots+\beta_{1p}X_p}}$$

$$P[Y=2|X] = \frac{e^{\beta_{20}+\beta_{21}X_1+\beta_{22}X_2+\cdots+\beta_{2p}X_p}}{e^{\beta_{00}+\beta_{01}X_1+\cdots+\beta_{0p}X_p} + e^{\beta_{20}+\beta_{21}X_1+\cdots+\beta_{2p}X_p} + e^{\beta_{10}+\beta_{11}X_1+\cdots+\beta_{1p}X_p}}$$

ou na forma matricial

$$P[Y=0|X] = \frac{e^{X\beta_0}}{e^{X\beta_0} + e^{X\beta_1} + e^{X\beta_2}}$$

$$P[Y=1|X] = \frac{e^{X\beta_1}}{e^{X\beta_0} + e^{X\beta_1} + e^{X\beta_2}}$$

$$P[Y=2|X] = \frac{e^{X\beta_2}}{e^{X\beta_0} + e^{X\beta_1} + e^{X\beta_2}}$$

Assim, percebe-se que este modelo consiste num conjunto de k modelos logísticos. Porém, por existir mais de uma combinação de que conduzem à mesmas probabilidades, o sistema é indeterminado.

Para corrigir, Marôco, J. (2021) sugere normalizar o sistema a uma categoria da variável dependente e um dos coeficientes referentes a uma das classes tem de ser zero ($\beta_0 = 0$), possibilitando o modelo ser reescrito da seguinte forma:

$$P[Y = 0|X] = \frac{1}{1 + e^{X\beta_1} + e^{X\beta_2}}$$

$$P[Y = 1|X] = \frac{e^{X\beta_1}}{1 + e^{X\beta_1} + e^{X\beta_2}}$$

$$P[Y = 2|X] = \frac{e^{X\beta_2}}{1 + e^{X\beta_1} + e^{X\beta_2}}$$

Considerando a normalização anterior, o modelo $Logit$ é dado por:

$$Ln\left[\frac{P[Y = 1|X]}{P[Y = 0|X]}\right] = X\beta_1 \quad | \quad Ln\left[\frac{P[Y = 2|X]}{P[Y = 0|X]}\right] = X\beta_2$$

O modelo é ajustado pelo método de máxima verosimilhança e as medidas da qualidade de ajuste são as mesma da regressão logística binária.

O autor acrescenta ainda que o rácio das chances ($Odds\ ratio$) é calculado para cada uma das $k - 1$ classes relativamente a classe de referência 0. Para a classe $c(c = 1, 2, \dots, k - 1)$ da variável dependente relativamente à variável independente $i(i = 1, 2, \dots, p)$ o rácio das chances é dado por:

$$OR(c, 0|X_i) = e^{X\beta_{ci}} = \frac{P(Y = c|X_i = x_i + 1)/P(Y = 0|X_i = x_i + 1)}{P(Y = c|X_i = x_i)/P(Y = 0|X_i = x_i)}$$

De salientar que, diferente da regressão logística binaria, na regressão logística multinomial, o rácio das chances é sempre relativo à classe de referência.

2.6.1 Classificação por recurso a Regressão Logística Multinomial

A semelhança da regressão logística binária, é possível classificar o sujeito na classe de referência, porém neste tipo de regressão, um determinado sujeito pode ser classificado em mais de duas classes de variável dependente.

Assim, para um sujeito j, a classificação é dada por:

$$P(Y = 0|\mathrm{X}_j) = \frac{1}{1 + \sum_{i=1}^{k} e^{\beta_{i0} + \beta_{i1}X_{1j} + \beta_{i2}X_{2j} + \cdots + \beta_{ip}X_{pj}}}$$

E se for a probabilidade de observar uma classe c em comparação à classe de referência no sujeito j, a classificação é dada por:

$$P(Y_j = c|\mathrm{X}_j) = \frac{e^{x_j\beta_c}}{1 + \sum_{i=1}^{k} e^{x_j\beta_i}} = \frac{e^{\beta_{c0}+\beta_{c1}X_{1j}+\beta_{c2}X_{2j}+\cdots+\beta_{cp}X_{pj}}}{1 + \sum_{i=1}^{k} e^{\beta_{i0}+\beta_{i1}X_{1j}+\beta_{i2}X_{2j}+\cdots+\beta_{ip}X_{pj}}}$$

De salientar que o sujeito é classificado na classe onde a sua probabilidade de ocorrência for maior.

Os testes de diagnóstico para este modelo são os mesmos aplicados para a regressão logística binária. Porém, para avaliar a qualidade do teste, Pestana, M. H., & Gageiro, J. N. (2020) recomenda os **rácios de verosimilhança** (*likelihood ratios*), LR, que consiste em informar sobre o resultado do teste aumentar ou diminuir a verosimilhança da caraterística de interesse ou resultado positivo ou de referência.

Se $LR > 1$, significa que o teste de diagnóstico aumenta a verosimilhança da caraterística de interesse;

Se $0 < LR < 1$, significa que o teste de diagnóstico diminui a verosimilhança da caraterística de interesse;

Se $LR \approx 1$, significa que o teste de diagnóstico não altera substancialmente a verosimilhança da caraterística de interesse.

O rácio de verosimilhança pode ser positivo $(+\,LR)$ ou negativo $(-\,LR)$, sendo:

$$+\,LR = \frac{Sensibilidade}{1 - Especificidade} \qquad | \qquad -\,LR = \frac{1 - Sensibilidade}{Especificidade}$$

CAPÍTULO III – APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Neste capítulo é apresentado a aplicação dos modelos de regressão linear e de regressão categorial com o objectivo de identificar as variáveis socioeconómicas ou regionais e académicas que influenciam a admissão dos estudantes na USTP.

Na análise dos dados, começou-se com a análise descritiva dados, para conhecer a constituição da base de dados, no que respeita a distribuição das informações recolhidas. De seguida, procurou-se testar algumas hipóteses subjacentes às variáveis que estão na constituição da base de dados, recorrendo a alguns testes de comparação paramétrica e não paramétrica.

Após conhecer a estrutura e constituição da base de dados, procedeu-se à construção do modelo de regressão, verificando todos os prossupostos de forma a poder realizar com precisão a previsão e inferência.

Estas análises foram realizadas com recursos ao *software* IBM SPSS (*Statistical Package for the Social Sciences*) versão 28 para Windows, Microsoft 365 (Excel).

Devido a heterogeneidade das disciplinas que compõem os diversos cursos, a base de dados terá a seguinte configuração, dentre as particularidades:

- Base de Dados I: incluem os dados dos candidatos admitidos que realizaram as disciplinas de Matemática, Biologia e Física/Química;
- Base de Dados II: incluem os dados dos candidatos admitidos que realizaram as disciplinas de História, Sociologia/Psicologia e Direito.

Com esta característica, serão analisadas, ora em simultâneo (num tópico/subcapítulo) ora em separados, de forma a destacar os aspectos relevantes.

3.1 Apresentação dos Dados

A aplicação prática se concentrou no estudo do ingresso de estudantes finalistas do ensino secundário em diferentes cursos oferecidos pela Universidade de São Tomé e Príncipe (USTP) no período entre os anos letivos de 2021/2022 e 2023/2024.

A base de dados é composta por 14 variáveis, a saber: i) Sexo, ii) Idade, iii) Residência, iv) Escola de origem, v) Média Final do Secundário, vi) Ano de Conclusão do Secundário, vii) Curso Matriculado na USTP, viii) nota da disciplina de Língua Portuguesa, x) de História, xi) de Física/Química, xii) de Biologia, xiii) de Direito e xiv) de Sociologia/Psicologia.

Destaca-se que as variáveis v), viii), ix) e x) são de importância vital para a admissão de um estudante em um curso da USTP, uma vez que a estrutura curricular do segundo ciclo do ensino secundário em São Tomé e Príncipe não permite que um estudante inclua no seu currículo as disciplinas viii) e x), nem tampouco permite que aqueles que estudam viii) também estudem xi), xii) e xiii).

3.1.1 Características das variáveis qualitativas

i) **Curso Inscrito**

A variável "**Curso Inscrito**" é do tipo qualitativo policotômico, agrupando diversos cursos organizados por departamentos, a saber:

- **Departamento de Ciências da Natureza, da Vida e do Ambiente (DCNVA)**: Licenciaturas em Biologia e Enfermagem;
- **Departamento de Língua e de Ciências Humanas e Sociais (DL&CHS)**: Licenciaturas em Língua Francesa, Língua Inglesa, Língua Portuguesa, Gestão Cultural, Português e Alemão, História, Ciências da Comunicação e Direito;
- **Departamento de Ciências Económica e de Ciências Exactas e Engenharias (DCE&EE)**: Licenciaturas em Matemática, Economia, Contabilidade e Auditoria, Turismo, Gestão de Empresa, Gestão e Administração de Serviços de Saúde, Engenharia Informática e Química;
- **Departamento de Ciências da Educação (DCE)**: Licenciaturas em Ciências da Educação, Educação Básica e Educação de Infância.

ii) **Residência**

A variável residência refere-se aos distritos em que os estudantes vivem. Ela abrange os distritos de São Tomé e a Região Autónoma do Príncipe (RAP). No entanto, devido ao número reduzido de estudantes residentes em dois desses distritos, os locais/distritos de residência dos alunos foram reorganizados da seguinte forma: Água Grande, Cantagalo, Lobata, Lembá, Mé-Zóchi e Caué/RAP (os distritos de Caué e RAP foram agrupados).

Assim sendo, com base nos resultados apresentados na Tabela 7, constactou-se que os distritos mais populosos do país (Água Grande e Mé-Zochi), congrega maior número de estudantes, com mais de 71% e 66% nas Base de Dados I e Base de Dados II respectivamente.

Constactou-se também, um ligeiro aumento dos estudantes dos distritos de Cantagalo e Lembá na Base de Dados II em comparação com a Base de Dados I.

Tabela 7 – Distribuição por Distrito/Residência

	N	%
Água Grande	199	36,1%
Mé-Zochi	195	35,4%
Cantagalo	40	7,3%
Lobata	39	7,1%
Lembá	58	10,5%
Caué&RAP	20	3,6%

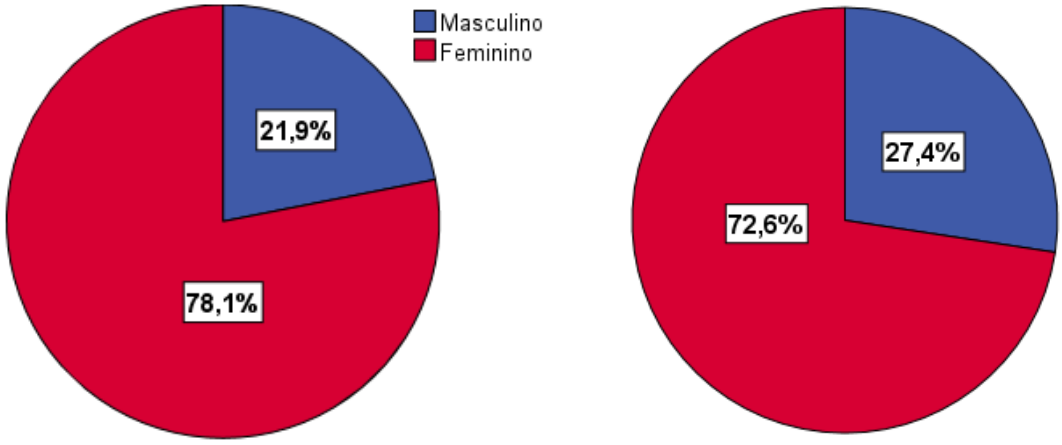
Base de Dados I

	N	%
Água Grande	180	31,8%
Mé-Zochi	208	36,7%
Cantagalo	72	12,7%
Lobata	23	4,1%
Lembá	67	11,8%
Caué&RAP	16	2,8%

Base de Dados II

iii) **Sexo**

A co-variável **Sexo**, do tipo binária, é composta pelo Masculino e Feminino, onde se verificou presença significativa de entrada das senhoras no ensino superior, com uma percentagem acima dos 70%, conforme ilustrado na Figura 2.



Base de Dados I Base de Dados II

Figura 2 – Distribuição por Sexo

iv) **Escola Proveniente**

A **Escola Proveniente**, também do tipo policotómica, está composta pelas escolhas secundárias, politécnicas ou de formação profissional do país, a saber: Liceu Nacional, Escola Secundária Maria Manuela Margarida, Escola Secundária de Santana, Escola Secundária Sun Mé Xinhô, Escola Secundária de Neves e Outras (Escolas e Centros de Formação Profissional do País).

Na Tabela 8, observamos que o Liceu Nacional possui a maior representatividade nos cursos oferecidos pela USTP, seguido pela Escola Secundária Maria Manuela Margarido.

Tabela 8 – Distribuição por Escolas ou Centros de Formação

	N	%
Liceu Nacional	260	47,2%
E. S. Maria Manuela Margarida	81	14,7%
E. S. Santana	19	3,4%
E. S. Neves	25	4,5%
E. S. Mé Xinhô	44	8,0%
Outras Escolas	122	22,1%

Base de Dados I

	N	%
Liceu Nacional	228	40,3%
E. S. Maria Manuela Margarida	110	19,4%
E. S. Santana	80	14,1%
E. S. Neves	62	11,0%
E. S. Mé Xinhô	17	3,0%
Outras Escolas	69	12,2%

Base de Dados II

Percebe-se também a baixa presença de estudantes da Escola Secundária de Santana na Base de Dados I (3,4%) e da Escola Secundária Mé Xinhô na Base de Dados II (3%).

Vale destacar a presença significativa de estudantes provenientes de escolas politécnicas e centros de formação, que buscam prosseguir seus estudos superiores na USTP.

v) **Ano de Conclusão do Ensino Secundário**

Na Figura 3, nota-se uma quantidade significativa de estudantes que ingressaram na USTP pelo menos três anos após concluírem o ensino secundário.

