Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Matemática

Uma proposta de análise de desempenho de estudantes do ENEM

Autor: Adenilson Almeida de Novaes

Orientador: Prof. Dr. Jean Piton Gonçalves

Disciplina: Trabalho de Conclusão de Curso

Profs Responsáveis: Profa. Dra. Luciene Nogueira Bertoncello
Profa. Dra. Natalia Andrea Viana Bedoya
Prof. Dr. Wladimir Seixas

São Carlos, 17 de dezembro de 2020.

Uma proposta de análise de desempenho de estudantes do ENEM

Autor: Adenilson Almeida de Novaes

Orientador: Prof. Dr. Jean Piton Gonçalves

Disciplina: Trabalho de Conclusão de Curso

Profs Responsáveis: Profa. Dra. Luciene Nogueira Bertoncello
Profa. Dra. Natalia Andrea Viana Bedoya
Prof. Dr. Wladimir Seixas

Este trabalho foi desenvolvido no Ensino Não Presencial Emergencial referente ao primeiro semestre de 2020 para o desenvolvimento da disciplina Trabalho de Conclusão de Curso.

São Carlos, 17 de dezembro de 2020.

Agradecimentos

Agradeço à minha família e, principalmente, aos meus pais por todo o apoio, incentivo e segurança que me proporcionaram. Todo conhecimento e oportunidade que tive foi graças a seus sacrifícios.

Também agradeço ao meu orientador, Prof. Dr. Jean Piton Gonçalves, pelo excelente trabalho e auxílio, bem como a todos os professores que permitiram a minha evolução até o momento.

Não poderia deixar de agradecer também aos meus amigos e aos colegas de curso, os quais me incentivaram e auxiliaram ao longo desta jornada.

Resumo

No campo educacional internacional, uma preocupação é predizer o desempenho do estudante para futuros cursos e/ou disciplinas. Nesse aspecto, os trabalhos de Rajalaxmi *et al.* (2019), Abledu (2012) e Gadhavi, Patel (2017) demonstram que uma abordagem estatístico-matemática via regressão múltipla geram modelos preditivos que podem auxiliar políticas educacionais, professores e gestores em tomadas de decisão. Partindo da Ciência de Dados Educacional, esse Trabalho de Conclusão de Curso utiliza dados reais do Exame Nacional do Ensino Médio (ENEM). Partindo de uma amostra de 26.731 participantes, mostra-se que é possível modelar o desempenho do estudante, de forma a correlacionar, multidimensionalmente, as quatro áreas de conhecimento do ENEM e, somado à isso, prover uma discussão do desempenho de estudantes oriundos de escolas públicas e particulares. Resultados mostram que a Ciência de Dados Educacional é um campo fundamental para a formação de professores de Matemática, elucidando, do ponto de vista quantitativo, as variáveis educacionais.

Palavras-chave: Enem, previsão de desempenho, regressão múltipla, formação de professores.

Abstract

In the international education field, it is relevant to predict the students' performance for future courses and/or disciplines. In that matter, Rajalaxmi *et al.* (2019), Abledu (2012) and Gadhavi, Patel (2017) works show a statistic-mathematical approach by multiple regression which generates predictives models that may help educational politics, professors and managers in decision making. Starting from Educational Data Science, this Undergraduate Thesis uses real data from the Exame Nacional do Ensino Médio (ENEM). Starting from a sample of 26,731 participants, it shows that it is possible to model the student performance, to then correlate multidimensionally the four knowledge pillars of the ENEM, and, on top of that, provide a discussion about the students' performance from public and private schools. Results show that Educational Data Science is a fundamental field to Mathematics professors' formation, elucidating, from the quantitative point of view, the educational variables.

Keywords: Enem, performance prediction, multiple regression, professors' formation.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

No campo educacional, uma preocupação de professores e gestores é o desempenho do estudante durante uma disciplina ou um curso. Na escola básica, a qualidade e quantidade de conhecimento aprendido pelo estudante durante sua trajetória escolar pode estar relacionada com o tipo da escola (pública ou particular), região onde a escola se encontra ou até mesmo a sua idade, por exemplo. Relativo ao desempenho de estudantes do Ensino Médio, o Exame Nacional do Ensino Médio (ENEM), avalia nacionalmente cerca de 6 milhões de inscritos (média dos últimos 5 anos) candidatos à vagas em diversas universidades públicas bem como faculdades particulares.

O ENEM é o segundo maior exame em larga escala do mundo, perdendo apenas para o Gaokao[1] (China). Quando recorremos à literatura, os trabalhos relacionados ao ENEM estão sempre ligados à uma análise quantitativa ou qualitativa dos itens da prova (SILVA; SANTIAGO; DOS SANTOS, 2013; TRAVITZKI, 2017) ou é visto como um instrumento pedagógico que auxilia no ensino, enquanto instrumento reflexivo para o professor e estudantes (WIENBUSCH, 2012).

Por outro lado, estudos que auxiliem na predição do desempenho dos estudantes do ENEM não existem no Brasil. No sentido da predição, Rajalaxmi *et al.* (2019), Abledu (2012) e Gadhavi, Patel (2017), buscam modelos preditivos que auxiliam em políticas educacionais, professores, gestores e até os próprios estudantes, fornecendo devolutivas para o ensino e a aprendizagem. Uma vez que não são foram encontrados trabalhos que estudem os resultados do ENEM do ponto de vista da modelagem do desempenho de estudantes, este TCC traz uma abordagem inédita sobre o assunto.

Foi motivado pela necessidade de explorar novas alternativas que possam contribuir positivamente com a Educação brasileira e com a formação como professor. Apesar do conteúdo de Regressão Linear não fazer parte do curso de Licenciatura em

[1] O exame de Gaokao é realizado na China e, diferente do ENEM, oferece apenas uma chance aos candidatos.

Matemática, através deste trabalho foi possível utilizá-lo dentro da perspectiva de previsão de desempenho de estudantes, gerando além dos resultados um conhecimento suficiente para abrir novas possibilidades no dia a dia escolar.

Partindo de um conjunto de dados reais de provas do ENEM ano 2012, este trabalho busca responder à seguinte questão: "é possível modelar o desempenho do estudante, de forma a correlacionar as quatro áreas de conhecimento do ENEM, permitindo analisar o desempenho entre escolas públicas e particulares"?

Buscando essa resposta, este TCC foi dimensionado da seguinte maneira:

- O Capítulo 2 aborda brevemente informações sobre a estrutura do ENEM e a matriz de competências da prova de Matemática e suas Tecnologias.
- O Capítulo 3 aborda o tema de previsão de desempenho, contextualizando desde a Ciência de Dados até as possibilidades do tema.
- O Capítulo 4 trata da Regressão Linear Simples e Múltipla, abordando a construção de modelos de regressão.
- O Capítulo 5 trata dos resultados obtidos após a aplicação de análise de regressão múltipla nos dados do Enem 2012.
- O Capítulo 6 aborda as conclusões obtidas pelo trabalho, relacionadas ao modelo de previsão e a questão escolar (pública versus particular).

Capítulo 2

ENEM

O Exame Nacional do Ensino Médio teve início no ano de 1998, idealizado pelo Ministério da Educação (MEC). O surgimento dele foi em um momento em que havia uma movimentação internacional pela padronização de avaliações, porém seguindo um caminho diferente ao cenário internacional, pois seu foco final era o aluno (TRAVITZKI, 2013).

Em 2009, o ENEM passou a ser utilizado como forma de ingresso para o Ensino Superior em universidades federais, as quais passaram pelo processo de mudança do vestibular individual da instituição para a utilização dos resultados do ENEM e se tornou o segundo maior exame do mundo, sendo o primeiro o Gaokao. A Tabela 2.1 traz o número de inscritos anualmente e os destaques daquele ano.

Ano	Inscritos	Destaque
2009	4.138.025	Mudança de formato e criação do SISU
2010	4.626.094	Resultados passam a ser utilizados pelo FIES
2011	5.366.949	53% dos participantes se declararam como negros e pardos
2012	5.791.066	Mudança nos critérios de isenção de taxa, resultou em 70% isentos
2013	7.173.910	ENEM se torna porta de acesso para quase todas as instituições públicas
2014	8.722.290	Expansão para Portugal, ENEM passa a ser aceito em 2 instituições
2015	7.792.024	Início da identificação de treineiros, chegando a 12% neste ano
2016	8.681.686	Lançamento de Aplicativo ENEM e maior segurança com biometria

2017	6.763.122	ENEM passa a ser aplicado em dois domingos
2018	5.513.712	Instituições portuguesas chegam a 35
2019	5.095.308	Maior índice de participação desde 2009, 77%

TABELA 2.1: Inscritos e destaques do ENEM (INEP, 2019).

Para o ano de 2020 o INEP iniciará testes com a realização do exame online, segundo o portal cerca de 100 mil candidatos irão participar dessa forma. Seguindo a evolução digital o Exame dá passos largos nesse sentido ao mesmo tempo que busca melhorar a segurança, como podemos observar no destaque de 2016 em que lançou um aplicativo de celular para auxiliar os participantes e melhorou a segurança através da implantação da biometria.

2.1 O Exame

O ENEM é um exame baseado em competências e habilidades. Até o ano de 2008 as 5 competências eram (INEP, 2018):

I - Dominar a norma culta da Língua Portuguesa e fazer uso das linguagens matemática, artística e científica.

II - Construir e aplicar conceitos das várias áreas do conhecimento para a compreensão de fenômenos naturais, de processos histórico-geográficos, da produção tecnológica e das manifestações artísticas.

III - Selecionar, organizar, relacionar, interpretar dados e informações representados de diferentes formas, para tomar decisões e enfrentar situações-problema.

IV - Relacionar informações, representadas em diferentes formas, e conhecimentos disponíveis em situações concretas, para construir argumentação consistente.

V - Recorrer aos conhecimentos desenvolvidos na escola para elaboração de propostas de intervenção solidária na realidade, respeitando os valores humanos e considerando a diversidade sociocultural.

Estas competências gerais que estavam presentes no Exame antigo permanecem como eixos cognitivos no Novo ENEM a partir de 2009: **dominar linguagens**, **compreender fenômenos**, **enfrentar situações-problema**, **construir argumentação** e **elaborar propostas**. O objetivo é o mesmo, porém atualmente são mais exploradas e direcionadas, deixando mais claro o que é esperado do candidato através

das competências de área (INEP, 2015). Estas competências devem ser aplicadas na realização da prova, a qual apresenta estrutura com 4 áreas do conhecimento e é composta por 180 questões e redação, a saber (INEP, 2009):

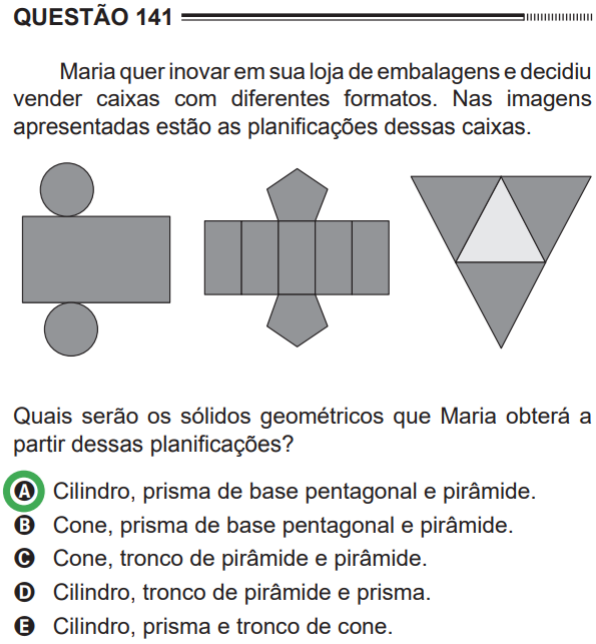
I - Linguagens, Códigos e suas Tecnologias.

II - Matemática e suas Tecnologias.

III - Ciências da Natureza e suas Tecnologias.

IV - Ciências Humanas e suas Tecnologias.

Cada área do conhecimento conta com 45 questões objetivas de múltipla escolha, de acordo com suas principais competências e habilidades[1]. As questões são estruturadas com a explanação da situação-problema ou contextualização do conteúdo seguido de 5 opções de resposta, como podemos ver os exemplos representados pelas Figuras 2.1 e 2.2 retiradas do Enem 2012.



QUESTÃO 141

Maria quer inovar em sua loja de embalagens e decidiu vender caixas com diferentes formatos. Nas imagens apresentadas estão as planificações dessas caixas.