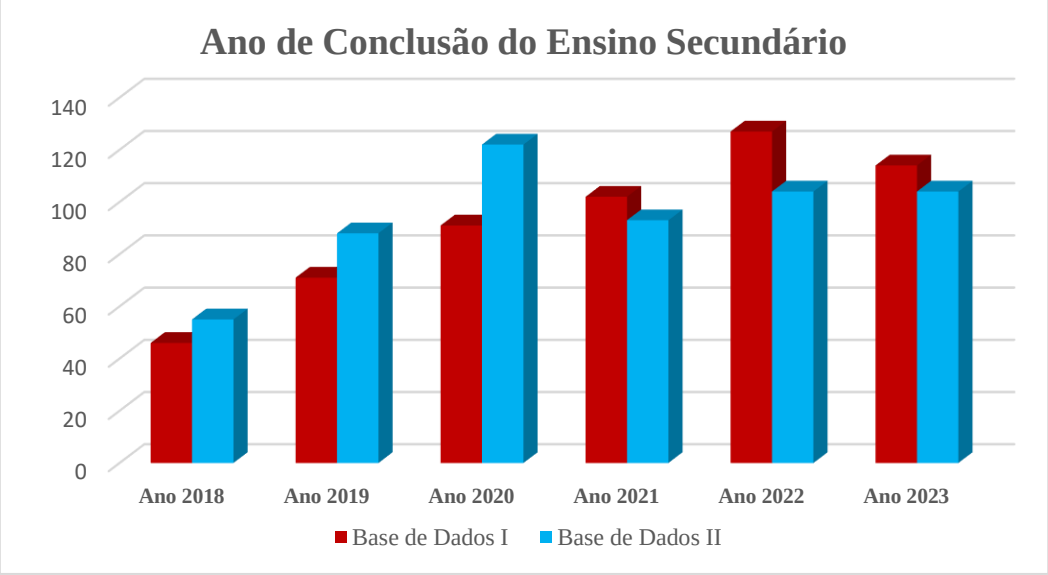


Figura 3 – Distribuição dos Estudantes por ano de conclusão do curso

Nota-se também que no ano lectivo 2020/2021, a USTP atraiu um maior número de estudantes formados em 2020, na área de letras, em comparação com os anos seguintes.

Além disso, observa-se que nos anos de 2018 a 2020, há uma predominância de estudantes na Base de Dados II em comparação com a Base de Dados I, enquanto nos anos seguintes essa situação é invertida.

3.1.2 Características das variáveis quantitativas

As variáveis métricas incluídas na base de dados são Idade, Média Final do 12º Ano e as notas de algumas disciplinas, nomeadamente Matemática, Língua Portuguesa, Biologia, Física/Química, Direito, Economia/Gestão e Sociologia/Psicologia.

Com base na Tabela 9, observa-se que a idade média dos candidatos estão em torno dos 21 anos com um desvio-padrão de 5 anos, sendo o estudante mais velho com 54 anos.

Quanto às notas, destaca-se a presença de excelentes estudantes, com média final do ensino secundário acima de 16 valores, particularmente nas disciplinas de Matemática, Direito e Física/Química, onde alcançaram 19 valores. No entanto, a média das notas nessas disciplinas fica abaixo de 12,5 valores, com uma variação de cerca de 2 valores.

Além disso, observa-se uma predominância de distribuição assimétrica positiva em todas as variáveis métricas desta Base de Dados.

Tabela 9 – Estatísticas das variáveis métricas da Base de Dados I

Estatísticas

		Idade	Média Final do 12º Ano	Matemática	L. Portuguesa	Biologia	Física/Química
N	Válido	551	551	551	551	551	551
	Omisso	0	0	0	0	0	0
Média		21,25	12,4988	10,9764	11,5014	12,2831	12,4319
Mediana		20,00	12,0000	10,0000	11,0000	12,0000	12,0000
Moda		18	12,00	10,00	10,00	11,00	10,00
Desvio Padrão		4,226	1,52227	1,54136	1,57887	2,04310	2,14018
Variância		17,861	2,317	2,376	2,493	4,174	4,580
Assimetria		2,574	,959	1,893	1,059	,915	,754
Curtose		10,411	,470	3,636	,592	,139	-,144
Mínimo		17	10,00	10,00	10,00	10,00	10,00
Máximo		54	18,00	19,00	17,00	19,00	19,00
Percentis	25	18,00	11,0000	10,0000	10,0000	11,0000	11,0000
	75	23,00	13,0000	12,0000	12,0000	14,0000	14,0000

Com relação à Base de Dados II, cujas estatísticas apresentam-se na Tabela 10, a média das idades é de 23 anos, sendo 19 anos a idade mais frequente, com uma variabilidade de 5 anos em torno do valor central (desvio padrão de 5). Observa-se que a maioria dos estudantes são jovens, com 75% deles tendo idades iguais ou inferiores a 25 anos.

Exceto na disciplina de Psicologia/Sociologia, a média das notas das disciplinas está abaixo de 12 valores, apresentando uma ligeira variabilidade (desvio padrão inferior a 2 valores).

Nota-se também que 75% das notas nas disciplinas de Língua Portuguesa e História são iguais ou inferiores a 11 e 12 valores, respectivamente.

Tabela 10 – Estatísticas das variáveis métricas da Base de Dados II

Estatísticas

		Idade	Média Final do 12º Ano	L. Portuguesa	História	Direito	Psicologia e Sociologia
N	Válido	566	566	566	566	566	566
	Omisso	0	0	0	0	0	0
Média		22,90	11,8498	10,7208	11,1678	11,9470	12,5954
Mediana		22,00	12,0000	10,0000	11,0000	12,0000	12,0000
Moda		19	12,00	10,00	10,00	11,00	12,00
Desvio Padrão		5,085	,93149	1,04745	1,31503	1,52409	1,57320
Variância		25,861	,868	1,097	1,729	2,323	2,475
Assimetria		1,699	,883	1,913	1,164	,674	,498
Curtose		3,083	1,274	4,760	1,194	,284	,106
Mínimo		17	10,00	10,00	10,00	10,00	10,00
Máximo		46	16,00	16,00	16,00	19,00	18,00
Percentis	25	19,00	11,0000	10,0000	10,0000	11,0000	11,0000
	75	25,00	12,0000	11,0000	12,0000	13,0000	14,0000

3.1.3 Testes de Comparação – Análise Inferencial

As bases de dados são compostas por informações de diversas escolas secundárias e incluem alunos de ambos os sexos. A USTP não possui a capacidade de admitir todos os candidatos, resultando em uma maior probabilidade de admissão e matrícula para os alunos com as melhores médias.

Portanto, é relevante testar algumas hipóteses relacionadas aos candidatos que efetivaram a matrícula, considerando suas médias de conclusão do ensino secundário por sexo e por escola. Isso permitirá entender a distribuição dos melhores alunos e identificar as escolas com as maiores médias.

- ✓ **H1** (Hipótese 1): As médias do ensino secundário dos candidatos do sexo masculino são as mesmas das de sexo feminino?

Nesta hipótese, busca-se comparar duas amostras utilizando o teste t-Student ou o teste U de Mann-Whitney, caso o pressuposto de normalidade não seja verificado.

A hipótese de normalidade para os grupos (sexo masculino e sexo feminino) na Base de Dados I foi rejeitada pelo teste de Kolmogorov-Smirnov ($D_M(124) = 0{,}211; valor\text{-}p < 0{,}01$ e $D_F(442) = 0{,}235; valor\text{-}p < 0{,}01$, ver Apêndice A). De forma semelhante, a hipótese foi

rejeitada na Base de Dados II, apresentando os seguintes resultados para o teste de Kolmogorov-Smirnov: ($D_M(151) = 0{,}284; valor\text{-}p < 0{,}01$ e $D_F(400) = 0{,}209; valor\text{-}p < 0{,}01$, ver Apêndice B).

No entanto, cada grupo contém mais de 50 estudantes, com a Base de Dados I contendo 124 estudantes do sexo masculino e 442 do sexo feminino, e a Base de Dados II contendo 151 estudantes do sexo masculino e 400 do sexo feminino. Tendo em consideração o Teorema do Limite Central, prosseguiremos com a análise, tomando as devidas precauções, dado que as dimensões dos grupos não são iguais ou semelhantes.

Embora reconhecido o tipo de teste a aplicar, é indispensável a verificação da hipótese de homogeneidade de variância. Assim, constatou-se que em ambas Bases de Dados a homogeneidade de variância não foi rejeitada a 5% de significância (Base de Dados I: teste de Levene: $T = 0{,}157; valor\text{-}p = 0{,}693 > 0{,}05$, ver Apêndice I; Base de Dados II: $T = 0{,}970; valor\text{-}p = 0{,}325 > 0{,}05$, ver Apêndice J).

Prosseguindo com o teste, verificou-se que não existem diferenças significativas na média do ensino secundário entre candidatos do sexo masculino e feminino na Base de Dados I, pois a hipótese nula (H_0) não foi rejeitada para $\alpha = 0{,}05$ ($t_{(549)} = -0{,}256; valor\text{-}p = 0{,}798 > \alpha$, ver Apêndice I). Da mesma forma, essa diferença também não foi observada na Base de Dados II, pois a hipótese nula (H_0) não foi rejeitada para $\alpha = 0{,}05$ ($t_{(564)} = 1{,}817; valor\text{-}p = 0{,}070 > \alpha$, ver Apêndice J).

- ✓ **H2** (Hipótese 2): As médias das disciplinas de Matemática e de História dos candidatos do sexo masculino são as mesmas das de sexo feminino?

Ao investigar possíveis diferenças nas médias das disciplinas de Matemática entre candidatos do sexo masculino e feminino, constatou-se que não há diferenças significativas. Isso é evidenciado pelo fato de que a hipótese nula (H_0) não foi rejeitada para $\alpha = 0{,}05$ ($t_{(549)} = -0{,}213; valor\text{-}p = 0{,}832 > 0{,}05 = \alpha$, ver Apêndice I).

Por outro lado, ao analisar as médias das notas da disciplina de História, a hipótese de nulidade é rejeitada ao mesmo nível de significância ($t_{(185{,}64)} = 2{,}537; valor\text{-}p = 0{,}012 < \alpha$, ver Apêndice J) indicando a existência de diferença significativa entre as médias das notas da disciplina de História entre candidatos do sexo masculino e feminino. Este resultado demonstra que a média das notas da disciplina de História dos candidatos do sexo masculino é significativamente superior à dos candidatos do sexo feminino.

É importante destacar que, para a disciplina de História, a hipótese de homogeneidade de variância foi rejeitada (teste de Levene: $T = 5{,}029; valor\text{-}p = 0{,}025 < 0{,}05$, ver Apêndice J). Em contraste, para a disciplina de Matemática, a homogeneidade de variância foi confirmada (teste de Levene: $T = 1{,}193; valor\text{-}p = 0{,}275 > 0{,}05$, ver Apêndice I).

- ✓ **H3** (Hipótese 3): As médias do ensino secundário dos candidatos são iguais nas diferentes Escolas?

Nesta análise, temos como objetivo comparar mais de duas amostras utilizando o teste ANOVA ou o teste de Kruskal-Wallis, dependendo da verificação dos pressupostos de normalidade e homogeneidade de variâncias.

Na Base de Dados I, constatou-se a rejeição da hipótese de normalidade em todos os grupos pelo teste de Kolmogorov-Smirnov ($valor\text{-}p < 0{,}01$, ver Apêndice A). Entretanto, é importante notar que três dos seis grupos possuem uma amostra com mais de 50 observações.

Quanto à hipótese de homogeneidade de variância, ela não foi rejeitada (teste de Levene baseado na mediana, $T = 1{,}142; valor\text{-}p = 0{,}337 > 0{,}05$, ver Apêndice G).

Da mesma forma, na Base de Dados II, a normalidade foi rejeitada em todos os grupos pelo teste de Kolmogorov-Smirnov ($valor\text{-}p < 0{,}01$, ver Apêndice B), sendo que nesta base de dados temos apenas um grupo com menos de 50 observações.

Em relação à homogeneidade de variância, ela também não foi rejeitada (teste de Levene baseado na mediana: $T = 0{,}686; valor\text{-}p = 0{,}634 > 0{,}05$, e o teste de Levene baseado na média: $T = 0{,}469; valor\text{-}p = 0{,}799 > 0{,}05$, ver Apêndice H).

Diante da robustez dos testes paramétricos, optou-se pela aplicação da ANOVA, onde foi possível constatar diferenças significativas na média de conclusão do ensino secundário em pelo menos duas escola da Base de Dados I, uma vez que a hipótese nula foi rejeitada para $\alpha = 0{,}05$, conforme confirmado na Tabela 11.

Na Tabela 12, é possível verificar as escolas nas quais existem diferenças nas médias de conclusão do ensino secundário. É relevante notar que há uma significância marginal na discrepância das médias de conclusão do ensino secundário entre o Liceu Nacional e a Escola Secundária Mé Xinhô.

Tabela 11 – Tabela da ANOVA a um Factor (Base de Dados I)

ANOVA

Média Final do 12º Ano

	Soma dos Quadrados	df	Quadrado Médio	F	Sig.
Entre Grupos	44,548	5	8,910	3,948	,002
Nos grupos	1229,974	545	2,257		
Total	1274,523	550			

Tabela 12 – Testes de Comparação Múltipla (Base de Dados I).

Comparações múltiplas

Variável dependente: Média Final do 12º Ano

	(I) Escola Proveniente	(J) Escola Proveniente	Diferença média (I-J)	Erro Padrão	Sig.
Scheffe	E. S. Mé Xinhô	Liceu Nacional	-,80112	,24489	,059
		E. S. Santana	-1,64833	,41240	,007

Realizando o mesmo teste, utilizando a estatística de Kruskal-Wallis, um teste não paramétrico, chegamos às mesmas conclusões: a hipótese nula é rejeitada para $\alpha = 0{,}05$ (estatística de teste de Kruskal-Wallis, $H = 21{,}020; valor\text{-}p < \alpha$).

As diferenças nas médias de conclusão do ensino secundário entre as escolas são evidenciadas na figura seguinte, no qual a significância foi ajustada pela correção de Bonferroni.

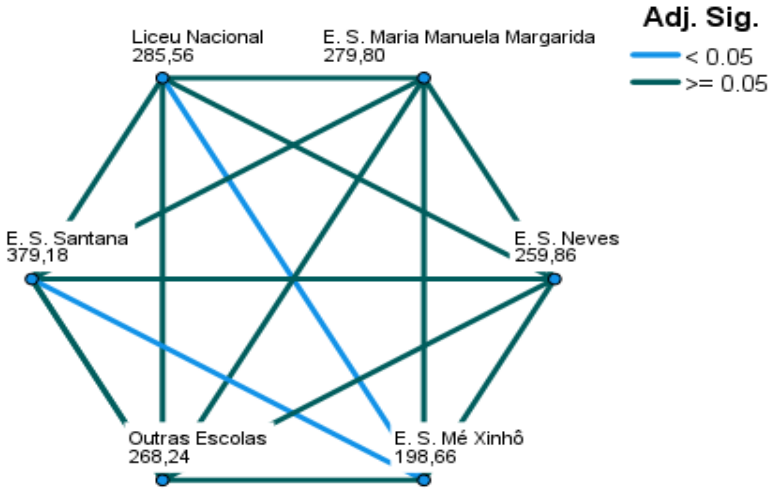


Figura 4 – Comparações por Método *Pairwise* de Escola Proveniente (Base de Dados I)

Em relação a outra base de dados, a Base de Dados II, observou-se que há diferenças significativas na média de conclusão do ensino secundário de pelo menos uma escola, pois a hipótese nula foi rejeitada para $\alpha = 0{,}05$, como é demonstrado na Tabela 13.

Tabela 13 – Tabela da ANOVA a um Factor (Base de Dados II)

ANOVA

Média Final do 12º Ano

	Soma dos Quadrados	df	Quadrado Médio	F	Sig.
Entre Grupos	12,400	5	2,480	2,907	,013
Nos grupos	477,835	560	,853		
Total	490,235	565			

No entanto, o teste de comparação múltipla de Scheffé revela uma significância marginal nas diferenças encontradas nas médias de conclusão do ensino secundário entre a Escola Secundária de Santana e Neves & Escola Secundária de Santana e Liceu Nacional, como demonstrado na Tabela 14.

Tabela 14 – Testes de Comparação Múltipla (Base de Dados II).

Comparações múltiplas

Variável dependente: Média Final do 12º Ano

Scheffe

(I) Escola Proveniente	(J) Escola Proveniente	Diferença média (I-J)	Erro Padrão	Sig.
E. S. Santana	Liceu Nacional	,38246	,12003	,073
	E. S. Neves	,50484	,15630	,066

Dessa forma, ao realizar o teste de Kruskal-Wallis, chegamos às mesmas conclusões sobre a significância dessas diferenças ($H_{(5)} = 16{,}630$; $valor\text{-}p = 0{,}005 < 0{,}05$). Adicionalmente, as discrepâncias nas médias entre as escolas podem ser confirmadas pela figura subsequente.

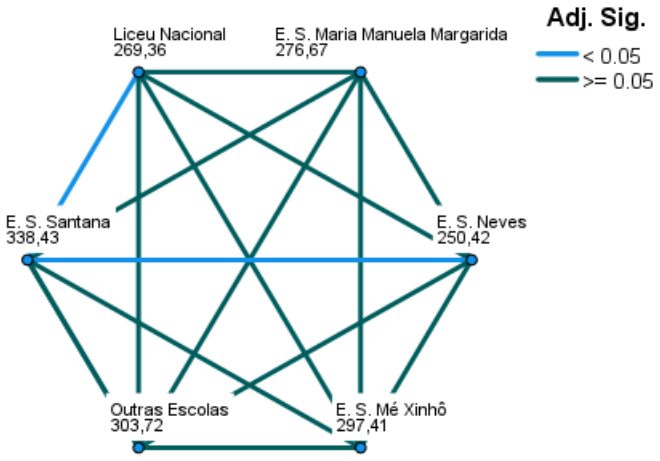


Figura 5 – Comparações por Método *Pairwise* de Escola Proveniente (Base de Dados II)

- ✓ **H4** (Hipótese 4): As médias finais do ensino secundário dos candidatos são iguais nos diferentes Cursos afectos aos departamentos matriculados?

No contexto desta hipótese, observou-se na Base de Dados I que a homogeneidade de variância foi rejeitada para $\alpha = 0{,}05$, conforme evidenciado na tabela subsequente.

Tabela 15 – Testes de Homogeneidade de Variância (Base de Dados I).

Teste de Homogeneidade de Variância

		Estatística de Levene	gl1	gl2	Sig.
Média Final do 12º Ano	Com base em média	23,762	3	547	<,001
	Com base em mediana	19,154	3	547	<,001
	Com base em mediana e com gl ajustado	19,154	3	490,476	<,001
	Com base em média aparada	23,111	3	547	<,001

Na Base de Dados II, observou-se um desequilíbrio significativo (base de dados desbalanceada), onde se constatou um grupo com apenas 2 candidatos, outro com 28 candidatos, em comparação com restantes grupos com mais de 200 candidatos, conforme demonstrado a seguir:

Tabela 16 – Composição da Base de dados por Departamentos (Base de Dados II).

		Válido		Total	
	Curso Inscrito	N	Porcentagem	N	Porcentagem
Média Final do 12º Ano	DCNVA	2	100,0%	2	100,0%
	DL&CHS	291	100,0%	291	100,0%
	DCE&EE	28	100,0%	28	100,0%
	DCE	245	100,0%	245	100,0%

Considerando ao reduzido número de estudantes matriculados nos cursos pertencentes ao departamento DCNVA, optou-se por excluí-los da análise a fim de preservar a robustez dos testes estatísticos.

Prosseguindo, a análise não paramétrica revelou diferenças estatisticamente significativas nas médias do ensino secundário entre candidatos matriculados em cursos associados a pelo menos dois departamentos distintos. A hipótese nula foi rejeitada para $\alpha = 0{,}05$ (H de Kruskal-Wallis $= 159{,}458$; $df = 3$; $valor\text{-}p < 0{,}001$), conforme indicado no Quadro 5.

Quadro 4 – Resumo do teste Kruskal-Wallis (Base de Dados I)

Sumarização de Teste de Hipótese

	Hipótese nula	Teste	Sig.[a,b]	Decisão
1	A distribuição de Média Final do 12º Ano é igual nas categorias de Curso Inscrito.	Amostras Independentes de Teste de Kruskal-Wallis	,000	Rejeitar a hipótese nula.

a. O nível de significância é ,050.

b. A significância assintótica é exibida.

Essa diferença é observada nas médias do ensino secundário dos estudantes matriculados nos cursos afetos a seguintes pares de departamento: *i)* Departamento de Ciências da Educação e o

Departamento de Línguas e Ciências Humanas e Sociais, *ii)* Departamento de Ciências da Educação e o Departamento de Ciências da Natureza, Vida e Ambiente, *iii)* Departamento de Ciências da Natureza, Vida e Ambiente e o Departamento de Línguas e Ciências Humanas e Sociais, e *iv)* Departamento de Ciências da Natureza, Vida e Ambiente e o Departamento de Ciências Económicas, Ciências Exatas e Engenharia.

Esses resultados são confirmados na figura seguinte, utilizando a significância ajustada pela correção de Bonferroni.

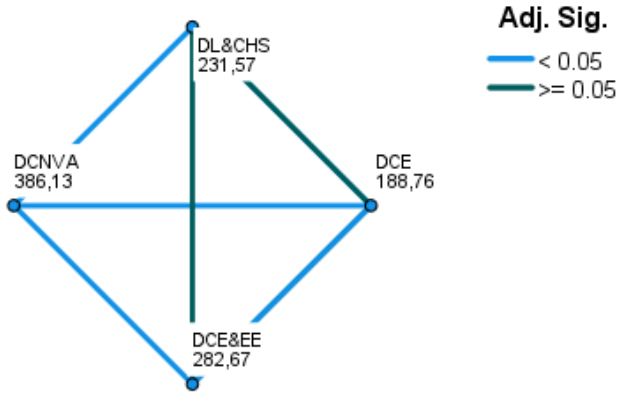


Figura 6 – Comparações por Método Pairwise de Curso Inscrito (Base de Dados I)

Ao continuar a análise na Base de Dados II, constatou-se a existência de diferenças significativas nas médias do ensino secundário dos candidatos em cursos de pelo menos dois departamentos, onde a hipótese nula foi rejeitada para $\alpha = 0{,}05$ (H de Kruskal-Wallis $= 9{,}479$; $df = 2$; $valor\text{-}p = 0{,}009 < \alpha$), conforme apresentado no Quadro 6.

Quadro 5 – Resumo do teste Kruskal-Wallis (Base de Dados II)

Sumarização de Teste de Hipótese

	Hipótese nula	Teste	Sig.[a,b]	Decisão
1	A distribuição de Média Final do 12º Ano é igual nas categorias de Curso Inscrito.	Amostras Independentes de Teste de Kruskal-Wallis	,009	Rejeitar a hipótese nula.

a. O nível de significância é ,050.

b. A significância assintótica é exibida.

Essa diferença é observada na média do ensino secundário dos estudantes matriculados nos cursos do Departamento de Ciências Económicas & Ciências Exatas e Engenharia em comparação com os matriculados nos cursos do Departamento de Línguas e Ciências Humanas e Sociais, conforme ilustrado no Apêndice C.

3.2 Aplicação de Regressão Linear

Após examinar as características descritivas das variáveis nas bases de dados e testar algumas hipóteses, procede-se à estimação de modelos de regressão com o intuito de identificar e analisar as contribuições dos candidatos matriculados em diversos cursos da USTP em relação ao sexo, idade, escolas de proveniência (instituições onde concluíram o ensino secundário), notas obtidas em disciplinas específicas e o local de residência, para a explicação da média final do ensino secundário.

Dado que essa média é crucial para a admissão nos cursos da Universidade de São Tomé e Príncipe, é essencial que os estudantes e a comunidade académica conheçam as variáveis significativas que afetam as notas dos alunos que concluem o ensino secundário e desejam ingressar na USTP.

Devido à natureza das variáveis binárias presentes na base de dados, foi necessário categorizá-las antes da modelação. Para isso, foram escolhidas as seguintes categorias de referência com base nas características das variáveis nominais: *i)* Sexo-Masculino, *ii)* Residência-Água Grande, *iii)* Escola Proveniente-Liceu Nacional.

Inicialmente, realizou-se uma análise minuciosa da base de dados com o intuito de identificar eventuais falhas no modelo ou pressupostos que não fossem válidos, visando assim identificar e eliminar suas possíveis causas.

Esta análise tinha como foco os *outliers*., uma vez que sua análise é fundamental para a construção de um modelo capaz de realizar inferências consistentes, uma vez que estes têm um forte impacto em outros pressupostos, como a homogeneidade e (multi)colinearidade. Além disso, essa análise contribuirá significativamente para simplificar o processo de estimação do modelo final (que atenda aos pressupostos da regressão linear).

3.2.1 Estimação da Média Final do Ensino secundário dos estudantes dos Cursos de Ciências (Matemática - Base de Dados I)

Da análise desta Base de Dados, onde destacam-se as notas das disciplinas de Matemática, Biologia e Física/Química. Nestas, foram detectados alguns *outliers* que foram considerados plausíveis de serem excluídos, o que resultou na redução do tamanho da amostra em cerca de 52 observações em comparação com o tamanho utilizado na seção anterior.

Em seguida foi conduzido o modelo de regressão múltipla considerando a variável dependente a Média Final do Ensino Secundário e as variáveis preditoras o Sexo, a Idade, a Residência, a

Escola de origem e as notas nas disciplinas de Língua Portuguesa, História, Física/Química, Biologia, Direito e Sociologia/Psicologia. Após esta análise, os outputs do IBM SPSS Statistics, no método "*Stepwise ou Regressão Stepwise*", revelaram um modelo com correlação muito forte entre os dados, onde 88,8% ($R = 0{,}943$; $R_a^2 = 0{,}888$) da variação da média do ensino secundário pode ser explicada pelas variáveis exógenas incluídas no modelo.

Adicionalmente, os resultados do teste de ajustamento do modelo indicam que pelo menos uma variável exógena presente no modelo é significativa, ou seja, é diferente de zero e capaz de explicar a variação observada ($F_{(7,\ 491)} = 567{,}829$; $valor\text{-}p < 0{,}01$).

Tabela 17 – Teste de significância do Modelo de Regressão.

ANOVA[a]

Modelo		Soma dos Quadrados	df	Quadrado Médio	F	Sig.
7	Regressão	903,452	7	129,065	567,829	<,001[h]
	Resíduo	111,602	491	,227		
	Total	1015,054	498			

a. Variável Dependente: Média Final do 12º Ano

h. Preditores: (Constante), Biologia, Física/Química, L. Portuguesa, Matemática, Idade, Resid=Lembá, Resid=Cantagalo

Na Tabela 18, observa-se que os coeficientes estatisticamente significativos no modelo incluem as notas obtidas em Biologia, Física e Química, Língua Portuguesa e Matemática, além das variáveis idade e residência nos distritos de Cantagalo ou Lembá. Verificou-se que estas variáveis, excetuando as três últimas, apresentam contribuições esperadas para a Média Final do Ensino Secundário, considerando que esta é calculada como a média aritmética dessas disciplinas.

Contudo, é crucial salientar que, com base nos coeficientes padronizados, as notas da disciplina de Língua Portuguesa exerceram a maior contribuição na determinação da Média Final. Em contrapartida, as notas da Matemática apresentaram a menor contribuição relativa, sugerindo resultados comparativamente inferiores nesta disciplina em relação às demais avaliadas.

Por outro lado, observou-se que a Média Final do Ensino Secundário diminui, em média, 0,211 pontos para cada ano adicional de idade, o que pode refletir impactos adversos de fatores associados ao envelhecimento dos estudantes durante o período de ensino secundário.

Assim, conforme indicado na Tabela 18, o modelo de regressão que explica a média final do ensino secundário é descrito pelas seguintes equações:

Para um estudante residente no distrito de Água Grande (ou noutros distritos além do distrito de Cantagalo ou Lembá):

$$\boldsymbol{MF12Ano} = 3{,}500 - 0{,}037 \cdot \text{Idade} + 0{,}114 \cdot \text{Matem} + 0{,}291 \cdot \text{L. Port} + 0{,}212 \cdot \text{Bio} + 0{,}206 \cdot \text{FisQuim}$$

Para um estudante residente no distrito de Cantagalo (ou noutros distritos além do distrito de Água Grande ou Lembá):

$$\boldsymbol{MF12Ano} = 3{,}673 - 0{,}037 \cdot \text{Idade} + 0{,}114 \cdot \text{Matem} + 0{,}291 \cdot \text{L. Port} + 0{,}212 \cdot \text{Bio} + 0{,}206 \cdot \text{FisQuim}$$

Para um estudante residente no distrito de Lembá (ou noutros distritos além do distrito de Água Grande ou Cantagalo):

$$\boldsymbol{MF12Ano} = 3{,}289 - 0{,}037 \cdot \text{Idade} + 0{,}114 \cdot \text{Matem} + 0{,}291 \cdot \text{L. Port} + 0{,}212 \cdot \text{Bio} + 0{,}206 \cdot \text{FisQuim}$$

Tabela 18 – Coeficientes e significâncias do modelo de regressão.

Coeficientes[a]

Modelo		Coeficientes não padronizados		Coeficientes padronizados			Estatísticas de colinearidade	
		B	Std. Error	Beta	t	Sig.	Tolerância	VIF
7	(Constante)	3,500	,312		11,220	<,001		
	Biologia	,212	,017	,294	12,299	<,001	,393	2,543
	Física/Química	,206	,017	,296	12,099	<,001	,374	2,675
	L. Portuguesa	,291	,022	,309	13,105	<,001	,402	2,490
	Matemática	,114	,022	,114	5,113	<,001	,450	2,224
	Idade	-,037	,007	-,084	-4,885	<,001	,748	1,336
	Resid=Lembá	-,211	,069	-,047	-3,043	,002	,934	1,071
	Resid=Cantagalo	,173	,080	,033	2,175	,030	,973	1,028

a. Variável Dependente: Média Final do 12º Ano

Observa-se que a variável que indica a escola onde o estudante conclui o ensino secundário não apresenta significância estatística para predizer o aumento ou diminuição de sua média final. Isso sugere uma uniformidade nos resultados, indicando possível consistência nas metodologias educacionais entre as escolas consideradas.

Com um nível de significância de 5%, verificou-se que a variável "Residência" (excluindo Água Grande, Cantagalo e Lembá) foi excluída do modelo por não contribuir de maneira estatisticamente significativa para sua explicação, conforme detalhado na tabela subsequente.

Tabela 19 – Variáveis excluídas do modelo de regressão.