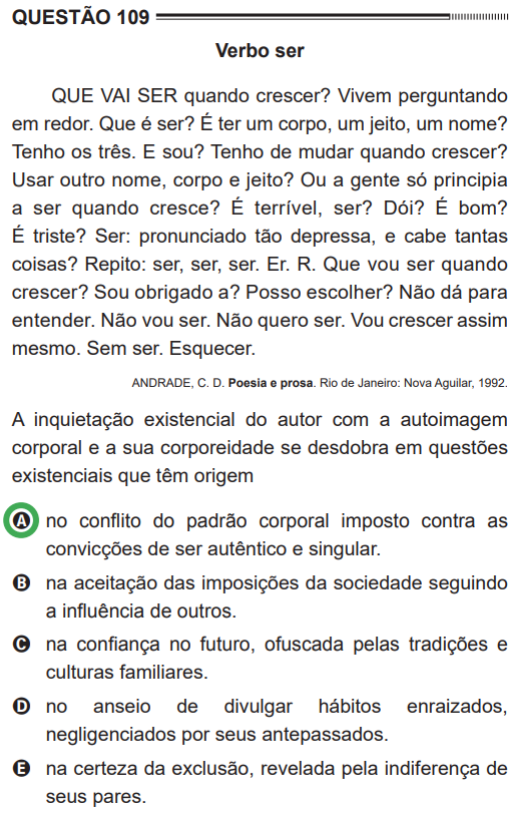
Quais serão os sólidos geométricos que Maria obterá a partir dessas planificações?

A Cilindro, prisma de base pentagonal e pirâmide.
B Cone, prisma de base pentagonal e pirâmide.
C Cone, tronco de pirâmide e pirâmide.
D Cilindro, tronco de pirâmide e prisma.
E Cilindro, prisma e tronco de cone.

FIGURA 2.1: Questão da prova de Matemática e suas Tecnologias no Enem 2012 (INEP, 2012).

Realizada em dois dias a prova tem tempo máximo de 4 horas e 30 minutos para a parte objetiva e, no segundo dia em que é realizada a redação é destinada mais 1 hora, resultando em 5 horas e 30 minutos. Além disso, a prova é dividida em 4 cadernos por dia, os quais dispõe de cores e números diferentes para identificação, sendo a relação do Exame de 2012 a seguinte: 1 - azul, 2 - amarelo, 3 - branco, 4 - rosa, 5 - amarelo,

[1]As habilidades são observadas de acordo com cada área do conhecimento e totalizam 30 por área.



QUESTÃO 109

Verbo ser

QUE VAI SER quando crescer? Vivem perguntando em redor. Que é ser? É ter um corpo, um jeito, um nome? Tenho os três. E sou? Tenho de mudar quando crescer? Usar outro nome, corpo e jeito? Ou a gente só principia a ser quando cresce? É terrível, ser? Dói? É bom? É triste? Ser: pronunciado tão depressa, e cabe tantas coisas? Repito: ser, ser, ser. Er. R. Que vou ser quando crescer? Sou obrigado a? Posso escolher? Não dá para entender. Não vou ser. Não quero ser. Vou crescer assim mesmo. Sem ser. Esquecer.

ANDRADE, C. D. **Poesia e prosa**. Rio de Janeiro: Nova Aguilar, 1992.

A inquietação existencial do autor com a autoimagem corporal e a sua corporeidade se desdobra em questões existenciais que têm origem

A) no conflito do padrão corporal imposto contra as convicções de ser autêntico e singular.
B) na aceitação das imposições da sociedade seguindo a influência de outros.
C) na confiança no futuro, ofuscada pelas tradições e culturas familiares.
D) no anseio de divulgar hábitos enraizados, negligenciados por seus antepassados.
E) na certeza da exclusão, revelada pela indiferença de seus pares.

FIGURA 2.2: Questão da prova de Linguagens e Códigos no Enem 2012 (INEP, 2012).

6 - cinza, 7 - azul e 8 - rosa. Esta divisão não implica que sejam questões diferentes, são os mesmos itens, porém organizados de 4 formas diferentes por dia resultando em 8 cadernos. Nos exemplos anteriores, Figuras 2.1 e 2.2, as questões foram retiradas do caderno amarelo número 5.

Quanto a pontuação no Exame, é calculada através da Teoria de Resposta ao Item (TRI) gerando a nota para cada uma das 4 áreas. Segundo Piton-Gonçalves (2020, p. 5):

> [...] a TRI propõe uma modelagem estatístico-matemática para as características latentes do examinado e modela a probabilidade de um indivíduo responder corretamente a um item em função do seu traço latente que é psicometricamente mapeado para um estimador θ.

2.2 Matemática e suas Tecnologias

Dentro das quatro áreas do ENEM iremos entender melhor sobre a estrutura da matriz de referência de Matemática. Nela temos o total de 7 competências de área, as quais guiam a formulação da prova objetiva e estão alinhadas aos conteúdos da educação básica. A organização é através de blocos temáticos: números, geometria, álgebra, grandezas e medidas, modelagem matemática, tratamento da informação e conhecimentos

de estatística e probabilidade. A partir desses blocos podemos elencar as competências (INEP, 2015, p. 5-7):

- **Competência de área 1**: construir significados para os números naturais, inteiros, racionais e reais.
- **Competência de área 2**: utilizar o conhecimento geométrico para realizar a leitura e a representação da realidade e agir sobre ela.
- **Competência de área 3**: construir noções de grandezas e medidas para a compreensão da realidade e a solução de problemas do cotidiano.
- **Competência de área 4**: construir noções de variação de grandezas para a compreensão da realidade e a solução de problemas do cotidiano.
- **Competência de área 5**: modelar e resolver problemas que envolvem variáveis socioeconômicas ou técnico-científicas, usando representações algébricas.
- **Competência de área 6**: interpretar informações de natureza científica e social obtidas da leitura de gráficos e tabelas, realizando previsão de tendência, extrapolação, interpolação e interpretação.
- **Competência de área 7**: compreender o caráter aleatório e não-determinístico dos fenômenos naturais e sociais e utilizar instrumentos adequados para medidas, determinação de amostras e cálculos de probabilidade para interpretar informações de variáveis apresentadas em uma distribuição estatística.

Por sua vez, as competências estão diretamente relacionadas aos blocos temáticos, mas não exclusivamente. A competência de área 1 se relaciona com o bloco **numérico**, podendo conter questões sobre identificação dos números ou operações, ordenando ou construindo códigos que se relacionam ao dia-a-dia, em revistas científicas, jornais ou jogos.

As habilidades que estão na competência de área 2 são relacionadas ao uso de **geometria**. Aqui o conteúdo das questões são relacionados a todas as possibilidades geométricas, como a sua presença na arte e arquitetura até no esporte. Também, relacionando com o cotidiano, o qual é repleto de geometria e possibilita ações como identificação, interpretação e o uso da percepção espacial para solucionar questões, a Figura 2.1 apresentada anteriormente exemplifica claramente este caso.

Na competência de área 3 temos habilidades relacionadas à noções de **grandezas e medidas**. São questões que variam do ato de selecionar instrumentos de medida para as situações até identificar e relacionar unidades de medida à grandezas.

Na área 4 temos questões similares, com objetivo de buscar no participante a habilidade de identificar interdependências entre duas grandezas nas situações-problema em que se busca uma solução.

A competência de área 5 busca habilidades relacionadas ao conteúdo **algébrico/geométrico** para resolução de situações-problemas através de generalizações e interpretações. Na área 6 as habilidades são exigidas para **tratamento da informação**, as quais muitas vezes são apresentadas como tabelas e gráficos.

Por fim, a competência de área 7 que traz questões relacionadas à exploração de fenômenos naturais e sociais, busca no participante as habilidades dentro do conteúdo de **probabilidade e estatística**. Esta separação é implícita, ou seja, não há uma separação das 45 questões pelas competências de área, além do fato de uma questão não ser apenas relativa à uma das competências.

Capítulo 3

Previsão de desempenho de estudantes

Pertencente à Ciência de Dados Educacionais, a área do conhecimento de previsão de desempenho acadêmico de estudantes (em inglês *Predicting Student Academic Performance*) permite diversas abordagens e utilizações para aprimorar cada vez mais a qualidade educacional, variando desde a descoberta de problemas em sistemas educacionais e processos de ensino até o desempenho de um estudante ao decorrer de um período letivo.

Identificar um problema de desempenho e agir sobre ele pode ser fundamental para melhorias em taxas de aprovação e metas de ensino, como por exemplo, o Plano Estadual de Educação de São Paulo[1]. A Ciência de Dados Educacionais é uma área de pesquisa recente e com raros trabalhos no país. Dessa forma será abordado brevemente a ciência de dados, a previsão baseado em dados finalizando com a contextualização sobre a aplicação destes conceitos no cenário educacional com alguns exemplos.

3.1 Ciência de dados

A Ciência de Dados (em inglês *Data Science*) surge da necessidade do processamento e a interpretação de uma grande quantidade de dados, que está ligado à diversos setores da economia, tais como o governamental, industrial, marketing, bancos, medicina, dentre outros (CAMILO, 2009, p. 2). Segundo Porto, Ziviani (2014, p. 2) sua base de formação é a ciência da computação, modelagem, estatística e matemática, de forma a desenvolver aptidão para enfrentar desafios como, organização e gerenciamento de dados, análise de dados e análise de redes complexas.

Essa evolução e o desafio de gerir grandes quantidades de dados já era esperada no ano de 2006, quando a Sociedade Brasileira de Computação divulgou o

[1]Disponível em https://www.fde.sp.gov.br/

relatório sobre a expectativa para os próximos 10 anos (PORTO; ZIVIANI, 2014, p. 5). Porém a evolução foi além da expectativa, ganhou força com a evolução da computação em nuvem (*cloud computing*) a partir dos anos 2000 junto com empresas como a Amazon, Google, Microsoft e até a rede de *streaming* Netflix, desenvolvendo e oferecendo este serviço a custos acessíveis para seus consumidores.

Com a evidência e necessidade deste profissional, foram surgindo mais campos de aplicação, basicamente o que pode gerar e armazenar dados também pode receber aplicações da Ciência de Dados. O principal destaque dentro dessa área do conhecimento hoje seria o profissional capaz de transformar os dados em informação, uma vez que, somos ricos em dados mas pobres em informação (HAN *et al.*, 2011). É uma crítica antiga que se aplica ao cenário atual, uma vez que o armazenamento evoluiu muito mais rápido que a formação de profissionais qualificados para o objetivo. Hoje temos, segundo o portal Brasil Escola[2], apenas 10 faculdades e universidades que oferecem este curso entre modalidade presencial e Educação à Distância (EAD).

Ocupam este cargo hoje profissionais das áreas correlatas, como ciência da computação, sistemas da informação e engenharia de computação. Porém ao sair do âmbito empresarial podemos encontrar aplicações para cientistas de dados em outros campos como na Educação.

Quando citamos a palavra "previsão"nos lembramos de algo mágico e ilógico esquecendo que nada mais é do que a identificação de padrões. Loh (2014, p. 17) reflete que identificar padrões é antigo e aparece na previsão de variações do tempo, fases da Lua e seus eclipses, plantio e colheita, dentre outros. Ainda segundo Loh (2014), a geração de um modelo pode ser utilizada para diversas ações, uma delas a previsão de eventos como já citado, uma vez que o modelo é uma generalização da realidade. Contudo, um modelo pode ser impreciso e não corresponder a uma visão correta da realidade. Este erro é esperado e gera a necessidade de realizar confirmações acerca do mesmo, podendo ser experimentos controlados ou identificação do grau de erro.

No que se refere às previsões do desempenho educacional de estudantes, os dados de origem podem ser os mais variados, tais como boletins, notas de avaliações, questionários, dentre outros. A partir de um determinado conjunto de dados é possível, por exemplo, inferir o desempenho de um estudante em um curso de graduação (certamente haverá uma margem de erro e esse é um ponto importante na Ciência de Dados) ou mesmo se um estudante que terminou o ensino médio atingirá bons níveis de desempenho enquanto calouro em uma universidade.