Variáveis excluídas[a]

Modelo		Beta In	t	Sig.	Correlação parcial	Estatísticas de colinearidade Tolerância
7	Sexo=Feminino	-,019[h]	-1,280	,201	-,058	,976
	Resid=Mé-Zochi	-,008[h]	-,479	,632	-,022	,856
	Resid=Lobata	-,014[h]	-,899	,369	-,041	,977
	Resid=Caué&RAP	,028[h]	1,878	,061	,085	,992
	EscProven=E. S. Maria Manuela Margarida	-,003[h]	-,166	,868	-,007	,943
	EscProven=E. S. Santana	,004[h]	,214	,831	,010	,730
	EscProven=E. S. Neves	,004[h]	,235	,815	,011	,634
	EscProven=E. S. Mé Xinhô	-,020[h]	-1,232	,219	-,056	,881
	EscProven=Outras Escolas	,005[h]	,326	,745	,015	,844

a. Variável Dependente: Média Final do 12º Ano

h. Preditores no Modelo: (Constante), Biologia, Física/Química, L. Portuguesa, Matemática, Idade, Resid=Lembá, Resid=Cantagalo

- Validação dos prossupostos

Para efeitos de inferências, o modelo de regressão ajustado deve verificar os pressupostos. Apesar do modelo atual demonstrar uma taxa de ajustamento superior a 85%, é imperativo validar esses pressupostos.

Para tanto, iniciou-se com a identificação de *outliers*, cuja análise, exclusão ou correção é crucial para a construção de um modelo capaz de realizar inferências de forma eficaz e consistente, pois o mesmo tem forte impacto noutros pressupostos, como o da homogeneidade.

A análise da Figura 7 revelou a inexistência de *outliers* no modelo.

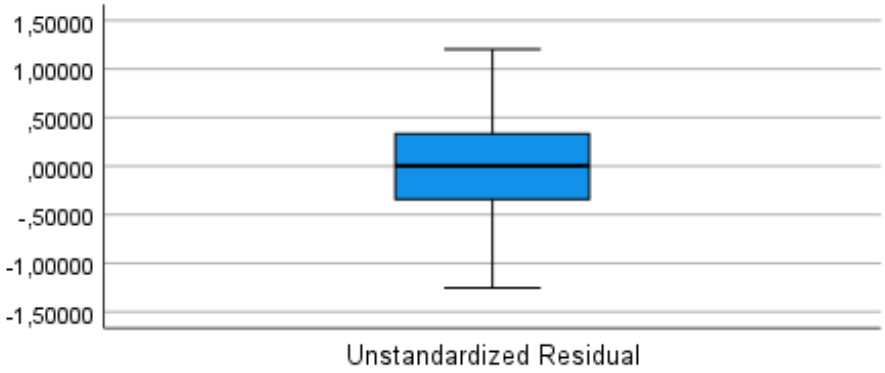


Figura 7 – Diagnóstico do *outliers* (Base de Dados I)

Esta percepção ou conclusão é validada pelos valores estimados da *Leverage*, conforme ilustrado na figura subsequente:

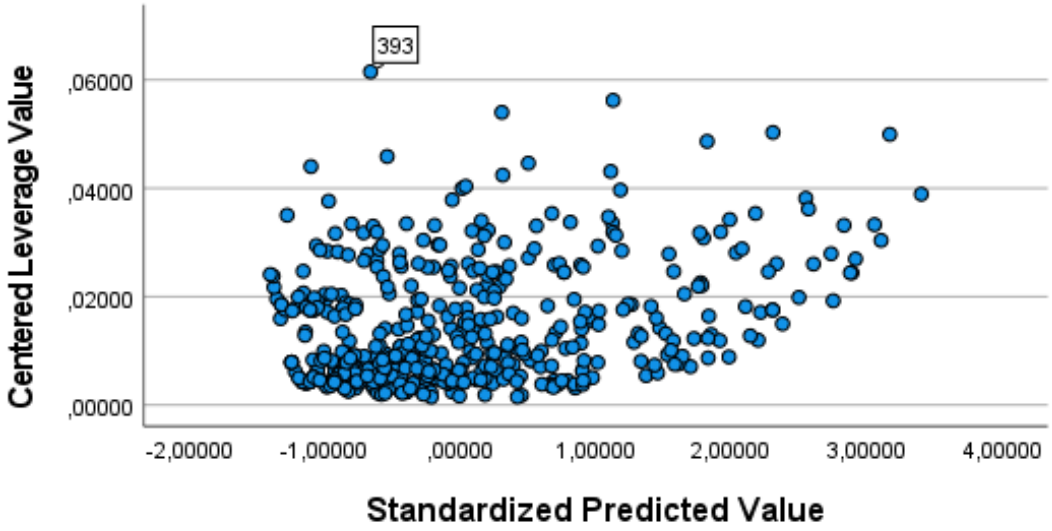


Figura 8 – Diagnóstico do *outliers* pela *Leverage* (Base de Dados I)

A observação 393 é a única com um alto valor de alavancagem centrada, embora inferior a 0,5, o que a exclui da classificação como *outlier*.

Em relação ao pressuposto da normalidade, podemos constatar graficamente (ver Apêndice F) que os resíduos exibem indícios de uma distribuição normal, pois estão muito próximos da linha reta.

Esta suposição é respaldada pelo teste de Kolmogorov-Smirnov, no qual não se rejeita a hipótese de normalidade dos resíduos para um nível de significância de 5% ($valor\text{-}p = 0{,}190 > 0{,}05$), conforme evidenciado na tabela a seguir:

Tabela 20 – Teste de Normalidade do resíduo.

Testes de Normalidade

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	gl	Sig.	Estatística	gl	Sig.
Unstandardized Residual	,028	499	,200*	,996	499	,190

*. Este é um limite inferior da significância verdadeira.

a. Correlação de Significância de Lilliefors

No que diz respeito à autocorrelação, os resultados da estatística de Durbin-Watson não oferecem evidências conclusivas sobre a presença ou ausência de autocorrelação nos resíduos. Para um tamanho de amostra $n = 499$ e $p = 8$ variáveis explicativas, os valores calculados de $d_U = 1{,}890 < \boldsymbol{DW = 1,917} < 4 - d_U = 4 - 1{,}890 = 2{,}11$ indicam que podemos afirmar com certeza a inexistência da autocorrelação.

Ao investigar os pressupostos de homogeneidade, considerando que os dados seguem uma distribuição normal, constatou-se que a hipótese de homogeneidade da variância é confirmada. Não há evidências significativas de heterocedasticidade, conforme indicado pelo $valor\text{-}p$ do teste de Breusch-Pagan ($valor\text{-}p > 0{,}05$), conforme demonstrado na Tabela 21. Adicionalmente, a estatística do teste de White também conduz à mesma conclusão ($\chi^2 = 123{,}611$; $valor\text{-}p = 0{,}072 > 0{,}05$).

Tabela 21 – Teste de Heteroscedasticidade.

Teste Breusch-Pagan para Heterocedasticidade[a,b]

Qui-quadrado	df	Sig.
,703	1	,402

a. Variável dependente: Média Final do 12º Ano

b. Testa a hipótese nula de que a variação dos erros não depende dos valores das variáveis independentes.

No diagnóstico de (multi)colinearidade, conforme evidenciado na Tabela 18, observa-se que nenhuma variável do modelo apresenta problemas relevantes de (multi)colinearidade, uma vez que todos os valores dos fatores de inflação da variância (VIF) são inferiores a 5 ($VIF < 5$)

Portanto, após a validação de todos os pressupostos dos modelos, nos quais as suas condições foram adequadamente atendidas (sem rejeição das hipóteses nulas), podemos afirmar que os modelos estimados são robustos e apropriados para os dados analisados.

3.2.2 Estimação da Média Final do Ensino secundário dos estudantes dos Cursos de Letras (História – Base de Dados II)

Durante essa análise desta Base de Dados, onde destacam-se, entre outras variáveis, as notas das disciplinas de História, Direito e Psicologia/Sociologia, foram detectados alguns *outliers* que resultou na redução do tamanho da amostra em cerca de 6 observações em comparação com o tamanho utilizado na seção anterior.

Utilizando as mesmas categorias de referência do modelo anterior, o modelo revelou fortes correlações entre os dados, utilizando o método de "*Stepwise ou Regressão Stepwise*". Verificou-se que aproximadamente 76,2% ($R = 0{,}875$; $R_a^2 = 0{,}762$) da variação na média do ensino secundário pode ser explicada pelas variáveis exógenas incluídas no modelo.

Além disso, os resultados do teste de significância do modelo indicam que pelo menos uma variável exógena presente no modelo é significativa, ou seja, é diferente de zero e capaz de explicar a variação observada, conforme demonstrado na Tabela 22.

Tabela 22 – Teste de significância do Modelo de Regressão.

ANOVA[a]

Modelo		Soma dos Quadrados	df	Quadrado Médio	F	Sig.
7	Regressão	357,661	7	51,094	254,645	<,001
	Resíduo	109,755	547	,201		
	Total	467,416	554			

a. Variável Dependente: Média Final do 12º Ano

A análise dos coeficientes do modelo revelou a significativa contribuição das notas de Direito, Psicologia/Sociologia, Língua Portuguesa e História para a média final do ensino secundário, evidenciando que um acréscimo de um ponto em qualquer destas disciplinas resulta num aumento correspondente na média final, conforme demonstrado na Tabela 23.

Entre as variáveis analisadas, os coeficientes padronizados indicam que as notas de Direito apresentaram a maior contribuição relativa na determinação da Média Final, ressaltando sua relevância no desempenho académico global. Em contraste, as notas de Língua Portuguesa exibiram a menor contribuição, apontando para um desempenho comparativamente inferior nesta disciplina em relação às demais avaliadas.

Além disso, verificou-se uma diferença estatisticamente significativa entre as médias dos estudantes residentes nos distritos de Cantagalo ou Lembá e dos estudantes do sexo masculino residentes em Água Grande e frequentadores do Liceu Nacional, sendo as médias dos estudantes de Lembá inferiores. Também foi observado que a média final do ensino secundário de um estudante aumenta em 0,009 pontos a cada ano adicional de idade.

Portanto, conforme indicado na Tabela 23, o modelo de regressão que explica a média final do ensino secundário pode ser representado pelas seguintes equações:

Para um estudante residente no distrito de Água Grande (ou noutros distritos além do distrito de Cantagalo ou Lembá):

$$\boldsymbol{MF12Ano} = 2{,}687 + 0{,}009 \cdot \text{Idade} + 0{,}189 \cdot \text{História} + 0{,}189 \cdot \text{L. Port} + 0{,}210 \cdot \text{Direito} + 0{,}181 \cdot \text{PscicolSociol}$$

Para um estudante residente no distrito de Cantagalo (ou noutros distritos além do distrito de Água Grande ou Lembá):

$$\boldsymbol{MF12Ano} = 2{,}871 + 0{,}009 \cdot \text{Idade} + 0{,}189 \cdot \text{História} + 0{,}189 \cdot \text{L. Port} + 0{,}210 \cdot \text{Direito} + 0{,}181 \cdot \text{PscicolSociol}$$

Para um estudante residente no distrito de Lembá (ou noutros distritos além do distrito de Água Grande ou Cantagalo):

$$\boldsymbol{MF12Ano} = 2{,}565 + 0{,}009 \cdot \text{Idade} + 0{,}189 \cdot \text{História} + 0{,}189 \cdot \text{L. Port} + 0{,}210 \cdot \text{Direito} + 0{,}181 \cdot \text{PscicolSociol}$$

Tabela 23 – Coeficientes e significâncias do modelo de regressão.

Coeficientes[a]

Modelo		Coeficientes não padronizados		Coeficientes padronizados			Estatísticas de colinearidade	
		B	Std. Error	Beta	t	Sig.	Tolerância	VIF
7	(Constante)	2,687	,260		10,343	<,001		
	Direito	,210	,015	,350	14,258	<,001	,713	1,403
	Psicologia e Sociologia	,181	,015	,311	12,360	<,001	,678	1,476
	História	,189	,018	,272	10,430	<,001	,630	1,588
	L. Portuguesa	,189	,022	,216	8,597	<,001	,679	1,473
	Resid=Cantagalo	,184	,059	,067	3,098	,002	,922	1,085
	Idade	,009	,004	,051	2,322	,021	,900	1,111
	Resid=Lembá	-,122	,060	-,043	-2,052	,041	,973	1,027

a. Variável Dependente: Média Final do 12º Ano

Observe-se que, assim como no modelo anterior, a variável que indica a escola onde o estudante concluiu o ensino secundário não apresenta significância estatística para predizer alterações na sua média final.

Com um nível de significância de 5%, as variáveis "Residência" (excluindo Água Grande, Cantagalo e Lembá) e "Sexo-Feminino" foram excluídos do modelo por não apresentarem uma contribuição estatisticamente significativa para a explicação da média final. As categorias de referência utilizadas foram: *i)* Sexo-Masculino, *ii)* Residência-Água Grande, *iii)* Escola Proveniente-Liceu Nacional, o conforme detalhado no Apêndice E.

- Validação dos prossupostos

Seguindo os procedimentos aplicados no modelo anterior, procedeu-se à análise de *outliers*, sendo que, conforme evidenciado na Figura 9, não foram identificados nenhum.

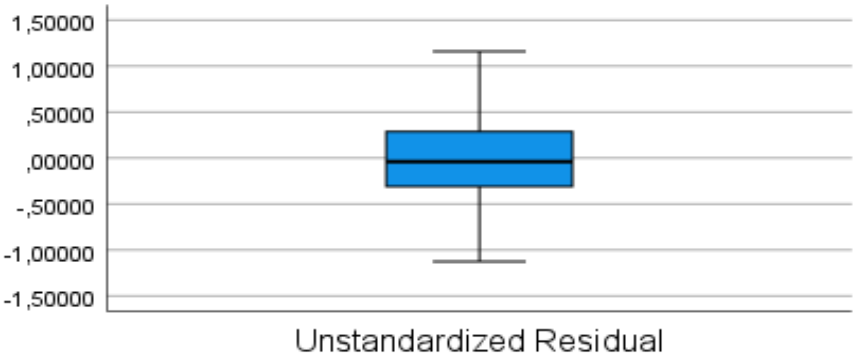


Figura 9 – Diagnóstico do *outliers* (Base de Dados II)

Esta percepção ou conclusão é corroborada pelos valores estimados da *Leverage*, conforme ilustrado na figura subsequente:

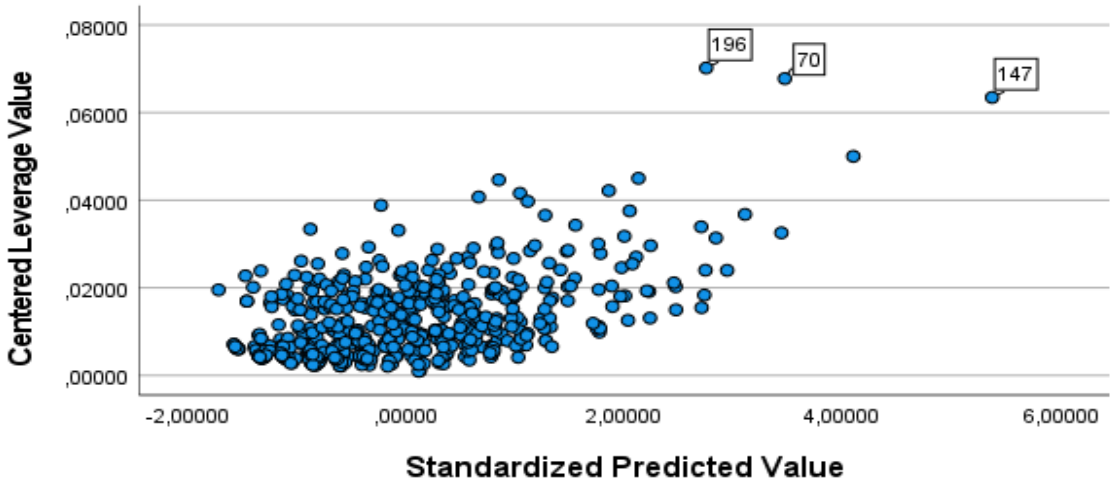


Figura 10 – Diagnóstico do *outliers* pela *Leverage* (Base de Dados II)

Embora as observações 70, 147 e 196 consistam em manter distâncias dos demais valores de alavancagem centralizada, estas não são consideradas *outliers* (valores com alavancagem menor que 0,5) e não apresentam evidências de serem pontos influentes, conforme confirmado pelo gráfico $DfFits$ vs $\hat{y}_j$.