Prever o desempenho do estudante possibilita uma tomada de ação que

[2]Disponível em https://vestibular.brasilescola.uol.com.br/profissoes-futuro/ciencia-de-dados.htm, 02/09/2020.

melhore o cenário atual. Esta melhora estará diretamente relacionada à realidade do estudo, se for relativo ao desempenho de um estudante ao longo de uma disciplina, curso ou etapa de ensino. Conhecendo o desempenho, há possibilidade de intervenção para evitar a evasão, problema ainda atual e relevante no Brasil (SANTOS, 2020), ou a reprovação deste estudante. Esse tipo de intervenção pode ser feito pelos professores ou gestores escolares e, para elucidar esta proposta, temos como exemplo o projeto realizado na Universidade do Sul de Illinois (Carbondale), o qual foi explanado no trabalho de Soule (2017) sobre a previsão de sucesso estudantil através da regressão logística, tópico abordado no Capítulo 3, Seção 2.

A literatura traz outras pesquisas que utilizam-se da regressão linear enquanto ferramenta da análise de desempenho. Citamos Rajalaxmi *et al.* (2019), que utiliza informações do uso da internet por graduandos em Engenharia para prever seu desempenho. Abledu (2012) faz estudo buscando por problemas administrativos e técnicos que possam influenciar nas avaliações de cursos Politécnicos em Gana, Gadhavi, Patel (2017), buscam por um modelo que possa ajudar os próprios alunos a preverem seu desempenho em determinado conteúdo.

Além da regressão linear há outras formas de se trabalhar com previsão de desempenho, talvez menos utilizadas por serem mais robustas como a rede neural artificial[3] e árvore de decisão[4]. No trabalho de Ibrahim, Rusli (2007) é feita a comparação destes três modelos para previsão de desempenho de estudantes, o objetivo é prever a métrica "média de pontos cumulativos"(CGPA[5]) após a graduação através da inserção de variáveis preditoras como o perfil demográfico e a pontuação acumulada no primeiro ano de curso. Para isso é utilizado o *software SAS Enterprise Miner*[6], ferramenta mais robusta que permite análises avançadas tanto descritivas quanto preditivas.

Após aplicar os três modelos, Ibrahim e Rusli fazem uma comparação entre eles utilizando como balizador a média de erro quadrático e encontra o mais efetivo como sendo a rede neural artificial. Vale ressaltar que todos os modelos foram excelentes e a diferença entre eles é pequena, todos apresentaram mais de 80% de precisão. É possível

[3]Em termos intuitivos, redes neurais artificiais (RNAs) são modelos matemáticos inspirados nos princípios de funcionamento dos neurônios biológicos e na estrutura do cérebro. Estes modelos têm capacidade de adquirir, armazenar e utilizar conhecimento experimental e buscam simular computacionalmente habilidades humanas tais como aprendizado, generalização, associação e abstração (GOLDSCHMIDT, 2010, p.72).

[4]"Árvores de decisão (ou de classificação) são usadas para catalogar uma instância ou objeto dentro de um grupo pré-definido de classes com base nos atributos que ele tem, sendo uma técnica útil na exploração dos conjuntos de dados", segundo Silva Osses (2020 apud ROKACH; MAIMON, 2008).

[5]Do inglês *Cumulative Grade Point Average*, esta métrica é a média de pontuações acumuladas ao longo do curso, no Brasil seria a média geral final do graduado.

[6]O SAS Enterprise Miner é uma ferramenta de software point-and-click que permite aos usuários ampliar a sua capacidade de análise, aplicando algoritmos sofisticados para as funcionalidades das análises descritivas e preditivas (IESB, Centro Universitário).

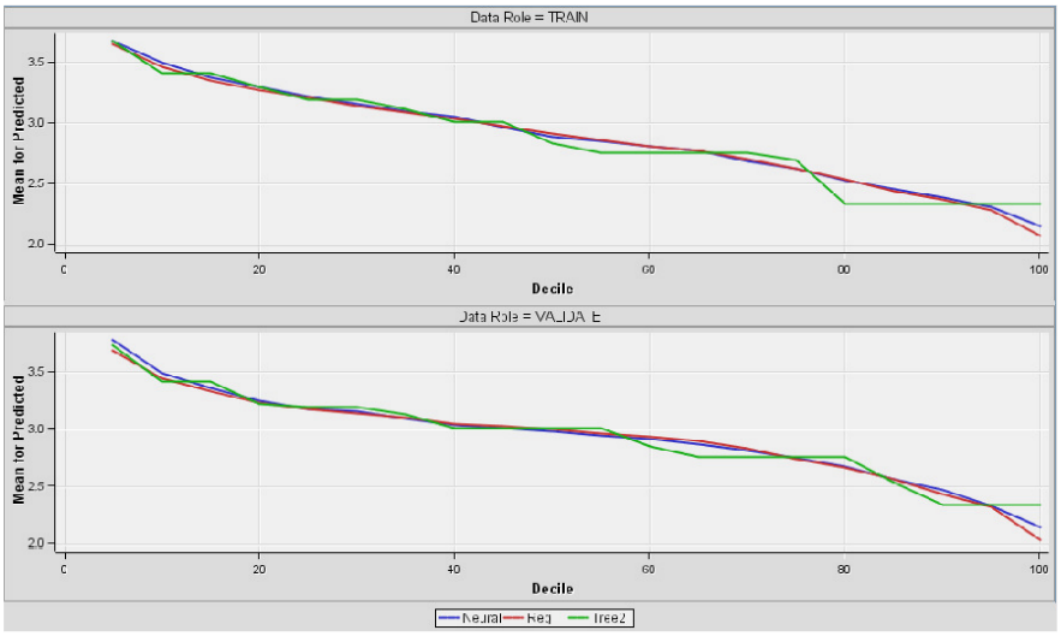


FIGURA 3.1: Comparação de modelos (IBRAHIM; RUSLI, 2007, p. 7).

seguir com qualquer um deles e, em um cenário diferente onde não há ferramentas robustas como o *software SAS*, possivelmente a opção mais adequada seria a regressão linear, presente nos demais estudos e também ferramenta de análise deste presente estudo.

Capítulo 4

Regressão Linear

De acordo com Maroco (2003), a Regressão Linear (RL) se refere ao conjunto de métodos estatísticos utilizados com o objetivo de modelar relações entre variáveis, as quais possibilitam a análise de comportamentos dentro de um conjunto de dados e pode se estender até algo mais complexo como realizar a predição de uma ou mais variáveis.

Tomemos alguns exemplos para facilitar a compreensão da RL:

a) em seu trabalho, Rajalaxmi *et al.* (2019) utilizou a RL para prever o desempenho de estudantes de Engenharia através de dados relacionados à forma como consomem internet, seu tempo em redes sociais e a nota acumulada no curso;

b) o trabalho de Bastos, Guimarães, Severo (2015) mostrou como o uso da RL pode auxiliar em tomadas de decisões empresariais, como a mudança na forma de alocar investimentos para manter o crescimento da empresa;

c) em Tavares, Pacheco, Borges (2016) o modelo de RL é aplicado para explicar os preços das diárias de quartos em hotéis a beira-mar, utilizando dados como o tipo de quarto, a época do ano e as vistas (se é para o mar ou para a terra).

As relações estudadas entre variáveis na RL são dividas em duas frentes: Regressão Linear Simples (RLS) e Regressão Linear Múltipla (RLM). Na RLS a relação estudada é entre uma variável dependente, Y, e uma variável independente, X. Por outro lado, na RLM essa relação é entre a variável dependente, Y, e 2 ou mais variáveis independentes $(X_1, X_2, ..., X_n)$. Neste capítulo iremos adentrar à RLS, iniciando pelo modelo teórico, passando em campos como a estimação dos parâmetros do modelo e correlação linear.

4.1 Regressão Linear Simples

A análise de regressão linear simples mostra a relação entre duas variáveis, chamadas de resposta e preditora ou variável dependente e variável independente, respectivamente. Essa relação simples é representada por um modelo estatístico, ou seja, uma equação que associa n pares de valores X_i e Y_i de forma linear, sendo:

$$Y_i = \alpha + \beta X_i + u_i \tag{4.1}$$

onde,

- Y_i é a variável dependente, com $i = 1, 2, \ldots, n$;
- X_i é a variável independente, com $i = 1, 2, \ldots, n$;
- α é parâmetro, também denominado coeficiente linear ou termo constante da equação;
- β é parâmetro, também denominado coeficiente de regressão ou inclinação da reta regressora;
- u_i representa o erro entre o modelo e o valor real.

Estabelecendo o modelo de RLS, pressupomos que:

1. a relação entre X e Y é linear;
2. os valores de X são fixos, dessa forma X não pode ser uma variável aleatória;
3. a média do erro é nula, ou seja, $E(u_i) = 0$, em que E é a Esperança Matemática;
4. para algum valor de X a variância do erro u é sempre σ^2, ou seja, $E(u_i^2) = \sigma^2$;
5. os erros são independentes, ou seja, o erro de uma observação não tem correlação com outro: $E(u_i u_j) = 0$ para $i \neq j$, com $i = 1, \ldots, n$ e $j = 1, \ldots, n$;
6. os erros seguem uma distribuição normal[1].

De acordo com os pressupostos 3, 4 e 6, podemos concluir que os erros u_i seguem uma distribuição normal com média 0 e variância σ^2, ou seja, $u_i \sim N(0, \sigma^2)$.

Além desses pressupostos, há um requisito básico para se estabelecer a análise de regressão: são necessários no mínimo 3 observações, uma vez que com apenas

[1]Distribuição normal é um padrão de comportamento, o qual é caracterizado por uma curva que em seu ponto máximo se torna simétrica. Pode-se encontrar mais em Morettin e Bussab (2017).

2 observações a determinação da reta se torna um problema de geometria analítica, não sendo necessário estimar os parâmetros.

Ao iniciar a análise de regressão, o primeiro passo será a determinação dos parâmetros α e β através de estimativas a e b. Para estimar os valores é necessário uma amostra com n pares de observações, X_i e Y_i onde i varia de 1 até n. Com isso teremos:

$$\hat{Y}_i = a + bX_i \tag{4.2}$$

onde $\hat{Y}_i$, a e b são as estimativas de $E(Y_i) = \alpha + \beta X_i$, α e β respectivamente.

Através de cada par de valores é possível estabelecer o desvio:

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i) \tag{4.3}$$

Graficamente teremos n pontos no gráfico (de dispersão) com coordenadas X_i e Y_i, onde os pontos descrevem uma reta considerando os parâmetros estimados a e b. Uma forma de estimar tais parâmetros é por meio do Método dos Mínimos Quadrados.

4.1.1 Mínimos quadrados para RLS

O método dos mínimos quadrados, que foi idealizado por Carl Friedrich Gauss ente 1777 e 1855, baseia-se na minimização das somas dos quadrados dos desvios (e_i), ou seja:

$$Z = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2 \tag{4.4}$$

Para encontrar os valores que minimizam a Equação (4.4) tem-se como condição necessária a identificação dos pontos críticos, ou seja, os valores em que a aplicação das derivadas parciais de Z em relação a a e b resultam em zero. Após a identificação de tais pontos, será possível encontrar os mínimos, logo:

$$\frac{\partial Z}{\partial a} = -2 \sum_{i=1}^{n} [Y_i - (a + bX_i)] = 0 \tag{4.5}$$

$$\frac{\partial Z}{\partial b} = 2 \sum_{i=1}^{n} [Y_i - (a + bX_i)](-X_i) = 0 \tag{4.6}$$

O somatório dos desvios resulta em zero, $\Sigma e_i = 0$, uma vez que a Equação (4.5) é condição necessária na estimação dos parâmetros. A Figura 4.1 mostra uma

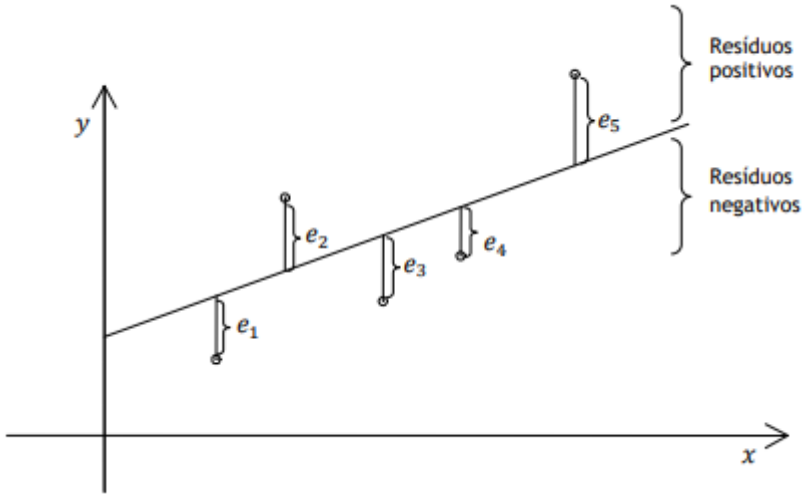


FIGURA 4.1: Representação gráfica dos resíduos (RODRIGUES, 2012)

representação gráfica dos desvios.