Quanto à suposição de normalidade, verificamos através do teste de Kolmogorov-Smirnov que os resíduos seguem uma distribuição normal. A 5% de significância, a hipótese de nulidade não foi rejeitada, conforme evidenciado na tabela a seguir:

Tabela 24 – Teste de Normalidade do resíduo.

Testes de Normalidade

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	gl	Sig.	Estatística	gl	Sig.
Unstandardized Residual	,037	555	,074	,995	555	,074

a. Correlação de Significância de Lilliefors

No que diz respeito à autocorrelação, os resultados do teste de Durbin-Watson não evidenciaram a presença de autocorrelação nos resíduos. Com uma de amostra de $n = 555$ e $p = 7$ variáveis explicativas, os valores calculados de $d_U = 1{,}828 < \boldsymbol{DW = 1{,}951} < 4 - d_U = 4 - 1{,}828 = 2{,}172$, não fornece base suficiente para eliminar uma hipótese nula. Portanto conclui-se que os resíduos são não autocorrelacionados.

Ao investigar os pressupostos de homogeneidade, a semelhança do modelo anterior, constatou-se que a hipótese de homogeneidade da variância é confirmada. Não há evidências

significativas de heterocedasticidade, conforme indicado pelo $valor\text{-}p$ do teste de Breusch-Pagan ($\chi^2 = 3{,}396;\ valor\text{-}p = 0{,}065 > 0{,}05$). Igualmente, a estatística do teste de White conduz à mesma conclusão ($\chi^2 = 125{,}532;\ valor\text{-}p = 0{,}107 > 0{,}05$).

No diagnóstico de (multi)colinearidade, conforme evidenciado na Tabela 23, observa-se que nenhuma variável do modelo apresenta problemas relevantes de (multi)colinearidade, uma vez que todos os valores dos fatores de inflação da variância (VIF) são inferiores a 5 ($VIF < 5$)

Portanto, após a validação de todos os pressupostos dos modelos, nos quais as suas condições foram adequadamente atendidas (sem rejeição das hipóteses nulas), podemos afirmar que os modelos estimados são robustos e apropriados para os dados analisados.

3.3 Aplicação de Regressão Logística

Neste subcapítulo, procede-se à estimação de modelos de regressão que investigam as contribuições de diversas variáveis em diferentes níveis ou categorias com o objectivo de propor abordagens alternativas para classificar os estudantes nos diferentes cursos afectos aos diferentes departamentos da USTP em relação ao sexo, idade, escolas de proveniência (instituições onde concluíram o ensino secundário), notas obtidas em disciplinas específicas e o local de residência, com base nas observações disponíveis.

Esses modelos proporcionarão aos estudantes o conhecimento de potenciais cursos ou departamentos onde as chances de admissão são maiores, bem como a comunidade académica a afetação correcta dos mesmos ou auxílio na definição dos critérios de selecção.

3.3.1 Estimação dos Modelos para a afetação dos estudantes dos Cursos de Ciências (Matemática) nos departamentos da USTP

A estimação dos modelos desta Base de Dados tem como a variável dependente o "Curso Inscritos", a qual possui quatro categorias correspondentes aos diferentes departamentos da USTP designadamente DCNVA, DL&CHS, DCE&EE e DCE que é a categoria de referência. A Residência (categoria Caué&RAP) apresenta apenas duas observações correspondentes a dois candidatos. A inclusão dessas observações na análise resultaria em aumento dos intervalos de confianças e as razões de chances, o que justifica a decisão de excluí-las da análise.

Ao analisar as métricas de ajuste do modelo, constatou-se que o modelo ajustado é estatisticamente significativo ($G(48) = 332{,}229$ e $valor\text{-}p < 0{,}001$, ver Tabela 25). Isso indica que pelo menos uma variável independente tem um impacto significativo na admissão dos estudantes nos cursos vinculados aos respectivos departamentos.

Os Critérios de Informação de Akaike (AIC) e Bayesiano (BIC) indicam que o modelo final apresenta um ajuste melhor do que o modelo com apenas o intercepto (ordenada na origem). Além disso, a verossimilhança -2LL evidencia a superioridade e adequação do modelo final.

Tabela 25 – Critérios de ajuste do modelo (Base de Dados I)

Informações de ajuste do modelo

	Critérios de ajuste do modelo			Testes de razão de verossimilhança		
Modelo	AIC	BIC	Verossimilhança de log -2	Qui-quadrado	df	Sig.
Somente intercepto	1306,471	1319,097	1300,471			
Final	1070,242	1284,880	968,242	332,229	48	<,001

Em termos da qualidade de ajuste dos dados, a Tabela 26 exibe as estatísticas de teste e os resultados de significância do teste Qui-Quadrado de Pearson e da *Deviance*.

Para a hipótese de que o modelo ajusta-se adequadamente aos dados, não se rejeita a hipótese nula (H_0), dado que a *deviance* apresenta os seguintes resultados: $D(1401) = 955{,}530$ e $valor\text{-}p = 1{,}000 > 0{,}05$.

Tabela 26 – Adequação do Ajuste do Modelo (Base de Dados I)

Adequação do ajuste

	Qui-quadrado	df	Sig.
Pearson	4511,677	1401	,000
Desvio	955,530	1401	1,000

A qualidade do modelo é avaliada pelos pseudo-R^2. No entanto, o teste $R^2_{MF} = 1 - \frac{-2LL_c}{G^2+(-2LL_c)} = 0{,}253$ revela um ajuste medíocre do modelo aos dados, enquanto o pseudo-R^2 de Nagelkerke $(0{,}525)$ indica uma qualidade razoável.

Assim como nos modelos de regressão linear, é fundamental verificar a significância dos parâmetros (ver Tabela 27) do modelo com rigor estatístico, a fim de construí-lo de maneira robusta e capaz de explicar adequadamente a variável dependente.

Tabela 27 – Testes de Razão de Verosimilhança (Base de Dados I).

Testes de razão de verossimilhança

	Critérios de ajuste do modelo			Testes de razão de verossimilhança		
Efeito	AIC do modelo reduzido	BIC do modelo reduzido	Verossimilhança de log -2 do modelo reduzido	Qui-quadrado	df	Sig.
Intercepto	1070,242	1284,880	968,242[a]	,000	0	.
Idade	1103,993	1306,005	1007,993	39,751	3	<,001
Média Final do 12º Ano	1065,547	1267,559	969,547	1,305	3	,728
Matemática	1073,776	1275,788	977,776	9,534	3	,023
L. Portuguesa	1066,355	1268,367	970,355	2,113	3	,549
Biologia	1097,287	1299,300	1001,287	33,046	3	<,001
Física/Química	1065,795	1267,807	969,795	1,553	3	,670
Género	1089,514	1291,526	993,514	25,272	3	<,001
Residência	1062,299	1226,434	984,299	16,057	12	,189
Escola Proveniente	1077,926	1229,435	1005,926	37,684	15	,001

A estatística qui-quadrado é a diferença no log de verossimilhanças -2 entre o modelo final e um modelo reduzido. O modelo reduzido é formado pela omissão de um efeito do modelo final. A hipótese nula significa que todos os parâmetros desse feito são 0.

a. Esse modelo reduzido é equivalente ao modelo final porque a omissão do efeito não aumenta os graus de liberdade.

A tabela acima evidencia que a Idade, as notas das disciplinas de Matemática e Biologia, o Género (com Masculino como classe de referência), e a escola de conclusão do ensino

secundário (com Liceu Nacional como classe de referência) possuem efeitos estatisticamente significativos sobre o *logit* da probabilidade de inscrição em um dos cursos associados ao departamento de Ciências de Educação. Em resumo, essas variáveis influenciam significativamente a escolha de um curso na área de ciências na USTP.

Em contraste, as notas das disciplinas de Língua Portuguesa, a Média Final do Ensino Secundário, a disciplina de Física/Química, e o local de residência não demonstram significância no teste de razão de verossimilhança com um nível de significância $\alpha = 0{,}05$. No entanto, o modelo que inclui essas variáveis apresenta o melhor ajustamento de qualidade de acordo com o teste -2LL.

Prosseguindo com a estimativa do modelo de Regressão Logística Multinomial, com base nos resultados apresentados na Tabela 28, podemos encontrar as estimativas dos coeficientes do modelo para as variáveis exógenas em relação à classe de referência (Departamento de Ciências de Educação).

- Probabilidade de ingressar num dos cursos do Departamento de Ciências de Educação:

$$P[Y = \text{"DCE"}] = \frac{1}{1 + e^{-6{,}93-0{,}28\cdot\text{Idad}+0{,}206\cdot\text{MF12Ano}+0{,}11\cdot\text{Mat}+\cdots+0{,}66\cdot\text{M}+0{,}25\cdot\text{AG}+\cdots-0{,}37\cdot\text{ESNeves}-1{,}08\cdot\text{ESMXinho}} + e^{0{,}738-0{,}08\cdot\text{Idad}+0{,}227\cdot\text{MF12Ano}-0{,}443\cdot\text{Mat}+\cdots+0{,}86\cdot\text{M}+0{,}152\cdot\text{AG}+\cdots-0{,}898\cdot\text{ESNeves}-0{,}90\cdot\text{ESMXinho}} + e^{4{,}447-0{,}34\cdot\text{Idad}-0{,}069\cdot\text{MF12Ano}+0{,}237\cdot\text{Mat}+\cdots+1{,}57\cdot\text{M}+2{,}156\cdot\text{AG}+\cdots+0{,}052\cdot\text{ESNeves}-2{,}279\cdot\text{ESMXinho}}}$$

- Probabilidade de ingressar num dos cursos do Departamento de Ciências da Natureza, da Vida e do Ambiente (DCNVA):

$$P[Y = \text{"DCNVA"}] = \frac{e^{-6{,}93-0{,}28\cdot\text{Idad}+0{,}206\cdot\text{MF12Ano}+0{,}11\cdot\text{Mat}+\cdots+0{,}66\cdot\text{M}+0{,}25\cdot\text{AG}+\cdots-0{,}37\cdot\text{ESNeves}-1{,}08\cdot\text{ESMXinho}}}{1 + e^{-6{,}93-0{,}28\cdot\text{Idad}+0{,}206\cdot\text{MF12Ano}+0{,}11\cdot\text{Mat}+\cdots+0{,}66\cdot\text{M}+0{,}25\cdot\text{AG}+\cdots-0{,}37\cdot\text{ESNeves}-1{,}08\cdot\text{ESMXinho}} + e^{0{,}738-0{,}08\cdot\text{Idad}+0{,}227\cdot\text{MF12Ano}-0{,}443\cdot\text{Mat}+\cdots+0{,}86\cdot\text{M}+0{,}152\cdot\text{AG}+\cdots-0{,}898\cdot\text{ESNeves}-0{,}90\cdot\text{ESMXinho}} + e^{4{,}447-0{,}34\cdot\text{Idad}-0{,}069\cdot\text{MF12Ano}+0{,}237\cdot\text{Mat}+\cdots+1{,}57\cdot\text{M}+2{,}156\cdot\text{AG}+\cdots+0{,}052\cdot\text{ESNeves}-2{,}279\cdot\text{ESMXinho}}}$$

- Probabilidade de ingressar num dos cursos do Departamento de Língua e de Ciências Humanas e Sociais (DL&CHS):

$$P[Y = \text{"DL\&DCHS"}] = \frac{e^{0{,}738-0{,}08\cdot\text{Idad}+0{,}227\cdot\text{MF12Ano}-0{,}443\cdot\text{Mat}+\cdots+0{,}86\cdot\text{M}+0{,}152\cdot\text{AG}+\cdots-0{,}898\cdot\text{ESNeves}-0{,}90\cdot\text{ESMXinho}}}{1 + e^{-6{,}93-0{,}28\cdot\text{Idad}+0{,}206\cdot\text{MF12Ano}+0{,}11\cdot\text{Mat}+\cdots+0{,}66\cdot\text{M}+0{,}25\cdot\text{AG}+\cdots-0{,}37\cdot\text{ESNeves}-1{,}08\cdot\text{ESMXinho}} + e^{0{,}738-0{,}08\cdot\text{Idad}+0{,}227\cdot\text{MF12Ano}-0{,}443\cdot\text{Mat}+\cdots+0{,}86\cdot\text{M}+0{,}152\cdot\text{AG}+\cdots-0{,}898\cdot\text{ESNeves}-0{,}90\cdot\text{ESMXinho}} + e^{4{,}447-0{,}34\cdot\text{Idad}-0{,}069\cdot\text{MF12Ano}+0{,}237\cdot\text{Mat}+\cdots+1{,}57\cdot\text{M}+2{,}156\cdot\text{AG}+\cdots+0{,}052\cdot\text{ESNeves}-2{,}279\cdot\text{ESMXinho}}}$$

- Probabilidade de ingressar num dos cursos do Departamento de Ciências Económica e de Ciências Exactas e Engenharias (DCE&EE):

$$P[Y = \text{"DCE\&EE"}]$$
$$= \frac{e^{4{,}447-0{,}34\cdot\text{Idad}-0{,}069\cdot\text{MF12Ano}+0{,}237\cdot\text{Mat}+\cdots+1{,}57\cdot\text{M}+2{,}156\cdot\text{AG}+\cdots+0{,}052\cdot\text{ESNeves}-2{,}279\cdot\text{ESMXinho}}}{1 + e^{-6{,}93-0{,}28\cdot\text{Idad}+0{,}206\cdot\text{MF12Ano}+0{,}11\cdot\text{Mat}+\cdots+0{,}66\cdot\text{M}+0{,}25\cdot\text{AG}+\cdots-0{,}37\cdot\text{ESNeves}-1{,}08\cdot\text{ESMXinho}} + e^{0{,}738-0{,}08\cdot\text{Idad}+0{,}227\cdot\text{MF12Ano}-0{,}443\cdot\text{Mat}+\cdots+0{,}86\cdot\text{M}+0{,}152\cdot\text{AG}+\cdots-0{,}898\cdot\text{ESNeves}-0{,}90\cdot\text{ESMXinho}} + e^{4{,}447-0{,}34\cdot\text{Idad}-0{,}069\cdot\text{MF12Ano}+0{,}237\cdot\text{Mat}+\cdots+1{,}57\cdot\text{M}+2{,}156\cdot\text{AG}+\cdots+0{,}052\cdot\text{ESNeves}-2{,}279\cdot\text{ESMXinho}}}$$

Em relação à significância dos coeficientes, avaliada pelo teste de Wald, constatou-se que, para o Departamento de Ciências da Natureza, da Vida e do Ambiente (DCNVA), nenhuma categoria de residência dos estudantes, assim como algumas escolas onde estudaram, foi estatisticamente significativa em comparação com o Departamento de Ciências de Educação.