Simplificando as Equações (4.5) e (4.6) teremos:

$$\begin{cases} na + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{cases} \tag{4.7}$$

Com a resolução do sistema chega-se a:

$$a = \frac{(\sum_{i=1}^{n} X_i^2)(\sum_{i=1}^{n} Y_i) - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} X_i Y_i)}{n(\sum_{i=1}^{n} X_i^2) - (\sum_{i=1}^{n} X_i)^2} \tag{4.8}$$

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n(\sum_{i=1}^{n} X_i^2) - (\sum_{i=1}^{n} X_i)^2} \tag{4.9}$$

Por padrão encontra-se primeiro o valor de b e depois substitui seu valor na Equação (4.8) que pode ser simplificada em:

$$a = \frac{\sum_{i=1}^{n} Y_i}{n} - b\frac{\sum_{i=1}^{n} X_i}{n} \tag{4.10}$$

ou em

$$a = \overline{Y} - b\overline{X} \tag{4.11}$$

Lembrando que os índices $\overline{Y}$ e $\overline{X}$ representam as médias aritméticas de X_i e Y_i.

Com isso, estabelecemos as seguintes propriedades:

a) $\sum_{i=1}^{n} X_i e_i = 0$, a soma dos produtos dos desvios pelos respectivos valores da variável independente é nula.

b) $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$, a soma dos produtos dos desvios pelos respectivos valores estimados da variável dependente é nula.

4.1.2 Correlação linear

A análise de Correlação Linear (CL) busca explicar a relação entre duas variáveis e como elas estão relacionadas, ou seja, o quanto uma depende da outra. Na RLS é necessário distinguir a variável dependente da independente, por outro lado, na CL não será necessário.

A relação que iremos ver a seguir é definida pelo coeficiente de correlação (r) para uma amostra de n pares de valores X_i e Y_i, com $i = 1, 2, \ldots, n$. Buscando a medida de correlação de forma a evitar influências de média e variância (tendência central e dispersão, respectivamente), serão utilizadas as seguintes variáveis reduzidas:

$$v_i = \frac{X_i - \overline{X}}{s(X)} = \frac{x_i}{\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n-1}}} \tag{4.12}$$

e

$$z_i = \frac{Y_i - \overline{Y}}{s(Y)} = \frac{y_i}{\sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n-1}}} \tag{4.13}$$

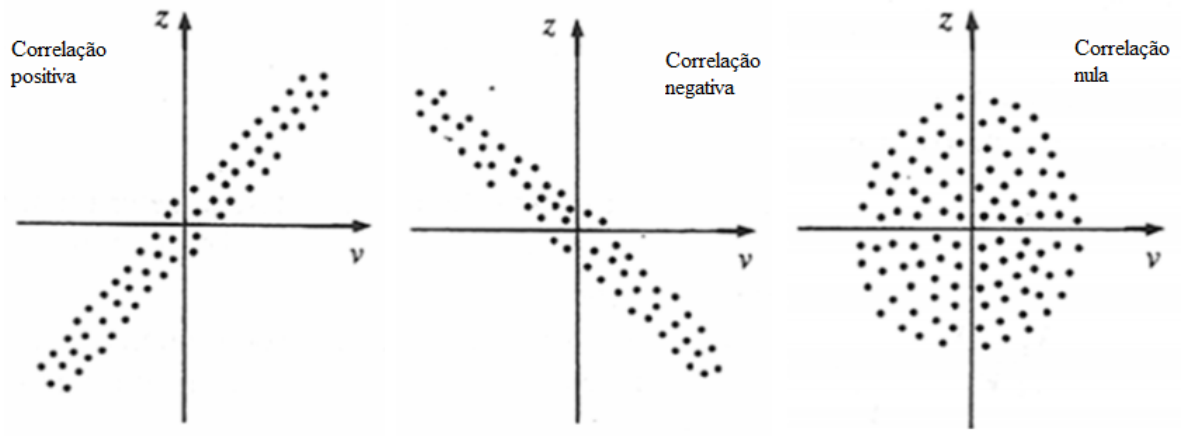


FIGURA 4.2: Classificação de correlações (HOFFMANN, 2016)

Com o resultado anterior, podemos representar através de gráficos de dispersão os pontos $(v_i,\ z_i)$ de acordo com a escala de valores para o coeficiente de correlação, o qual varia de -1 a 1. Tomando os exemplos de Hoffmann (2016), temos os gráficos da Figura (4.2), em que a correlação positiva mostra que os valores de X e

Y tendem a variar no mesmo sentido, a negativa mostra que tendem a variar no sentido contrário e ainda temos o valor de correlação igual ou próximo de zero, o qual demonstra a ausência de correlação linear. O coeficiente de correlação pode ser apresentado de forma simplificada como:

$$r = \frac{\sum_{i=1}^{n} v_i z_i}{n-1} \tag{4.14}$$

Considerando as Equações (4.12) e (4.13) podemos simplificar o cálculo em:

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{4.15}$$

Além de positiva ou negativa, a CL pode ser classificada de acordo com a sua força, ou seja, quanto mais próxima dos extremos maior a certeza de que há relação entre as duas variáveis. Na tabela a seguir, veremos a classificação da CL de Pearson.

Valores do coeficiente	Classificação da CL
$r = 1,0$	Perfeita positiva
$0,8 \leq r < 1,0$	Forte positiva
$0,5 \leq r < 0,8$	Moderada positiva
$0,1 \leq r < 0,5$	Fraca positiva
$0 \leq r < 0,1$	Ínfima positiva
$0,0$	Nula
$0,1 \leq r < 0,0$	Ínfima negativa
$0,5 \leq r < 0,1$	Fraca negativa
$0,8 \leq r < 0,5$	Moderada negativa
$1,0 \leq r < 0,8$	Forte negativa
$r = -1,0$	Perfeita negativa

TABELA 4.1: Interpretação do coeficiente de correlação de Pearson (RODRIGUES, 2012)

4.1.3 Exemplos para RLS

Com o objetivo de aplicar o Método dos Mínimos Quadrados e suas interpretações das correlações, criamos três exemplos, mostrados a seguir.

Exemplo 1) Iremos analisar se há correlação entre a porcentagem de abstenção

no ENEM 2012 e a faixa etária dos inscritos. Os dados para essa análise estão na Tabela (4.2), representados graficamente na Figura (4.3) e foram retirados do Relatório Pedagógico de 2011-2012.

Faixa etária (em anos)	Abstenção (%)
Até 16	13,8
17	12,6
18	18,7
19	25,6
20	30,5
21	34,2
22	37,2
23	39,4
24	41,1
25 a 29	43,8
30 a 39	45,3
40 a 49	42,6
Acima de 50	39,6

TABELA 4.2: Dados de abstenção por faixa etária ENEM 2012 (INEP, 2015)

Definido os dados e o objetivo, vamos calcular a correlação através da Equação (4.16) e encontrar o modelo de RLS através das Equações (4.8) e (4.9), mas primeiro calculamos os parâmetros individualmente:

$\sum_{i=1}^{n} X_i = 16 + 17 + \cdots + 49 + 50 = 1.155$

$\sum_{i=1}^{n} Y_i = 0,138 + 0,126 + \cdots + 0,411 + 0,411 + \cdots + 0,411 + 0,438 = 13,910$

$\sum_{i=1}^{n} X_i Y_i = (16 \cdot 0,138) + \cdots + (40 \cdot 0,426) + \cdots + (49 \cdot 0,426) + (50 \cdot 0,396) = 477,829$

$\sum_{i=1}^{n} X_i^2 = 16^2 + 17^2 + \cdots + 50^2 = 41.685$

$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = (16^2 + \cdots + 50^2) - \frac{1.155^2}{35}$
$\sum_{i=1}^{n} x_i^2 = 41.685 - 38.115 = 3.570$

$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = (0,138^2 + \cdots + 0,396^2) - \frac{13,91^2}{35}$

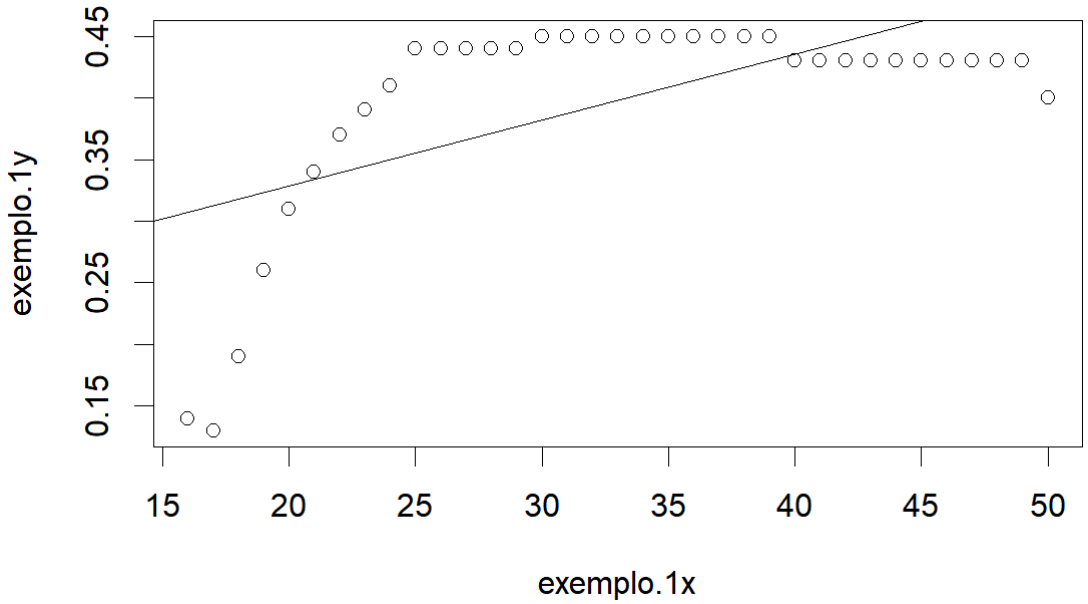


FIGURA 4.3: Representação gráfica do Exemplo 1. Fonte: O trabalho.

$\sum_{i=1}^{n} y_i^2 = 5,793 - 5,528 = 0,265$

$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} = (16 \cdot 0,138 + \cdots + 50 \cdot 0,396) - \frac{1.155 \cdot 13,91}{35}$
$\sum_{i=1}^{n} x_i y_i = 477,829 - 459,030 = 18,799$

Finalmente, iremos verificar se há correlação.

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \frac{18,799}{\sqrt{3.570 \cdot 0,265}} = 0,611$$

Com isso, conseguimos concluir que há correlação e podemos classificá-la como moderada positiva, pois $0,5 \leq r <$ 0,8.

A seguir calcularemos os parâmetros do modelo de RLS para encontrar o modelo para estes dados.

$$a = \frac{(41.685)(13,910) - (1.155)(477,829)}{(35)(41.685) - (1.155)^2} = 0,224$$

$$b = \frac{(35)(477,829) - (1.155)(13,910)}{(35)(41.685) - (1.155)^2} = 0,005$$

Portanto o modelo de RLS pode ser escrita como $\hat{Y}_i = 0,224 + 0,005X_i$. É notável pela correlação que conforme se aumenta a idade há uma tendência em aumentar a porcentagem de abstenção, este fato fica claro pela linha do modelo de RLS no gráfico

da Figura (4.3) e pela função ser crescente. Também podemos destacar que a linha que representa o modelo não segue a distribuição dos pontos tão precisamente, uma vez que a correlação é moderada.

Exemplo 2) Vamos analisar a amostra de dados referentes a altura e idade da Tabela (4.3) e representados no gráfico da Figura (4.4).