Para o Departamento de Língua e de Ciências Humanas e Sociais (DL&CHS), nenhuma das escolas onde os estudantes terminaram o ensino secundário apresentou significância estatística.

Em relação às variáveis métricas, verificou-se que a média do ensino secundário, bem como as notas em Língua Portuguesa, Matemática (exceto no DL&CHS, que apresentam uma significância marginal), Física/Química e Biologia (exceto no DCNVA), não demonstraram diferenças estatisticamente significativas entre os diversos departamentos em comparação com o departamento de referência. Em outras palavras, essas variáveis não permitem distinguir os outros departamentos em relação ao departamento de referência.

Na tabela apresentada, observa-se uma lista das categorias de cursos, onde se destaca um número limitado de variáveis significativas em comparação com o curso de referência, sugerindo que várias variáveis independentes não conseguem discernir adequadamente as probabilidades entre as classes de cursos em relação ao departamento de referência.

A probabilidade de transição da classe de referência para a classe DCNVA é significativamente influenciada pela nota na disciplina de Biologia ($\chi^2_W = 20{,}322;\ valor\text{-}p < 0{,}001$). A cada incremento na nota desta disciplina, as chances de inscrição em um curso do DCNVA aumentam em 85% $(100\% \times (1{,}85 - 1))$. Para a variável idade, as chances de inscrição em um dos cursos afetos ao DCNVA, DL&CHS ou DCE&EE diminuem respectivamente em $24{,}5\%$, $8{,}1\%$, $28{,}9\%$ por ano adicional de idade.

Além disso, constata-se que não há diferenças estatisticamente significativas nas médias do ensino secundário, nem nas notas de Língua Portuguesa, Matemática e Física/Química, no que se refere às chances de inscrição em um dos cursos do DCE em comparação com DCNVA, DL&CHS ou DCE&EE.

Tabela 28 – Estimação dos parâmetros (Base de Dados I)

Estimativas de Parâmetro

Curso Inscrito[a]		B	Erro Erro	Wald	df	Sig.	Exp(B)	95% Intervalo de Confiança para Exp(B)	
								Limite inferior	Limite superior
DCNVA	Intercepto	-6,928	2,883	5,776	1	,016			
	Idade	-,281	,068	16,926	1	<,001	,755	,661	,863
	Média Final do 12º Ano	,206	,305	,456	1	,500	1,229	,675	2,236
	Matemática	,105	,170	,381	1	,537	1,111	,796	1,551
	L. Portuguesa	,110	,178	,378	1	,539	1,116	,787	1,582
	Biologia	,615	,137	20,322	1	<,001	1,850	1,416	2,418
	Física/Química	,027	,126	,047	1	,829	1,028	,802	1,316
	[Género=1]	,658	,355	3,446	1	,063	1,931	,964	3,869
	[Género=2]	0[b]	.	.	0	.	.	.	.
	[Residência=1]	,245	,717	,116	1	,733	1,277	,313	5,212
	[Residência=2]	,835	,716	1,360	1	,244	2,304	,567	9,370
	[Residência=3]	,016	,833	,000	1	,985	1,016	,198	5,198
	[Residência=4]	,555	,816	,462	1	,497	1,742	,352	8,630
	[Residência=5]	0[b]	.	.	0	.	.	.	.
	[Escola Proveniente=1]	-,691	,470	2,156	1	,142	,501	,199	1,260
	[Escola Proveniente=2]	-2,539	,664	14,605	1	<,001	,079	,021	,290
	[Escola Proveniente=3]	1,421	1,268	1,255	1	,263	4,140	,345	49,723
	[Escola Proveniente=4]	-,365	,862	,180	1	,672	,694	,128	3,757
	[Escola Proveniente=5]	-1,078	,734	2,154	1	,142	,340	,081	1,436
	[Escola Proveniente=6]	0[b]	.	.	0	.	.	.	.
DL&CHS	Intercepto	,738	3,250	,052	1	,820			
	Idade	-,084	,053	2,534	1	,111	,919	,829	1,020
	Média Final do 12º Ano	,227	,318	,511	1	,475	1,255	,673	2,342
	Matemática	-,443	,242	3,359	1	,067	,642	,400	1,031
	L. Portuguesa	,282	,199	2,011	1	,156	1,325	,898	1,955
	Biologia	-,106	,152	,485	1	,486	,899	,667	1,212
	Física/Química	-,013	,145	,008	1	,927	,987	,742	1,312
	[Género=1]	,858	,328	6,818	1	,009	2,358	1,238	4,488
	[Género=2]	0[b]	.	.	0	.	.	.	.
	[Residência=1]	,152	,738	,042	1	,837	1,164	,274	4,945
	[Residência=2]	,691	,728	,901	1	,342	1,996	,479	8,311
	[Residência=3]	-,666	1,075	,384	1	,535	,514	,062	4,222
	[Residência=4]	,948	,761	1,552	1	,213	2,579	,581	11,451
	[Residência=5]	0[b]	.	.	0	.	.	.	.
	[Escola Proveniente=1]	-,093	,495	,035	1	,852	,912	,345	2,407
	[Escola Proveniente=2]	-,368	,603	,373	1	,542	,692	,212	2,255
	[Escola Proveniente=3]	1,206	1,592	,574	1	,449	3,341	,148	75,658
	[Escola Proveniente=4]	-,898	1,074	,699	1	,403	,407	,050	3,344
	[Escola Proveniente=5]	-,903	,723	1,562	1	,211	,405	,098	1,670
	[Escola Proveniente=6]	0[b]	.	.	0	.	.	.	.
DCE&EE	Intercepto	,447	2,981	,022	1	,881			
	Idade	-,341	,070	23,657	1	<,001	,711	,620	,816
	Média Final do 12º Ano	-,069	,313	,049	1	,825	,933	,505	1,723
	Matemática	,237	,176	1,804	1	,179	1,267	,897	1,790
	L. Portuguesa	,155	,186	,692	1	,406	1,167	,811	1,680
	Biologia	-,008	,147	,003	1	,957	,992	,743	1,324
	Física/Química	,142	,129	1,208	1	,272	1,153	,895	1,485
	[Género=1]	1,572	,330	22,696	1	<,001	4,815	2,522	9,193
	[Género=2]	0[b]	.	.	0	.	.	.	.
	[Residência=1]	2,156	,907	5,647	1	,017	8,633	1,459	51,080
	[Residência=2]	2,335	,890	6,881	1	,009	10,331	1,805	59,146
	[Residência=3]	,890	1,033	,743	1	,389	2,436	,322	18,443
	[Residência=4]	1,531	1,078	2,017	1	,156	4,621	,559	38,212
	[Residência=5]	0[b]	.	.	0	.	.	.	.
	[Escola Proveniente=1]	-1,534	,477	10,341	1	,001	,216	,085	,549
	[Escola Proveniente=2]	-2,434	,612	15,810	1	<,001	,088	,026	,291
	[Escola Proveniente=3]	1,001	1,302	,591	1	,442	2,721	,212	34,945
	[Escola Proveniente=4]	,052	,984	,003	1	,958	1,054	,153	7,243
	[Escola Proveniente=5]	-2,279	,960	5,632	1	,018	,102	,016	,672
	[Escola Proveniente=6]	0[b]	.	.	0	.	.	.	.

a. A categoria de referência é: DCE.

b. Este parâmetro é definido para zero porque é redundante.

Os modelos de regressão multinomial previamente estimados permitem calcular a probabilidade de cada sujeito pertencente a cada uma das classes. Portanto, um estudante é direcionado a se matricular em um dos cursos onde a probabilidade de admissão é mais elevada.

Na Tabela 29, verifica-se que aproximadamente 38 estudantes estão classificados de forma inadequada no DCNVA, ou seja, deveriam ter se matriculado em cursos de outros departamentos. No entanto, observou-se que cerca de 118 estudantes (75,6%) estão matriculados no departamento onde as probabilidades de admissão são mais altas.

Para os departamentos DL&CHS e DCE&EE, constata-se um baixo índice de classificação correta, com menos de 50% dos estudantes que tinham maiores probabilidades de serem admitidos nesses departamentos efetuando a matrícula.

Globalmente, a percentagem de estudantes que efetuaram matrícula nos cursos dos departamentos onde as probabilidades eram mais altas é de 60,8%. Isso indica que um número considerável de estudantes optou por cursos cujas probabilidades de admissão eram menores.

Tabela 29 – Classificação dos candidatos por departamentos (Base de Dados I)

Classificação

		Previsto			
Observado	DCNVA	DL&CHS	DCE&EE	DCE	Porcentagem Correta
DCNVA	118	3	11	24	75,6%
DL&CHS	11	8	8	44	11,3%
DCE&EE	25	3	36	28	39,1%
DCE	17	6	15	140	78,7%
Porcentagem global	34,4%	4,0%	14,1%	47,5%	60,8%

A percentagem de estudantes matriculados nos cursos dos departamentos onde têm maiores probabilidades é de 28,14% $\left(((0{,}314)^2 + (0{,}143)^2 + (0{,}185)^2 + (0{,}358)^2) \times 100\%\right)$. Este resultado revela que o modelo alcança uma taxa de classificação correta significativamente superior (2,16 × maior) àquela que seria obtida por acaso.

3.3.2 Estimação dos Modelos para a afetação dos estudantes dos Cursos de Letras (História) nos departamentos da USTP)

Nesta Base de Dados, a variável dependente "Curso Inscritos" possui três categorias designadamente DL&CHS, DCE&EE e DCE, sendo este último o grupo de referência. Conforme detalhado no subcapítulo 3.1.3, o departamento DCNVA apresenta apenas duas observações correspondentes a dois candidatos. A inclusão dessas observações na análise

resultaria em aumento de níveis de variáveis dependentes por subpopulação com frequências zero, o que justifica a decisão de excluí-las da análise.

Ao analisar o ajuste do modelo, observamos que o mesmo é estatisticamente significante pelo teste de Razão de Verossimilhança ($G(34) = 167{,}663$; $valor\text{-}p < 0{,}001$), conforme apresentado na Tabela 30. Este resultado indica evidências claras sobre a adequação do modelo aos dados.

Embora o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC) tenham sugerido abordagens diferentes, a verossimilhança -$2LL$ do modelo final demonstrou ser superior ao modelo com apenas o intercepto.

Tabela 30 – Critérios de ajuste do modelo (Base de Dados II)

Informações de ajuste do modelo

	Critérios de ajuste do modelo			Testes de razão de verossimilhança		
Modelo	AIC	BIC	Verossimilhança de log -2	Qui-quadrado	df	Sig.
Somente intercepto	944,298	952,929	940,298			
Final	844,635	999,988	772,635	167,663	34	<,001

Apesar da contrariedade entre o Critérios de Informação de Akaike (AIC) e Bayesiano (BIC) dos modelos em termos de ajuste aos dados, os resultados da Tabela 31 indicam que o modelo se ajusta adequadamente aos dados, não havendo motivo para rejeitar a hipótese nula (H_0). Isso é respaldado pelos valores da *deviance*: $D(1056) = 768{,}476$ e $valor\text{-}p = 1{,}000 > 0{,}05$ e do Qui-quadrado de *Pearson*: $\chi^2(1056) = 1012{,}754$ e $valor\text{-}p = 0{,}826 > 0{,}05$.

Tabela 31 – Adequação do Ajuste do Modelo (Base de Dados II)

Adequação do ajuste

	Qui-quadrado	df	Sig.
Pearson	1012,754	1056	,826
Desvio	768,476	1056	1,000

Apesar do modelo ajustar aos dados, a avaliação de sua qualidade por meio do pseudo-R^2 $R^2_{MF} = 1 - \frac{-2LL_c}{G^2+(-2LL_c)} = 0{,}178$ revela um ajuste medíocre do modelo aos dados. Esse resultado é respaldado pelos pseudo-R^2 de Nagelkerke (0,319) e de pseudo-R^2 de Cox e Snell (0,262), indicando uma explicação limitada da variabilidade dos dados pelo modelo proposto.

Ao analisar a significância dos coeficientes, menos de metade das variáveis presente no modelo com efeitos estatisticamente significante o *logit* da probabilidade de inscrição em um dos

cursos associados ao departamento de Ciências de Educação. Entre essas variáveis estão a Idade, Género e as notas em Língua Portuguesa e Direito (ver Apêndice D). Apesar da baixa presença de variáveis significativas, o modelo demonstrou o melhor ajuste de acordo com o teste $-2LL$.