Idade (anos)	Altura (cm)
10	155
11	158
12	161
13	164
14	167
15	170
16	173

TABELA 4.3: Altura vs idade. Fonte: Autor

Aplicando as fórmulas individuais, temos:

$\sum_{i=1}^{n} X_i = 10 + 11 + \cdots + 15 + 16 = 91$

$\sum_{i=1}^{n} Y_i = 155 + 158 + \cdots + 170 + 173 = 1.148$

$\sum_{i=1}^{n} X_i Y_i = (10 \cdot 155) + (11 \cdot 158) + \cdots + (16 \cdot 173) = 15.008$

$\sum_{i=1}^{n} X_i^2 = 10^2 + 11^2 + \cdots + 16^2 = 1.211$

$\sum_{i=1}^{n} x_i^2 = 1.211 - \frac{8.281}{7} = 28$

$\sum_{i=1}^{n} y_i^2 = 188.524 - \frac{1.317.904}{7} = 252$

$\sum_{i=1}^{n} x_i y_i = 15.008 - \frac{104.468}{7} = 84$

Substituindo os valores na equação de correlação temos:

$$r = \frac{84}{\sqrt{28 \cdot 252}} = 1,0$$

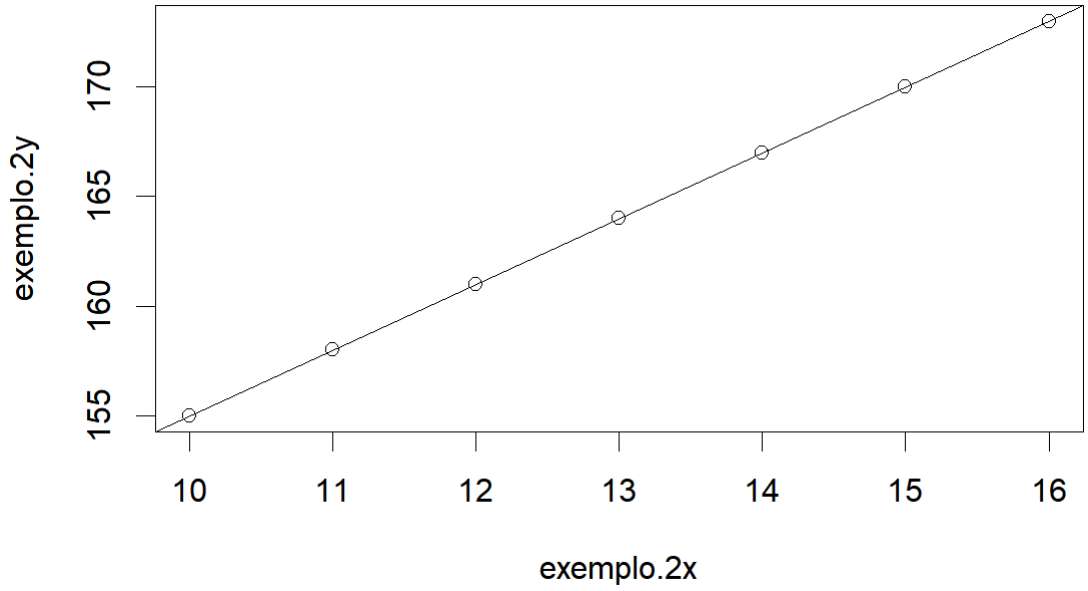


FIGURA 4.4: Representação gráfica do Exemplo 2. Fonte: O trabalho.

Após finalizar os cálculos temos a correlação $r = 1,0$, ou seja, perfeita positiva entre a altura e a idade da amostra. Para calcular os parâmetros do modelo de RLS vamos utilizar as Equações (4.9) e (4.10):

$$b = \frac{(7)(15.008) - (91)(1.148)}{(7)(1.211) - (91)^2} = 3,0$$

$$a = \frac{\sum_{i=1}^{n} Y_i}{n} - b\frac{\sum_{i=1}^{n} X_i}{n} = \frac{1.148}{7} - b\frac{91}{7} = 125,0$$

Finalmente, a equação de RLS para este exemplo é $\hat{Y} = 125 + 3X_i$ e através de sua representação gráfica na Figura (4.4) podemos entender melhor como acontece o ajuste perfeito entre os pontos e o modelo, uma vez que o valor do coeficiente de correlação foi $1,0$.

Exemplo 3) Neste último exemplo, iremos calcular a correlação entre a temperatura e a venda de cerveja, com a amostra de dados da Tabela (4.4) representada graficamente na Figura (4.5).

Temperatura (°C)	Vendas (R$)
20	1.400
22	2.000
25	1.300
27	2.050

30	1.000
32	1.250
33	2.200

TABELA 4.4: Temperatura do dia versus vendas de cerveja.
Fonte: Autor

Calculando os termos individualmente, temos:

$\sum_{i=1}^{n} X_i = 20 + 22 + \cdots + 32 + 33 = 189$

$\sum_{i=1}^{n} Y_i = 1.400 + 2.000 + \cdots + 1.250 + 2.200 = 11.200$

$\sum_{i=1}^{n} X_i Y_i = (20 \cdot 1.400) + (22 \cdot 2.000) + \cdots + (33 \cdot 2.200) = 19.255.000$

$\sum_{i=1}^{n} X_i^2 = 20^2 + 22^2 + \cdots + 33^2 = 5.251$

$\sum_{i=1}^{n} x_i^2 = 5.251 - \frac{35.721}{7} = 148$

$\sum_{i=1}^{n} y_i^2 = 19.255.000 - \frac{125.440.000}{7} = 1.335.000$

$\sum_{i=1}^{n} x_i y_i = 302.450 - \frac{2.116.800}{7} = 50$

Substituindo os valores na equação de correlação temos:

$$r = \frac{50}{\sqrt{148 \cdot 1.335.000}} = 0,004$$

Podemos concluir através do valor do coeficiente r que a correlação neste caso é aproximadamente zero, podendo ser classificada como nula.

Agora iremos calcular os parâmetros do modelo de RLS:

$$b = \frac{(7)(19.255.000) - (189)(11.200)}{(7)(5.251) - (189)^2} = 128.058,00$$

$$a = \frac{11.200}{7} - b\frac{189}{7} = -3.455,97$$

Finalmente chegamos no modelo de RLS $\hat{Y} = -3.455,97 + 128.058,00X_i$, e

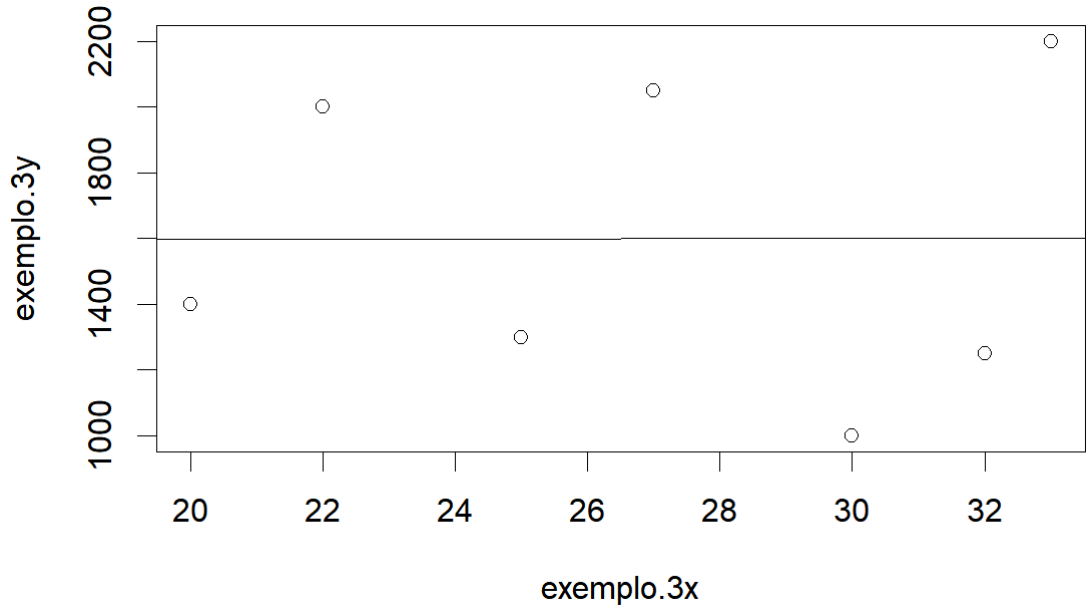


FIGURA 4.5: Representação gráfica do Exemplo 3. Fonte: O trabalho.

pela Figura (4.5) podemos notar que não foi um bom ajuste pois não há correlação entre as variáveis analisadas.

4.2 Regressão Linear Múltipla

Inicialmente vamos escrever o modelo em $\mathbb{R}^3$, tomando a equação do plano que representa o menor modelo de RLM com duas variáveis independentes: $ax + by + cz + d = 0$. Vamos reescrevê-la de forma que o Eixo Z seja uma relação linear dos demais tomando $c = -1$, assim:

$$z = ax + by + d \tag{4.16}$$

onde;

- z é a variável dependente ou resposta referente ao Eixo Z, variando de acordo com x e y;
- x e y são os valores das variáveis independentes ou preditoras, referentes aos Eixos X e Y;
- a, b e d são parâmetros do modelo a serem definidos.

Este modelo pode ser representado graficamente por um hiperplano, ou seja, um plano de equação linear em $\mathbb{R}^3$, conforme a Figura (4.6).

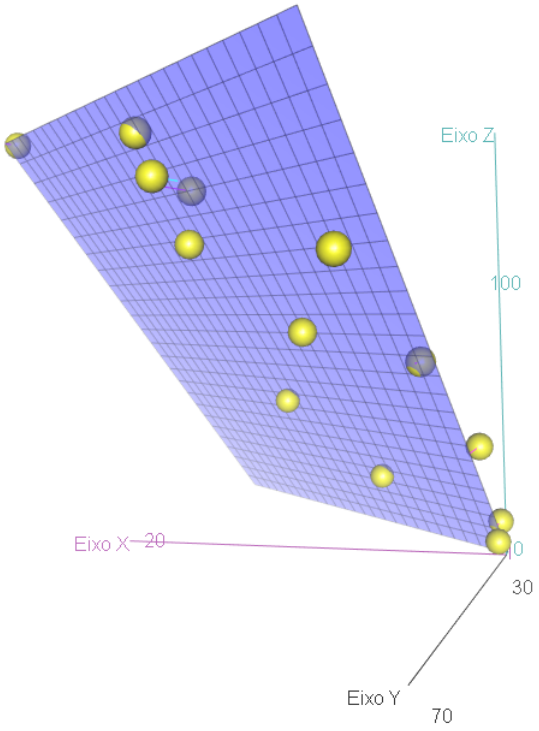


FIGURA 4.6: Hiperplano representando o modelo de RLM. Fonte: O trabalho.

Por outro lado, quando temos disponíveis k variáveis independentes, o modelo para a RLM pode ser escrito como:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + u_j, \quad j = 1, \ldots, n \tag{4.17}$$

ou

$$Y_j = \alpha + \sum_{i=1}^{k} \beta_i X_{ij} + u_j$$

onde,

- Y_j é a variável dependente ou resposta;
- $X_{1j}, X_{2j}, \ldots, X_{kj}$ são os valores das k variáveis independentes ou preditoras;
- $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ são os parâmetros do modelo;
- u_j representa os erros associados à resposta Y_j.

Com essa notação, a Equação (4.16) pode ser reescrita como $Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + \alpha + u_j$. Outra forma de representar o modelo de RLM é através de matrizes, a partir daqui iremos padronizar duas notações: vetores serão representados por letras em negrito (ex: $\mathbf{y} = (Y_1, \ldots, Y_n)$), e matrizes transpostas serão representadas

pela adição de aspas simples ao símbolo do vetor (ex: $\mathbf{X}'$ como transposta de $\mathbf{X}$). Sendo assim, a forma simplificada na notação matricial é dada por:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{4.18}$$

onde podemos expandir,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

com,

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Dessa forma temos que:

- $\mathbf{u}$ é um vetor de dimensão n x 1, onde os componentes são os erros $u_1, u_2, \dots, u_n$;
- $\mathbf{y}$ é um vetor n x 1, com componentes $Y_1, Y_2, \dots, Y_n$;
- $\mathbf{X}$ é uma matriz n x $(k+1)$, também conhecida como matriz do modelo;
- β é um vetor $(k+1)$ x 1 com os parâmetros do modelo (α, β_1, β_2, …, β_k).