No que concerne as estimativas dos coeficientes do modelo de Regressão Logística Multinomial em relação à classe de referência (Departamento de Ciências de Educação), tem-se a partir dos resultados apresentados na Tabela 32:

- Probabilidade de ingressar num dos cursos do Departamento de Ciências de Educação:

$$P[Y = \text{"DCE"}] = \frac{1}{1 + e^{-0{,}43-0{,}12\cdot\text{Idad}-0{,}420\cdot\text{MF12Ano}-0{,}10\cdot\text{Hist}+\cdots+1{,}16\cdot\text{M}+0{,}58\cdot\text{AG}+\cdots-1{,}55\cdot\text{ESNeves}-1{,}29\cdot\text{ESMXinho}} + e^{5{,}66-0{,}15\cdot\text{Idad}-0{,}797\cdot\text{MF12Ano}+0{,}04\cdot\text{Hist}+\cdots+1{,}65\cdot\text{M}+1{,}54\cdot\text{AG}+\cdots+1{,}10\cdot\text{ESNeves}-0{,}76\cdot\text{ESMXinho}}}$$

- Probabilidade de ingressar num dos cursos do Departamento de Língua e de Ciências Humanas e Sociais (DL&CHS):

$$P[Y = \text{"DL\&CHS"}] = \frac{e^{-0{,}43-0{,}12\cdot\text{Idad}-0{,}420\cdot\text{MF12Ano}-0{,}10\cdot\text{Hist}+\cdots+1{,}16\cdot\text{M}+0{,}58\cdot\text{AG}+\cdots-1{,}55\cdot\text{ESNeves}-1{,}29\cdot\text{ESMXinho}}}{1 + e^{-0{,}43-0{,}12\cdot\text{Idad}-0{,}420\cdot\text{MF12Ano}-0{,}10\cdot\text{Hist}+\cdots+1{,}16\cdot\text{M}+0{,}58\cdot\text{AG}+\cdots-1{,}55\cdot\text{ESNeves}-1{,}29\cdot\text{ESMXinho}} + e^{5{,}66-0{,}15\cdot\text{Idad}-0{,}797\cdot\text{MF12Ano}+0{,}04\cdot\text{Hist}+\cdots+1{,}65\cdot\text{M}+1{,}54\cdot\text{AG}+\cdots+1{,}10\cdot\text{ESNeves}-0{,}76\cdot\text{ESMXinho}}}$$

- Probabilidade de ingressar num dos cursos do Departamento de Ciências Económica e de Ciências Exactas e Engenharias (DCE&EE):

$$P[Y = \text{"DCE\&EE"}] = \frac{e^{5{,}66-0{,}15\cdot\text{Idad}-0{,}797\cdot\text{MF12Ano}+0{,}04\cdot\text{Hist}+\cdots+1{,}65\cdot\text{M}+1{,}54\cdot\text{AG}+\cdots+1{,}10\cdot\text{ESNeves}-0{,}76\cdot\text{ESMXinho}}}{1 + e^{-0{,}43-0{,}12\cdot\text{Idad}-0{,}420\cdot\text{MF12Ano}-0{,}10\cdot\text{Hist}+\cdots+1{,}16\cdot\text{M}+0{,}58\cdot\text{AG}+\cdots-1{,}55\cdot\text{ESNeves}-1{,}29\cdot\text{ESMXinho}} + e^{5{,}66-0{,}15\cdot\text{Idad}-0{,}797\cdot\text{MF12Ano}+0{,}04\cdot\text{Hist}+\cdots+1{,}65\cdot\text{M}+1{,}54\cdot\text{AG}+\cdots+1{,}10\cdot\text{ESNeves}-0{,}76\cdot\text{ESMXinho}}}$$

Os dados indicam que nenhuma variável métrica demostrou diferença estatisticamente significante entre cursos afectos ao departamento DCE&EE em comparação com os cursos do departamento DCE. Em outras palavras, essas variáveis não permitem distinguir o departamento DCE&EE em relação ao departamento de referência.

Além de não serem estatisticamente significantes, a cada incremento nessas variáveis resulta em uma redução nas chances de inscrição nos cursos afectos ao departamento DCE&EE em comparação com os cursos do departamento DCE.

Tabela 32 – Estimação dos parâmetros (Base de Dados II)

Estimativas de Parâmetro

Curso Inscrito[a]		B	Erro Erro	Wald	df	Sig.	Exp(B)	95% Intervalo de Confiança para Exp(B)	
								Limite inferior	Limite superior
DL&CHS	Intercepto	-,428	1,737	,061	1	,805			
	Idade	-,118	,023	25,932	1	<,001	,889	,849	,930
	Média Final do 12º Ano	-,420	,227	3,438	1	,064	,657	,421	1,024
	L. Portuguesa	,568	,134	17,963	1	<,001	1,764	1,357	2,294
	História	-,103	,108	,916	1	,339	,902	,730	1,114
	Direito	,301	,099	9,169	1	,002	1,351	1,112	1,641
	Psicologia e Sociologia	-,024	,087	,079	1	,779	,976	,823	1,157
	[Género=1]	1,163	,271	18,466	1	<,001	3,199	1,882	5,436
	[Género=2]	0[b]	.	.	0	.	.	.	.
	[Residência=1]	,582	,803	,526	1	,468	1,790	,371	8,636
	[Residência=2]	1,052	,812	1,676	1	,195	2,863	,582	14,069
	[Residência=3]	,938	,935	1,007	1	,316	2,556	,409	15,971
	[Residência=4]	,933	,925	1,016	1	,313	2,541	,415	15,574
	[Residência=5]	-1,105	1,162	,904	1	,342	,331	,034	3,231
	[Residência=6]	0[b]	.	.	0	.	.	.	.
	[Escola Proveniente=1]	-,629	,428	2,158	1	,142	,533	,230	1,234
	[Escola Proveniente=2]	-,826	,476	3,015	1	,082	,438	,172	1,112
	[Escola Proveniente=3]	-1,057	,615	2,950	1	,086	,348	,104	1,161
	[Escola Proveniente=4]	-1,545	,986	2,454	1	,117	,213	,031	1,474
	[Escola Proveniente=5]	-1,287	,716	3,231	1	,072	,276	,068	1,123
	[Escola Proveniente=6]	0[b]	.	.	0	.	.	.	.
DCE&EE	Intercepto	5,675	4,039	1,974	1	,160			
	Idade	-,151	,056	7,181	1	,007	,860	,771	,960
	Média Final do 12º Ano	-,797	,491	2,635	1	,105	,451	,172	1,180
	L. Portuguesa	,341	,312	1,195	1	,274	1,407	,763	2,594
	História	,038	,241	,024	1	,876	1,038	,648	1,665
	Direito	-,082	,212	,148	1	,700	,922	,608	1,397
	Psicologia e Sociologia	,147	,194	,573	1	,449	1,158	,792	1,693
	[Género=1]	1,647	,489	11,322	1	<,001	5,189	1,989	13,539
	[Género=2]	0[b]	.	.	0	.	.	.	.
	[Residência=1]	1,543	1,314	1,379	1	,240	4,679	,356	61,448
	[Residência=2]	2,077	1,359	2,335	1	,126	7,980	,556	114,514
	[Residência=3]	,363	1,875	,037	1	,847	1,437	,036	56,712
	[Residência=4]	1,103	1,687	,427	1	,513	3,013	,110	82,273
	[Residência=5]	-,763	2,546	,090	1	,764	,466	,003	68,482
	[Residência=6]	0[b]	.	.	0	.	.	.	.
	[Escola Proveniente=1]	-2,087	,684	9,299	1	,002	,124	,032	,474
	[Escola Proveniente=2]	-2,190	,816	7,204	1	,007	,112	,023	,554
	[Escola Proveniente=3]	-1,992	1,154	2,980	1	,084	,136	,014	1,309
	[Escola Proveniente=4]	-1,726	2,274	,576	1	,448	,178	,002	15,361
	[Escola Proveniente=5]	-1,866	1,317	2,008	1	,156	,155	,012	2,044
	[Escola Proveniente=6]	0[b]	.	.	0	.	.	.	.

a. A categoria de referência é: DCE.

b. Este parâmetro é definido para zero porque é redundante.

A probabilidade de transição da classe de referência para a classe DL&CHS é significativamente influenciada por alguns factores-chave, incluindo a nota nas disciplinas de Direito, Língua português, a Idade, o Sexo ($valor\text{-}p < 0{,}001$) e de forma marginal pela Média final do 12º Ano e pela localização de residência de alguns estudantes.

Cada aumento na nota das disciplinas Língua Portuguesa e Direito resulta em um aumento significativo nas chances de inscrição em um curso do DL&CHS, em 76,4% (100% ×

$(1{,}764 - 1))$ e $35{,}1\%$ $(100\% \times (1{,}351 - 1))$, respectivamente. Por outro lado, a cada incremento na Média Final do 12º Ano, as chances de inscrição em um curso do DL&CHS diminuem em $-34{,}3\%$. É relevante destacar que a probabilidade de homens se matricularem em cursos do DL&CHS ou DCE&EE é substancialmente maior, com percentagens de 219,9% e 418,9%, respectivamente, em comparação com o departamento de referência.

Por conseguinte, realizamos o processo de classificação dos estudantes nos cursos vinculados aos diferentes departamentos em comparação com a classificação ao acaso. Através dos modelos de regressão multinomial previamente ajustados, conseguimos calcular a probabilidade de cada aluno pertencer à classe na qual a chance de admissão é mais alta.

Ao analisar os dados da Tabela 33, percebemos que o grupo de estudantes do DCE&EE conta com cerca de 28 alunos, sendo que apenas um deles foi corretamente classificado. Por outro lado, observamos uma precisão mais elevada na classificação dos alunos do DL&CHS, onde mais de 80% dos estudantes foram classificados corretamente, ou seja, estão matriculados no departamento com maiores chances de admissão.

Para o departamento DCE, observamos uma quantidade considerável de classificações equivocadas, com mais de 40% dos alunos inscritos em cursos cujas chances de admissão eram inferiores.

De forma abrangente, aproximadamente 68,2% dos estudantes direcionaram suas escolhas para os cursos oferecidos pelos departamentos com as probabilidades de admissão mais elevadas.

Tabela 33 – Classificação dos candidatos por departamentos (Base de Dados II)

Classificação

		Previsto		
Observado	DL&CHS	DCE&EE	DCE	Porcentagem Correta
DL&CHS	235	0	53	81,6%
DCE&EE	18	1	9	3,6%
DCE	96	0	141	59,5%
Porcentagem global	63,1%	0,2%	36,7%	68,2%

Aproximadamente 45,81% dos estudantes estão matriculados nos cursos dos departamentos nos quais possuem as maiores probabilidades de admissão. Essa métrica é calculada com base nas proporções percentuais de classificação correta ao acaso, as quais são dadas por $\left(((0{,}521)^2 + (0{,}051)^2 + (0{,}429)^2) \times 100\%\right)$. O modelo empregado evidencia que a taxa de classificação correta é significativamente superior, excedendo em 20% aquela que seria esperada por pura casualidade.

CONCLUÕES

Este estudo sobre a modelação estatística aplicada ao processo de admissão de estudantes na Universidade de São Tomé e Príncipe (USTP) apresenta duas técnicas de análise: Regressão Linear Múltipla e Regressão Logística Multinomial. Ambas são precedidas por análises inferenciais de pressupostos subjacentes aos dados recolhidos, que revelaram a ausência de diferenças significativas nas médias de ensino secundário entre candidatos do sexo masculino e feminino a um nível de significância de 5%.

Na aplicação da primeira técnica, Regressão Linear Múltipla, observou-se uma forte correlação entre as variáveis do modelo, com mais de 75% da variação da média do ensino secundário explicada pelas variáveis exógenas incluídas. Utilizando o método *Stepwise*, os modelos apresentaram coeficientes estatisticamente significativos para as seguintes variáveis: notas em Biologia, Física/Química, Língua Portuguesa, Matemática, História, Direito, Psicologia/Sociologia, idade, e residência nos distritos de Cantagalo ou Lembá. Exceto pelas três últimas variáveis, um aumento de um ponto em qualquer uma dessas disciplinas resulta em um aumento na média final. Especificamente, os coeficientes padronizados evidenciam que as notas de Língua Portuguesa (notas de Direito na Base de Dados II) tiveram a maior contribuição para a média final, enquanto as de Matemática (notas de Língua Portuguesa na Base de Dados II) exerceram a menor contribuição. Por outro lado, verificou-se que a Média Final diminui em 0,211 pontos em uma base de dados e aumenta 0,009 pontos em outra à medida que a idade avança (conforme as Tabelas 18 e 23).

O modelo mostrou-se estatisticamente adequado, pois os pressupostos foram validados, tendo como categoria de referência estudantes do sexo masculino, residentes no distrito de Água Grande e que frequentaram o Liceu Nacional. Observou-se que as disciplinas selecionadas influenciam as decisões de seleção e seus impactos nos resultados académicos, contribuindo para a média final do ensino secundário. Notou-se também que algumas variáveis de residência e escolas não foram estatisticamente significativas para o modelo, considerando as categorias de referência utilizadas.

Na aplicação da segunda técnica, Regressão Logística Multinomial, foram consideradas quatro categorias correspondentes aos diferentes departamentos da USTP, nomeadamente DCNVA, DL&CHS, DCE&EE e DCE como categoria de referência. Ao avaliar as métricas de ajuste do modelo, constatou-se que pelo menos uma variável independente exerce um impacto significativo na admissão dos estudantes nos cursos vinculados aos respectivos departamentos (o $valor\text{-}p$ associado ao teste de Razão de Verossimilhança é inferior a 0,001).

Os modelos que incorporaram as variáveis idades, média final do 12º ano, notas em Matemática, Língua Portuguesa, História, Direito, Física/Química, Sociologia/Psicologia, género, residência e as escolas onde os estudantes concluíram o ensino secundário apresentam melhor ajuste em comparação ao modelo com apenas o intercepto (conforme as Tabelas 25 e 30). Entretanto, a qualidade do ajuste foi razoável na primeira base de dados e medíocre na segunda, conforme indicado pelos valores dos pseudo-R^2.

As variáveis idade, notas em Matemática, Direito e Biologia, género (com masculino como classe de referência), e a escola de conclusão do ensino secundário (com Liceu Nacional como classe de referência) demonstraram efeitos estatisticamente significativos sobre o *logit* da probabilidade de inscrição em um dos cursos dos departamentos de Ciências de Educação.

Na Base de Dados I, as chances de inscrição em cursos nos departamentos DCNVA, DL&CHS ou DCE&EE diminuem em 23,8%, 7,9% e 28,6%, respectivamente, por ano adicional de idade. Nos departamentos DL&CHS e DCE&EE, observou-se um baixo índice de classificação correta, com menos de 50% dos estudantes mais prováveis de serem admitidos efetivamente matriculando-se nesses departamentos.

Na Base de Dados II, nenhuma variável métrica permitiu distinguir o departamento DCE&EE do departamento de referência (DCE). Além disso, cada incremento nas variáveis analisadas resultou em uma redução nas chances de inscrição nos cursos do DCE&EE em comparação com os do DCE. Aproximadamente 68,2% dos estudantes direcionaram suas escolhas para os cursos oferecidos pelos departamentos com maiores chances de admissão. Observou-se que as chances de inscrição em um curso do DL&CHS diminuem em 34,3% para cada incremento na média final do 12º ano e aumentam em 76,4% e 35,1% para cada incremento nas notas de Língua Portuguesa e Direito, respectivamente. É importante destacar que o modelo empregado evidenciou que a taxa de classificação correta é significativamente superior, excedendo em 20% aquela esperada por pura casualidade.

Em suma, a modelação estatística aplicada ao processo de admissão na USTP, utilizando Regressão Linear Múltipla e Regressão Logística Multinomial, demonstra que as variáveis idade, género, notas em disciplinas específicas e escolas de origem dos candidatos possuem efeitos estatisticamente significativos tanto na média final do ensino secundário quanto na probabilidade de admissão nos diferentes departamentos, com ajustes estatísticos robustos e adequados que validam a eficácia dessas técnicas para otimizar os critérios de seleção da universidade.

RECOMENDAÇÕES, LIMITAÇÕES E PESQUISAS FUTURAS

- ✓ **Recomendações**

Em virtude dos resultados alcançados com este trabalho de pesquisa, é fundamental destacar que o número de variáveis utilizadas não esgota a possibilidade de explorar outras variáveis que possam ser determinantes e significativos para o sucesso académico.

Alem disso, propomos implementação de um Sistema Baseado em Dados, para que a USTP possa adotar um sistema de admissão orientado por dados e que incorpore análises estatísticas contínuas. Este sistema deve utilizar perfis preditivos que considerem múltiplas dimensões do candidato, superando a limitação das notas de exames de acesso ou dos dados obtidos a partir dos certificados dos candidatos.

- ✓ **Limitações**

Este estudo, apesar de suas contribuições, apresenta algumas limitações que merecem consideração a saber: i) Disponibilidade e Qualidade dos Dados: A análise foi limitada pela disponibilidade e qualidade dos dados históricos dos candidatos; ii) Generalização dos Resultados: A pesquisa focou exclusivamente na USTP, o que pode restringir a generalização dos resultados a outras instituições de ensino superior.

- ✓ **Pesquisas Futuras**

Ao reconhecer estas limitações e considerações, os resultados deste estudo revelam alguns pontos importantes a explorar, como *i)* Mudanças no Contexto Social e Educacional: A dinâmica dos sistemas educacionais e as necessidades do mercado de trabalho estão em constante evolução. Assim, é fundamental que os modelos e critérios utilizados sejam revistos periodicamente para refletir as mudanças sociais, económicas e tecnológicas; *ii)* Eficácia do Modelo: Avaliar a eficácia do modelo estatístico desenvolvido, validando-o com dados de admissões passadas e realizando simulações preditivas para testar sua robustez e precisão; *iii)* Explorar outros modelos lineares como os modelos multiníveis e comparar os resultados: compreender as variações em diferentes níveis; e *iii)* Inclusão de Novas Variáveis: explorar outras variáveis como desempenho escolar, antecedentes familiares, condição socioeconómica e região de origem dos estudantes que são determinantes significativos para o sucesso académico.

REFERÊNCIAS BIBLIOGRÁFICAS

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons. 3th Edition.

Costa, P. S. (2019). *Plano estratégico da Universidade de São Tomé e Príncipe – Quadriénio 2019-2023*. (Documento não publicado)

Decreto-lei n.º 9/2018 - *Estatuto da Universidade de S. Tomé e Príncipe*; Diário da República n.º 99 de 18 de Julho.

Demétrio, C. G. B., & Zocchi, S. S. (2006). Modelos de regressão. *Piracicaba: ESALQ*.

Fávero, L. P., & Belfiore, P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. Elsevier Brasil.

Ferreira, M. C. C. D. S. (2013). *Modelos de Regressão: uma aplicação em Medicina Dentária* (Dissertation). Mestrado em Estatística, Matemática e Computação (Ramo – Estatística Computacional) pela Universidade Aberta, Lisboa.

Gujarati, D. N., & Porter, D. C. (2008). *Basic Econometrics*, 5th Edition. Tradução: Denise Durante, Mônica Rosemberg e Maria Lúcia G. L. Rosa (2009). Econometria Básica, 5ª Edição. AMGH Editora. Porto Alegre: ISBN 978-0-07-337577-9

Hair, J. F. Jr et al. (2006). *Multivariate Data Analysis*, 6th Edition. Tradução: Adonai Schlup Sant'Anna (2009). Análise multivariada de dados. Porto Alegre: Bookman. ISBN 978-85-7780-534-1

Heckman, J.J., et al. (2006). *The Effects of Educational and Family Background on College Admission*. American Economic Review, 96(2), 517-521.

Hosmer Jr, DW, Lemeshow, S., & Sturdivant, RX (2013). *Applied Logistic Regression* 3rd Edition. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN 978-0-470-58247-3.

Jonhson, R. A., Wichern D. W. (2002) *Applied Multivariate Statistical Methods*, Prentice Hall

Kubrusly, J. (2014). *Notas de Aula Modelos Lineares* I-GET00138.

Lai n.º 4/2017 - *Lei do Regime Jurídico das Instituições do Ensino Superior*; Diário da República n.º 31 de 24 de Março.

Marôco, J., (2021). *Análise estatística com o SPSS Statistics*. 8. ed. Pêro Pinheiro, Portugal. Editora ReportNumber, ISBN: 978-989-96763-7-4.

Meyer, K., et al. (2020). *Factors Influencing University Admission: A Study in Higher Education*. Journal of Educational Research, 113(4), 504-517.

Montgomery, D. C., & Runger, G. C. (2009). *Estatística aplicada e probabilidade para engenheiros, 4ª. Ed. Rio de Janeiro: Editora LTC*.

Myers, R. H., Montgomery, D. C., Vining, G. G. & Robinson, T. J. (2012). *Modelos lineares generalizados: com aplicações em engenharia e ciências*. John Wiley e Filhos.

Nakamura, K. G. (2013). *Multicolinearidade em modelos de regressão logística* (Dissertation, Universidade de São Paulo).

Paula, G. A., (2004). *Modelos de regressão: com apoio computacional* (pp. 28-55). São Paulo: IME-USP.

Pestana, M. H., & Gageiro, J. N. (2020). *Análise de dados para ciências sociais: a complementaridade do SPSS*, 6ª Edição, 2ª Impressão – Lisboa, Outubro.

Reis, E., *et. al*. (2018). *Estatística Aplicada. vol. 2*, (6ª Edição). *Edições Sílabo*.

Ross, S. M (2010). *Introduction to probability models*. 10th ed. University of Southern California. Los Angeles, Califórnia

Vaz, F. E. D. C. R. (2020). *Modelação linear e extensões: aplicação da regressão logística no estudo de câncer da mama* (Dissertation).

Wooldridge, J. M. (2011). *Introdução à econometria: uma abordagem moderna*. Cengage Learning. Tradução da 4ª Edição Norte-Americana

APÊNDICES

A – Teste de Normalidade da Base de Dados I

Testes de Normalidade

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Género	Estatística	gl	Sig.	Estatística	gl	Sig.
Idade	Masculino	,203	124	<,001	,861	124	<,001
	Feminino	,184	442	<,001	,816	442	<,001
Média Final do 12º Ano	Masculino	,211	124	<,001	,877	124	<,001
	Feminino	,235	442	<,001	,845	442	<,001
História	Masculino	,213	124	<,001	,866	124	<,001
	Feminino	,243	442	<,001	,801	442	<,001

a. Correlação de Significância de Lilliefors

B – Teste de Normalidade da Base de Dados II

Testes de Normalidade

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Género	Estatística	gl	Sig.	Estatística	gl	Sig.
Idade	Masculino	,164	151	<,001	,814	151	<,001
	Feminino	,182	400	<,001	,750	400	<,001
Média Final do 12º Ano	Masculino	,284	151	<,001	,833	151	<,001
	Feminino	,209	400	<,001	,904	400	<,001
Matemática	Masculino	,368	151	<,001	,664	151	<,001
	Feminino	,308	400	<,001	,696	400	<,001

a. Correlação de Significância de Lilliefors

C - Testes de Comparação Múltipla (Base de Dados II).

Comparações por Método Pairwise de Curso Inscrito

Sample 1-Sample 2	Estatística de teste	Erro Padrão	Estatística de Teste Padrão	Sig.	Adj. Sig.[a]
DCE&EE-DCE	-59,614	30,491	-1,955	,051	,152
DCE&EE-DL&CHS	84,130	30,243	2,782	,005	,016
DCE-DL&CHS	24,516	13,253	1,850	,064	,193

Cada linha testa a hipótese nula em que as distribuições Amostra 1 e Amostra 2 são iguais. As significâncias assintóticas (teste de dois lados) são exibidas. O nível de significância é , 050.

a. Os valores de significância foram ajustados pela correção Bonferroni para vários testes.

D –Testes de Razão de Verosimilhança.

Testes de razão de verossimilhança

Efeito	Critérios de ajuste do modelo: AIC do modelo reduzido	BIC do modelo reduzido	Verossimilhança de log -2 do modelo reduzido	Testes de razão de verossimilhança: Qui-quadrado	df	Sig.
Intercepto	844,635	999,988	772,635[a]	,000	0	.
Idade	872,191	1018,913	804,191	31,556	2	<,001
Média Final do 12º Ano	845,406	992,128	777,406	4,771	2	,092
L. Portuguesa	859,738	1006,460	791,738	19,103	2	<,001
História	841,751	988,473	773,751	1,116	2	,572
Direito	852,240	998,962	784,240	11,604	2	,003
Psicologia e Sociologia	841,476	988,198	773,476	,841	2	,657
Género	864,580	1011,302	796,580	23,945	2	<,001
Residência	838,069	950,269	786,069	13,434	10	,200
Escola Proveniente	837,913	950,112	785,913	13,277	10	,209

A estatística qui-quadrado é a diferença no log de verossimilhanças -2 entre o modelo final e um modelo reduzido. O modelo reduzido é formado pela omissão de um efeito do modelo final. A hipótese nula significa que todos os parâmetros desse feito são 0.

a. Esse modelo reduzido é equivalente ao modelo final porque a omissão do efeito não aumenta os graus de liberdade.

E – Variáveis excluídas do modelo de regressão (Base de Dados II).

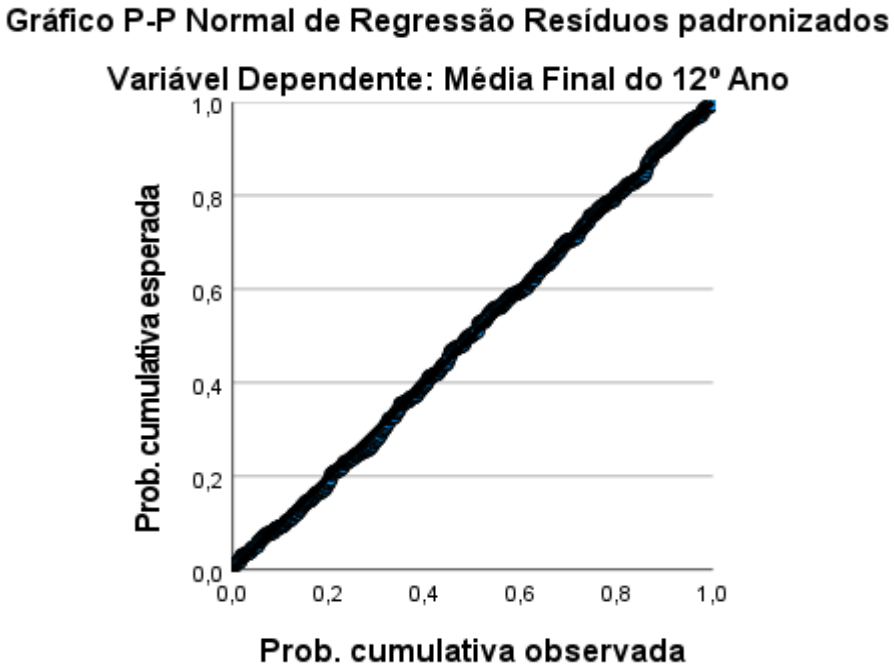
Variáveis excluídas[a]

Modelo		Beta In	t	Sig.	Correlação parcial	Estatísticas de colinearidade: Tolerância
7	Sexo=Feminino	-,034[h]	-1,619	,106	-,069	,967
	Resid=Mé-Zochi	,009[h]	,397	,692	,017	,784
	Resid=Lobata	,010[h]	,477	,633	,020	,967
	Resid=Caué&RAP	,002[h]	,111	,912	,005	,984
	EscProven=E. S. Maria Manuela Margarida	,014[h]	,642	,521	,027	,858
	EscProven=E. S. Santana	,007[h]	,188	,851	,008	,336
	EscProven=E. S. Neves	,047[h]	,870	,385	,037	,149
	EscProven=E. S. Mé Xinhô	-,020[h]	-,971	,332	-,042	,968
	EscProven=Outras Escolas	-,012[h]	-,562	,574	-,024	,898

a. Variável Dependente: Média Final do 12º Ano

h. Preditores no Modelo: (Constante), Direito, Psicologia e Sociologia, História, L. Portuguesa, Resid=Cantagalo, Idade, Resid=Lembá

F – Gráfico da Probabilidade Normal



G – Teste de Homogeneidade de variâncias (Base de Dados I).

Teste de Homogeneidade de Variância

		Estatística de Levene	gl1	gl2	Sig.
Média Final do 12º Ano	Com base em média	2,647	5	545	,022
	Com base em mediana	1,142	5	545	,337
	Com base em mediana e com gl ajustado	1,142	5	536,483	,337
	Com base em média aparada	2,294	5	545	,044

H – Teste de Homogeneidade de variâncias (Base de Dados II).

Teste de Homogeneidade de Variância

		Estatística de Levene	gl1	gl2	Sig.
Média Final do 12º Ano	Com base em média	,469	5	560	,799
	Com base em mediana	,686	5	560	,634
	Com base em mediana e com gl ajustado	,686	5	553,615	,634
	Com base em média aparada	,638	5	560	,670

I – Testes para amostras Independentes

Teste de amostras independentes

		Teste de Levene para igualdade de variâncias		teste-t para Igualdade de Médias							
						Significância				95% Intervalo de Confiança da Diferença	
		T	Sig.	t	df	Unilateral p	Bilateral p	Diferença média	Erro de diferença padrão	Inferior	Superior
Idade	Variâncias iguais assumidas	4,384	,037	-,199	549	,421	,842	-,080	,404	-,874	,713
	Variâncias iguais não assumidas			-,222	342,796	,412	,824	-,080	,362	-,793	,632
Média Final do 12º Ano	Variâncias iguais assumidas	,157	,693	-,256	549	,399	,798	-,03721	,14552	-,32305	,24863
	Variâncias iguais não assumidas			-,247	253,557	,402	,805	-,03721	,15044	-,33349	,25906
Matemática	Variâncias iguais assumidas	1,193	,275	-,213	549	,416	,832	-,03136	,14735	-,32079	,25807
	Variâncias iguais não assumidas			-,209	260,618	,417	,835	-,03136	,15011	-,32694	,26422

J – Testes para amostras Independentes

Teste de amostras independentes

		Teste de Levene para igualdade de variâncias		teste-t para Igualdade de Médias							
						Significância				95% Intervalo de Confiança da Diferença	
		T	Sig.	t	df	Unilateral p	Bilateral p	Diferença média	Erro de diferença padrão	Inferior	Superior
Média Final do 12º Ano	Variâncias iguais assumidas	,970	,325	1,817	564	,035	,070	,17165	,09447	-,01390	,35720
	Variâncias iguais não assumidas			1,728	184,60	,043	,086	,17165	,09936	-,02436	,36767
História	Variâncias iguais assumidas	5,029	,025	2,656	564	,004	,008	,35305	,13293	,09196	,61414
	Variâncias iguais não assumidas			2,537	185,64	,006	,012	,35305	,13918	,07848	,62763
Idade	Variâncias iguais assumidas	,204	,652	2,094	564	,018	,037	1,079	,515	,067	2,091
	Variâncias iguais não assumidas			2,081	195,63	,019	,039	1,079	,519	,056	2,102