Para ilustrar melhor a notação, iremos apresentar um exemplo com dados, resumidos, referentes a quantidade de calor liberada por grama de cimento e dois componentes do mesmo (HELENA, 2019), as variáveis da equação são:

$$\mathbf{y} = \begin{pmatrix} 78,5 \\ 74,3 \\ 104,3 \\ \vdots \\ 113,3 \\ 109,4 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 7 & 26 \\ 1 & 1 & 29 \\ 1 & 11 & 56 \\ & \vdots & \\ 1 & 11 & 66 \\ 1 & 10 & 68 \end{pmatrix} \quad \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{12} \\ u_{13} \end{pmatrix}$$

Escrevendo como equação temos:

$$\begin{pmatrix} 78,5 \\ 74,3 \\ 104,3 \\ \vdots \\ 113,3 \\ 109,4 \end{pmatrix} = \begin{pmatrix} 1 & 7 & 26 \\ 1 & 1 & 29 \\ 1 & 11 & 56 \\ & \vdots & \\ 1 & 11 & 66 \\ 1 & 10 & 68 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{12} \\ u_{13} \end{pmatrix}$$

4.2.1 Mínimos quadrados para RLM

Para a estimativa dos parâmetros será utilizada a forma matricial. Seja **b** o vetor da estimativa dos parâmetros e **e** o vetor dos desvios (erros):

$$\mathbf{b} = \begin{pmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} \quad \text{e} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Temos

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e} \tag{4.19}$$

e

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb} = \mathbf{y} - \hat{\mathbf{y}} \tag{4.20}$$

onde

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix}$$

Disso, tomemos a soma do quadrado dos desvios de forma a encontrar os valores de $\mathbf{b}$ que minimizam Z:

$$Z = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \tag{4.21}$$

$$Z = (\mathbf{y}' - \mathbf{b}'\mathbf{X}')(\mathbf{y} - \mathbf{Xb}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Xb} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}$$

Como $\mathbf{y}'\mathbf{Xb} = \mathbf{b}'\mathbf{X}'\mathbf{y}$, pois este produto resulta em um escalar e uma é a transposta da outra, chegamos em

$$Z = \mathbf{y}'\mathbf{y} - \mathbf{2b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \tag{4.22}$$

Tomando a diferencial identicamente nula da função Z para os valores de $\mathbf{b}$, encontramos o ponto de mínimo, logo

$$dZ = -(2d\mathbf{b}')\mathbf{X}'\mathbf{y} + (d\mathbf{b}')\mathbf{X}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{X}(d\mathbf{b}) \equiv 0 \tag{4.23}$$

Como $(d\mathbf{b}')\mathbf{X}'\mathbf{Xb} = \mathbf{b}'\mathbf{X}'\mathbf{X}(d\mathbf{b})$, uma vez que resultam em escalares e uma ser transposta da outra e simplificando a igualdade temos

$$\begin{gathered} -(2d\mathbf{b}')\mathbf{X}'\mathbf{y} + (2\mathbf{b}')\mathbf{X}'\mathbf{Xb} \equiv 0 \\ (d\mathbf{b}')(\mathbf{X}'\mathbf{Xb} - \mathbf{X}'\mathbf{y}) \equiv 0 \\ \mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y} \\ \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{gathered} \tag{4.24}$$

Com isso, ao encontrar o valor de $\mathbf{b}$ chegamos ao modelo final de regressão múltipla: $y = X\beta + u$, onde $(\mathbf{X}'\mathbf{X})^{-1}$ representa a inversa de $\mathbf{X}'\mathbf{X}$. Para realização dos

cálculos de estimativa, as matrizes da Equação 4.24 podem ser descritas como:

$$\mathbf{X'X} = \begin{pmatrix} n & \sum X_{1j} & \sum X_{2j} & \dots & \sum X_{kj} \\ \sum X_{1j} & \sum X_{1j}^2 & \sum X_{1j}X_{2j} & \dots & \sum X_{1j}X_{kj} \\ \sum X_{2j} & \sum X_{1j}X_{2j} & \sum X_{2j}^2 & \dots & \sum X_{2j}X_{kj} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum X_{kj} & \sum X_{1j}X_{kj} & \sum X_{2j}X_{kj} & \dots & \sum X_{kj}^2 \end{pmatrix}$$

$$\mathbf{X'y} = \begin{pmatrix} \sum Y_j \\ \sum X_{1j}Y_j \\ \sum X_{2j}Y_j \\ \vdots \\ \sum X_{kj}Y_j \end{pmatrix}$$

Essas matrizes são base da regressão linear múltipla.

4.2.2 Analisando a efetividade do modelo

Após calcular os parâmetros do modelo é necessário entender a efetividade do modelo encontrado e se será necessário realizar ajustes. No trabalho de Huang e Fang (2010) sobre predição da performance acadêmica, foram feitos e testados quatro modelos para identificação do melhor. Uma alternativa a este método é criar um único modelo com todas as variáveis disponíveis e testar a efetividade do modelo e das variáveis, retirando as que não contribuem.

São duas medidas principais, o teste F de significância e o coeficiente de determinação R^2. Vamos falar brevemente dos dois para aplicação no modelo e, para isso será necessário iniciar falando sobre Soma dos Quadrados Totais (SQT).

A SQT pode ser decomposta em outras duas somas, assim temos:

$$SQT = SQR + SQE$$

$$\sum_{i=1}^{n}(Y_j - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y_j} - \overline{Y})^2 + \sum_{i=1}^{n}(Y_j - \hat{Y_j})^2$$

Onde: SQR é a Soma dos Quadrados da Regressão e SQE é a Soma dos Quadrado dos Resíduos. Estabelecido tais notações, podemos testar a seguinte hipótese para RLM

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \exists : \beta_j \neq 0, \quad j = 1, \ldots, k \end{cases}$$

Tal hipótese testa se o modelo tem parâmetros significativos para o modelo ou não. Para a realização do teste é utilizado o quociente

$$F = \frac{\frac{SQR}{k}}{\frac{SQE}{n-k-1}} = \frac{QMR}{QME} \tag{4.25}$$

Sob a hipótese inicial H_0 temos que $\frac{SQR}{\sigma^2} \sim \chi^2_k$, $\frac{SQE}{\sigma^2} \sim \chi^2_{n-k-1}$ e como estas medidas são independentes e correspondem à razão anterior, concluímos que a estatística de teste segue uma distribuição F com k e $n-k-1$ graus de liberdade, ou seja,

$$F \sim F_{k,\ n-k-1}$$

Portanto, o teste F de significância, sob H_0, irá seguir uma distribuição $F_{k,\ n-k-1}$. Escolhendo-se o nível de significância θ, podemos rejeitar H_0 se $F > F_{(1-\theta:\ k,\ n-k-1)}$ e concluir que pelo menos uma das variáveis preditoras contribui para o modelo.

Após essa construção anterior, podemos sintetizar os dados observados na Tabela (4.5) conhecida como ANOVA (do inglês *analysis of variance*).

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

TABELA 4.5: Tabela da análise de variância - ANOVA (RODRIGUES, 2012).

Agora abordaremos o coeficiente de determinação, o qual pode ser sintetizado como "a fração da variância total em Y explicada pelo modelo de regressão com a da variância atribuída aos resíduos" (FARIA, 2011, p.14). Também podemos descrever esse valor como sendo "igual ao quadrado do coeficiente de correlação de Pearson" (RODRIGUES, 2012, p.31), e escrever da seguinte forma:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \tag{4.26}$$

$$0 < R^2 < 1$$

Com o coeficiente de determinação podemos quantificar a capacidade explicativa do modelo, ou seja, o quanto a variável resposta pode ser explicada pelas variáveis preditoras.

Contudo, um problema com essa medida está relacionado ao número de variáveis, de forma que quanto mais variáveis forem adicionadas (mesmo que não acrescentem ao resultado final) o coeficiente retorna valores mais altos. Para contornar este problema é preferível a utilização do coeficiente de determinação ajustado, o qual é uma medida que "penaliza" o número excessivo de variáveis do modelo de RLM.

O R^2 ajustado pode definido como:

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

Com esse ajuste, quando aumentamos o número de variáveis (k) o valor do coeficiente tende a diminuir, fazendo com que seja possível identificar o "ponto ótimo" do modelo, ou seja, o ponto que a adição de variáveis prejudica o ajuste.

Após analisar a capacidade preditiva e a existência de variáveis explicativas, será necessário realizar a análise de resíduos (diferença entre os valores projetados pelo modelo e os valores reais), que nomeamos também como erros anteriormente. A análise se dá pela validação de alguns dos pressupostos da RLS, e para RLM temos a adição de outro fator a se verificar, eles são: normalidade dos erros, independência dos erros, variância constante, a não existência de outliers, e exclusivo da RLM, verificar se não existe colinearidade ou multicolinearidade entre as variáveis.

Rodrigues (2012) aponta algumas formas de analisar e validar os pressupostos, sendo elas:

***Normalidade dos erros*:** para verificar a normalidade dos erros podem se usados diferentes gráficos, foram escolhidos dois: normal Q-Q, ou seja, distribuição dos resíduos quantil-quantil, e histograma dos resíduos estandardizados. Este último deve-se estar atento a quantidade de dados, sendo ele efetivo apenas para grandes quantidades.

***Independência dos erros*:** para identificar se os erros são independentes iremos utilizar o teste Durbin-Watson (DW), o qual se trata de um teste de hipótese: H_0: não existe correlação e H_1: existe correlação. A independência dos erros é aceita conforme aceita-se a hipótese H_0.

***Variância constante*:** nesta análise é comumente utilizado o gráfico de resíduos versus valores ajustados, onde a leitura do gráfico de dispersão permite a identificação de bom ajuste quando os pontos não apresentam nenhum padrão, bem como ajustes com valores extremos ou algum comportamento visível.

***Existência de Outliers*:** este tópico trata-se tanto de identificar outliers quanto entender se são valores influentes. Caso sejam valores influentes, ou seja, sua presença ou ausência muda consideravelmente os coeficientes do modelo, eles são preservados e o modelo pode ser definido. Caso não sejam influentes, podem ser removidos. O procedimento de identificação também pode ser realizado com gráficos, como o gráfico de valores estandardizados do modelo.

***Existência de colinearidade ou multicolinearidade*:** exclusivo da RLM, colinearidade ou multicolinearidade se refere ao fato das variáveis independentes estarem fortemente correlacionadas entre si, sendo o primeiro termo para duas variáveis e o segundo para mais do que duas variáveis. Para verificar a colinearidade pode ser utilizada o indicador VIF (Variance Inflation Factor), sendo que um $VIF > 5$ mostra que há problemas com colinearidade.

Finalizada a análise de resíduos, ou seja, validando os pressupostos o modelo estará completo. Contudo, há algumas formas de encontrar as melhores variáveis do modelo, especialmente quando se tem muitas, evitando assim problemas com multicolinearidade. A seguir vamos abordar os três principais métodos de seleção de variáveis: *Forward*, *Backward* e *Stepwise* (RODRIGUES, 2012 e MASIERO, 2011).

***Forward*:** método em que são adicionadas as variáveis uma por uma, começando apenas com a constante. Para seleção de qual variável incluir em cada passo pode ser utilizado o teste F ou o coeficiente de correlação, selecionado a que tem maior significância ou correlação. O movimento é finalizado quando não há mais variáveis significativas para adicionar.

***Backward*:** neste método a lógica é similar ao anterior, porém ao contrário de adicionar é retirado variáveis, ou seja, começa com todas as variáveis e retira-se as que têm menor significância para o modelo, caso haja alguma.

***Stepwise*:** este método pode ser descrito como uma combinação dos anteriores, começando com a mais significativa e, ao passo que é adicionada uma variável também pode ser retirada outra, sempre deixando apenas as mais significativas para o modelo.

Com o objetivo de mostrar como ocorrem os cálculos envolvidos em uma RLM, tomemos uma amostra (Tabela 4.6) com 5 observações (Hoffmann, 2016).

Y	X_1	X_2
16,5	1,0	2
14,0	3,5	3
6,0	4,0	4
10,0	7,5	5
3,5	9,0	6

TABELA 4.6: Amostra de 5 observações (HOFFMANN, 2016).

Vamos agora escrever a matriz do modelo, $\mathbf{X}$, e a matriz das variáveis respostas:

$$\mathbf{X} = \begin{pmatrix} 1,0 & 1,0 & 2 \\ 1,0 & 3,5 & 3 \\ 1,0 & 4,0 & 4 \\ 1,0 & 7,5 & 5 \\ 1,0 & 9,0 & 6 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 16,5 \\ 14,0 \\ 6,0 \\ 10,0 \\ 3,5 \end{pmatrix}$$

Como próximo passo, devemos encontrar as matrizes para cálculo de $\mathbf{b}$, as quais são: $\mathbf{X'X}$ e $\mathbf{X'y}$. Começaremos por $\mathbf{X'X}$:

$$\mathbf{X'X} = \begin{pmatrix} 1,0 & 1,0 & 1,0 & 1,0 & 1,0 \\ 1,0 & 3,5 & 4,0 & 7,5 & 9,0 \\ 2 & 3 & 4 & 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1,0 & 1,0 & 2 \\ 1,0 & 3,5 & 3 \\ 1,0 & 4,0 & 4 \\ 1,0 & 7,5 & 5 \\ 1,0 & 9,0 & 6 \end{pmatrix}$$

$$\mathbf{X'X} = \begin{pmatrix} 5,0 & 25,0 & 20,0 \\ 25,0 & 166,5 & 120,0 \\ 20,0 & 120,0 & 90,0 \end{pmatrix}$$

Agora vamos calcular $\mathbf{X'y}$:

$$\mathbf{X'y} = \begin{pmatrix} 1,0 & 1,0 & 1,0 & 1,0 & 1,0 \\ 1,0 & 3,5 & 4,0 & 7,5 & 9,0 \\ 2 & 3 & 4 & 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 16,5 \\ 14,0 \\ 6,0 \\ 10,0 \\ 3,5 \end{pmatrix}$$

$$\mathbf{X'y} = \begin{pmatrix} 50,0 \\ 196,0 \\ 170,0 \end{pmatrix}$$

Como sabemos, $\mathbf{b} = (\mathbf{X'X})^{-1}\,\mathbf{X'y}$, então basta calcularmos os parâmetros do modelo:

$$\mathbf{b} = \begin{pmatrix} 585,0 & -330,0 & 150,0 \\ -330,0 & 207,5 & -33,25 \\ 150,0 & -33,25 & 50,0 \end{pmatrix} \cdot \begin{pmatrix} 50,0 \\ 196,0 \\ 170,0 \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} 0,37973 \\ -0,27419 \\ 0,13767 \end{pmatrix}$$

Portanto, o nosso modelo para os dados disponibilizados será:

$$\hat{Y} = 0,37973 - 0,27419 \cdot X_1 + 0,13767 \cdot X_2$$

Para analisar o modelo vamos utilizar o teste F de significância e o coeficiente de determinação R^2 ajustado. Primeiro passo será construir a ANOVA:

Causas de Variação	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	F
Regressão	$114,0$	2	$QMR = 57,0$	$F = 45,6$
Erro (resíduo)	$2,5$	2	$QME = 1,25$	
Total	$116,5$	4	—	

TABELA 4.7: Tabela da análise de variância - ANOVA. Fonte: O trabalho.

Com ajuda do software R podemos ver que o valor associado à $F = 45,6$ com 2 e 2 graus de liberdade é 0,02146, ou seja, ao nível de significância de 5% o modelo é significativo.

Para calcular o valor de R^2 aplicamos a fórmula, dividindo SQR por SQT e encontramos $R^2 = 0,9785$. Porém o que nos interessa é o R_a^2:

$$R_a^2 = 1 - \left(\frac{4}{2} \right) (1 - 0,9785) = 0,9571$$

Com isso, temos que o modelo explica 95,7% do comportamento dos dados.

Capítulo 5

Resultados

Os microdados do ENEM contemplam uma grande quantidade de dados sobre os candidatos e resultados de provas. Os dados vão desde a sua localização e ambiente escolar até dados socioeconômicos. Para este TCC, adotamos os dados da Edição de 2012[1] que somam 3.8GB de dados de milhões de inscritos. Como o volume de dados é muito grande, coletamos uma amostra que corresponde aos candidatos da região de Ribeirão Preto, uma vez que o autor cresceu e realizou o ENEM na regional.

Segundo dados disponíveis no site da Empresa Paulista de Planejamento Metropolitano (EMPLASA), a região de Ribeirão preto é composta por 34 municípios, os quais são divididos em 4 sub-regiões, e conta com uma população maior que 1,7 milhão de habitantes segundo o Instituto Brasileiro de Geografia e Estatística (IBGE) em 2018. Em termos dos dados do ENEM 2012, são 26.731 candidatos, que é a amostra utilizada neste trabalho.

Sendo definida a região de trabalho, o próximo passo foi definir os dados a serem explorados: Idade, Sexo, Raça, Tipo conclusão, Ano conclusão, Tipo escola, Score da prova de matemática, Score da prova de ciências naturais, Score da prova de ciências humanas e Score da prova de linguagens e códigos.

Os dados citados acima são disponibilizados numericamente ou como caracteres, onde os números podem representar as opções, como Sexo sendo 0 - Masculino e 1 - Feminino além do símbolo de ponto (.) representando a ausência de informação. Para exemplificar a disposição dos dados, a seguir segue um pequeno exemplo:

Código município	Idade	...	Tipo conclusão	Ano conclusão	Tipo escola	score mt	...	score lc
3501004	38	...	2	2012	1	391.3	...	409.8

[1]Os dados foram minerados e disponibilizados pelo orientador, a partir de dados da pesquisa de Piton-Gonçalves, Souza, Lamonato (2017).

3501004	21	...	2	2012	1	551.1	...	461.7
3501004	18	...	2	2012	1	637.9	...	585.2
3501004	18	...	2	2012	1	463.6	...	485.1
3501004	18	...	2	2012	1	353.9	...	391.2
3501004	18	...	2	2012	1	530	...	394.9
3501004	18	...	2	2012	1	429.9	...	453.1
⋮	⋮		⋮	⋮	⋮	⋮		⋮

TABELA 5.1: Exemplo de dados para análise (INEP).

Com os dados em mãos deu-se início ao processo de análise. O primeiro passo foi definir quais modelos seriam testados. Como já visto no Capítulo 4 Seção 2, podemos apenas incluir todos os dados e pedir que o programa nos sinalize quais contribuem ou não para o modelo, porém, além disso seguimos por outro lado forçando a exploração e investigação, e assim definimos alguns modelos para teste.

Antes de apresentar os modelos, nos atentemos a nomenclatura que será utilizada: Idade (**idade**), Sexo (**sexo**), Raça (**raca**), Tipo conclusão (**tp_conclusao**), Ano conclusão (**ano**), Tipo escola (**tp_escola**), Score da prova de matemática (**mt**), Score da prova de ciências naturais (**cn**), Score da prova de ciências humanas (**ch**) e Score da prova de linguagens e códigos (**lc**).

Desta forma, os modelos definidos, no formato $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + u_j$, para serem testados e analisados são:

Teste 1) $\text{mt} = \alpha + \beta_1 \cdot \text{cn} + \beta_2 \cdot \text{ch} + \beta_3 \cdot \text{lc}$

Teste 2) $\text{mt} = \alpha + \beta_1 \cdot \text{idade} + \beta_2 \cdot \text{ano}$

Teste 3) $\text{mt} = \alpha + \beta_1 \cdot \text{idade} + \beta_2 \cdot \text{tp\_escola}$

Teste 4) $\text{mt} = \alpha + \beta_1 \cdot \text{idade} + \beta_2 \cdot \text{raca} + \beta_3 \cdot \text{sexo} + \beta_4 \cdot \text{tp\_escola}$

Teste 5) $\text{mt} = \alpha + \beta_1 \cdot \text{cn} + \beta_2 \cdot \text{ch} + \beta_3 \cdot \text{lc} + \beta_4 \cdot \text{tp\_escola}$

O teste consiste em encontrar os parâmetros do modelo e analisar se estes são significativos à confiança de 5% e se fornecem bom poder preditivo. Após estes primeiros passos confirmados, haveria a sequência pela análise de resíduos.

5.1 Gerando os modelos de RLM

Definidos os modelos a serem testados, o processo segue com o tratamento dos dados. Vale ressaltar que eles já foram disponibilizados previamente filtrados, minimizando o trabalho, o qual se reduziu a identificação de ausência de dados e a separação de acordo com o modelo testado. Por exemplo: para o Teste 1 pode-se utilizar 100% dos dados, porém o Teste 3 só é possível com 37,2% dos dados.

Após realizar as separações, obtivemos as seguintes percentagens do volume total de dados para os testes: Teste 1 - 100,0%, Teste 2 - 79,6%, Teste 3, 4 e 5 - 37,2% (pois todos são limitados pelo dado *tipo de escola*). O dado que explica o tipo de escola (pública, particular) do participante só foi coletado de quem estava concluindo o Ensino Médio em 2012, logo os dados da grande maioria dos participantes não entra nas análises que envolvem este dado. Para realizar os experimentos foi utilizado o software estatístico R.

5.1.1 Modelo para o Teste 1

Para o Teste 1, como já mencionado, foram utilizados todos os dados. O programa nos retorna diversas informações: os coeficientes de regressão estimados, o p-valor do Teste F de significância das variáveis, p-valor do modelo e o valor R^2 ajustado que explica o poder preditivo do modelo. Na Tabela 5.2 temos os coeficientes de regressão com valores:

Variável	Coeficiente	Valor	p-valor Teste F
Independente	α	-82,738272	2^{-16}
cn	β_1	0,706947	2^{-16}
ch	β_2	0,258790	2^{-16}
lc	β_3	0,281395	2^{-16}

TABELA 5.2: Parâmetros para Teste 1. Fonte: O trabalho.

Além destes valores individuais, o modelo é significativo pelo Teste F e tem poder de predição de 59,2%. Portanto, com este bom valor de predição do score em Matemática através dos demais scores e sendo significativo, o modelo se mostra adequado.

5.1.2 Modelo para o Teste 2

Neste teste iremos utilizar 79,6% dos dados que foram disponibilizados devido ao dado ano de conclusão (**ano**) não ter sido preenchido por todos. Após aplicar

a função de regressão múltipla temos:

Variável	Coeficiente	Valor	p-valor Teste F
Independente	α	12.545,32565	2^{-16}
idade	β_1	-4,6295	2^{-16}
ano	β_2	-5,9184	2^{-16}

TABELA 5.3: Parâmetros para Teste 2. Fonte: O trabalho.

Como podemos ver, as variáveis são significativas e o modelo também pelo teste de significância, porém ele apresenta apenas 2,3% de poder preditivo. Concluímos que apenas idade e ano de conclusão do Ensino Médio não são suficientes para gerar um modelo adequado.

5.1.3 Modelo para o Teste 3

Para o Teste 3, e consequentemente 4 e 5, teremos apenas 37,2% dos dados por conta da informação Tipo de Escola ser adicionada apenas pelos candidatos concluintes do Ensino Médio no ano de 2012. Aplicando a função de regressão múltipla obtemos os dados a seguir:

Variável	Coeficiente	Valor	p-valor Teste F
Independente	α	435,2969	2^{-16}
idade	β_1	-3,4389	2^{-16}
tp escola	β_2	129,4223	2^{-16}

TABELA 5.4: Parâmetros para Teste 3. Fonte: O trabalho.

O modelo é significativo assim como as variáveis presentes na Tabela 5.4. Por outro lado, pelo R^2 ajustado o poder preditivo é apenas 25,6%, o qual é relativamente mais significativo que o Teste 2, porém ainda longe de um valor adequado.

Além destas informações, neste teste surge pela primeira vez o dado de Tipo de escola. Este dado tem valores 1 para escola pública e 2 para escola particular, e nota-se que sua relevância na computação final do score da prova de Matemática é de 129,42 pontos para este modelo, ou seja, o participante de escola particular teria uma vantagem de 129,42 pontos contra um aluno de escola pública. Este é um indício inicial que iremos investigar mais a frente, inclusive nos próximos modelos.

5.1.4 Modelo para o Teste 4

O modelo para o Teste 4 é focado em dados sociais, o objetivo final é entender se há algum indício que gere outras discussões e investigações. Os resultados obtidos são:

Variável	Coeficiente	Valor	p-valor Teste F
Independente	α	485,2013	2^{-16}
idade	β_1	-3,4258	2^{-16}
raca	β_2	-8,8910	$9,82^{-14}$
sexo	β_3	-47,2156	2^{-16}
tp escola	β_4	122,2288	2^{-16}

TABELA 5.5: Parâmetros para Teste 4. Fonte: O trabalho.

Todos os valores são significativos, assim como o modelo. Porém não evoluímos muito no poder preditivo em relação ao anterior, sendo o deste modelo apenas 29,6%, ainda longe de um valor adequado.

Como pontos de destaque observados neste modelo temos: escola e sexo. O tipo de escola notamos antes e iremos entender melhor, porém além dele vemos uma diferença significativa entre o sexo masculino e feminino. Com este indício pode-se trazer outros estudos mais profundos e específicos que não adentaremos aqui, como o realizado por Teresa Vernaglia (2020) no campo de saneamento básico em que chegou a conclusão de que "meninas sem banheiro em casa tem notas 25% menores no Enem".

5.1.5 Modelo para o Teste 5

Adentrando ao modelo para o Teste 5 iremos abordar as notas com o tipo de escola, assim temos os seguintes resultados:

Variável	Coeficiente	Valor	p-valor Teste F
Independente	α	-89,09328	2^{-16}
cn	β_1	0,61788	2^{-16}
ch	β_2	0,27488	2^{-16}
lc	β_3	0,29014	2^{-16}
tp escola	β_4	28,47132	2^{-16}

TABELA 5.6: Parâmetros para Teste 5. Fonte: O trabalho.

Neste último teste temos também variáveis e modelo significativos ao nível

de 5%. Além disso, pelo R^2 ajustado o modelo prevê 62,1% do desempenho de um candidato na prova de Matemática e suas Tecnologias. Como estávamos também de olho na quesito tipo de escola, vemos que mesmo diminuindo o volume de dados o valor total ainda é significativo este dado para o score final.

Este modelo é similar ao primeiro, porém com a adição do tipo de escola limitamos os dados a 37,2% dos dados e mesmo assim alcançamos um poder preditivo maior que o Teste 1! Isso nos mostra que este volume de dados já é suficiente para gerar um bom modelo e que os diferenciais se tornam a inclusão de outras variáveis significativas. O tipo de escola, apesar de ser uma variável binária acrescenta ao modelo de forma positiva e o torna em nosso melhor modelo.

5.2 Analisando quanto ao tipo de escola

Buscando compreender a relação entre candidatos e tipo de escola cursada (pública ou particular), adotamos o gráfico de Draftman, que a partir de gráficos matricialmente distribuídos, busca-se interpretar a relação entre as variáveis em questão. Nesse caso, são os scores, a idade e o quesito particular versus pública.

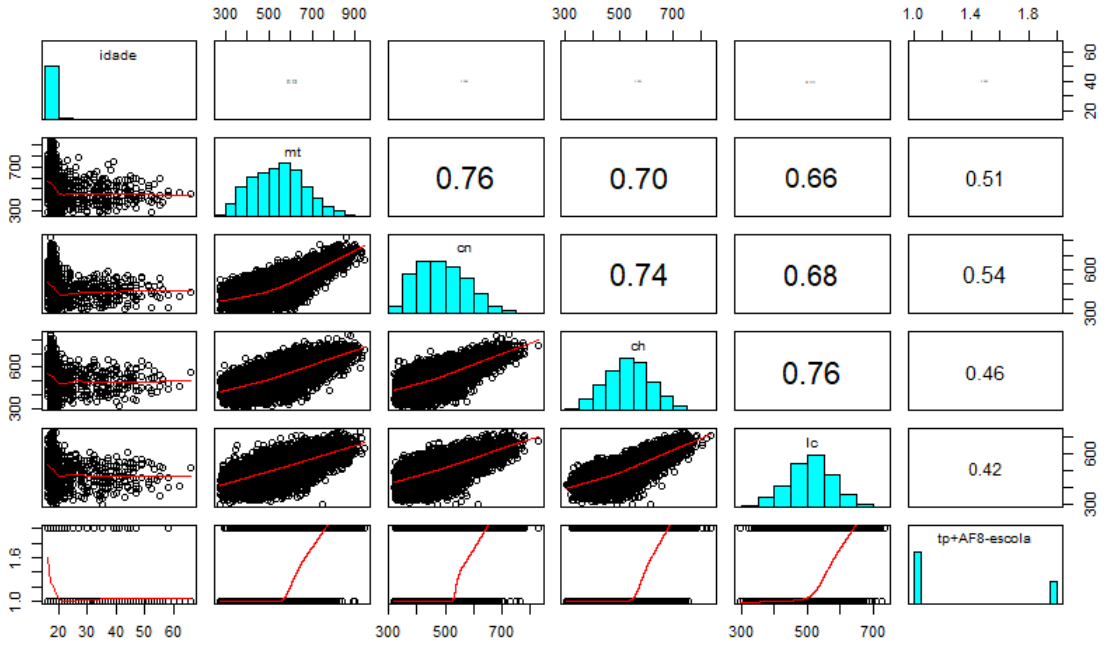


FIGURA 5.1: Gráfico de Draftman. Fonte: O trabalho.

Pela Figura 5.1 podemos notar na parte superior a diagonal os valores da correlação de Pearson, onde os valores são apresentados de forma a seu tamanho ser relativo a força da correlação. É possível notar que o valor de correlação entre matemática e língua portuguesa é o menor dentre os conteúdos, indicando que as notas

estão relacionadas com menor intensidade, ou seja, o resultado baixo em matemática não pode ser atribuído ao desempenho baixo de língua portuguesa assim como o cenário oposto com resultados altos. Há dentre os candidatos aqueles que tiveram maior nota em matemática e menor em língua portuguesa e vice versa.

Na parte inferior temos os gráficos de dispersão com uma linha de suavização em vermelho e, na diagonal, um histograma com a distribuição dos dados. Através destes dados notamos a correlação forte entre as áreas, sendo as correlações mais significativas entre Matemática / Ciências Naturais e Linguagens / Ciências Humanas.

Além deste fator, através dos dados contidos nos histogramas notamos diferenças significativas em relação a forma como são distribuídos os scores, uma vez que nas áreas exatas temos grande parte dos dados com valores inferiores à mediana e nas áreas humanas esse comportamento não acontece, sendo melhor distribuídos os scores.

Agora, repetiremos o processo para os dados separados entre o tipo de escola, o que irá permitir entender melhor os números obtidos nos modelos da seção anterior e ainda verificar as discrepâncias vistas nesta primeira etapa.

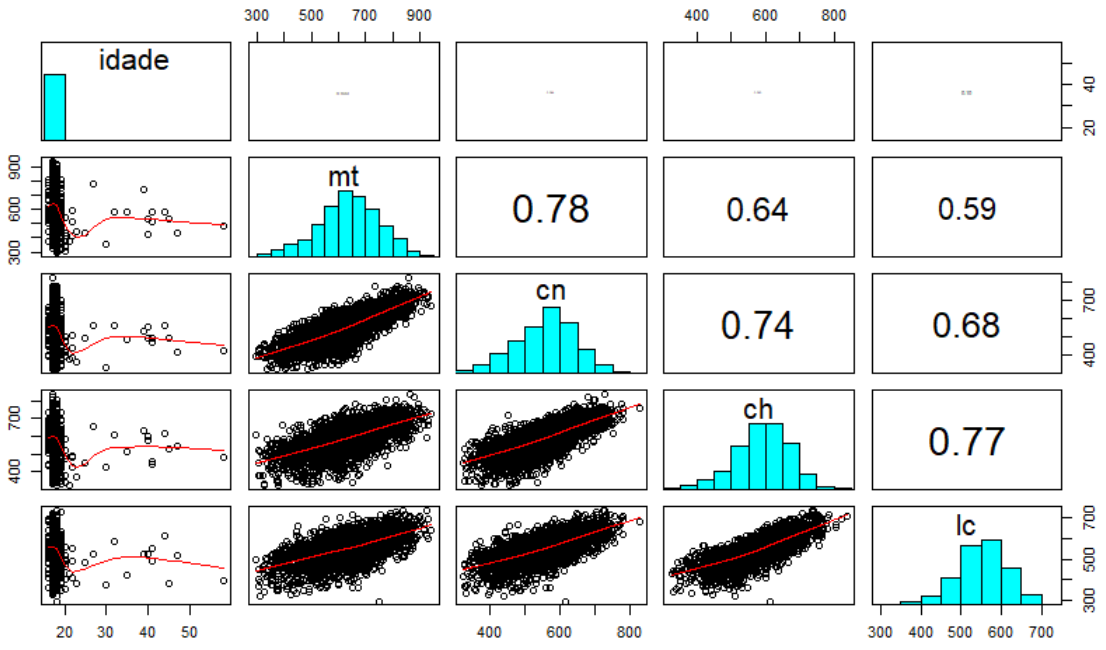


FIGURA 5.2: Gráfico de Draftman para escola particular. Fonte: O trabalho.

Através das Figuras 5.2 e 5.3 podemos notar que a diferença entre as correlações das áreas de ensino é menor para escola pública, o que leva a hipótese dos resultados não serem tão uniformes quanto a particular. Esse ponto é reforçado pelos gráficos de dispersão, os quais nos trazem outra informação quanto a escola pública: há uma faixa de idade considerável fugindo do padrão que gira em torno de 20 anos, de onde podemos inferir quanto ao tipo de ensino ser Educação de Jovens e Adultos (EJA). Ao

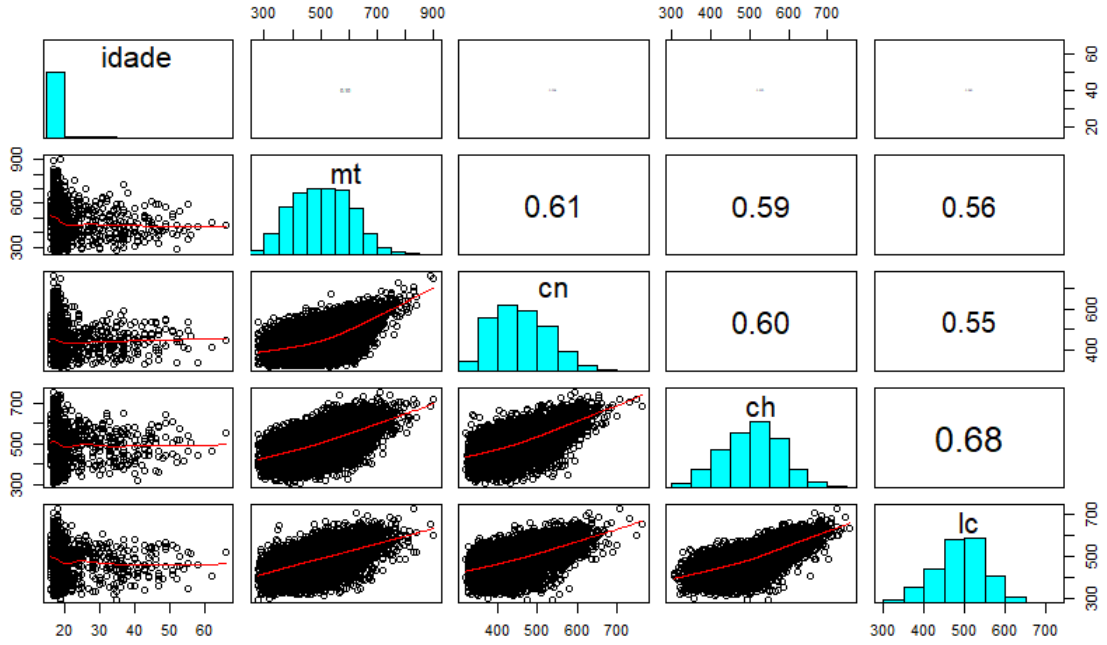


FIGURA 5.3: Gráfico de Draftman para escola pública. Fonte: O trabalho.

olhar mais afundo os dados, notamos que este total representa menos de 3% do volume total dos dados de escola pública, não interferindo significativamente para este tipo de visão.

Com os histogramas fica muito claro as diferenças apontadas pelas correlações. Especificamente em Matemática vemos que o comportamento visto anteriormente nos dados gerais é potencializado pelos dados de escola pública, o qual se assemelha muito àquela visão; para escola particular a distribuição em Matemática é mais uniforme e o maior volume se encontra após a mediana, então iremos sobrepor os gráficos por área do conhecimento para que a comparação fique clara e possamos entender melhor.

No primeiro gráfico (Figura 5.4) temos a distribuição de Linguagens e Códigos, a qual apresenta média para particular de 552,1 e 489,9 para pública, o que gera uma diferença no score de aproximadamente 11% a menos para pública. Olhando para a mediana os valores são bem próximos, sendo 554,5 e 493,8 para particular e pública respectivamente.

A distribuição de Ciências Humanas (Figura 5.5) apresenta médias de 590,6 e 506,9 para particular e pública respectivamente, mostrando uma diferença maior e ainda mais significativa de aproximadamente 14% a menos para pública. Além disso, a mediana neste caso é 595,4 e 509,5 para particular e pública, respectivamente.

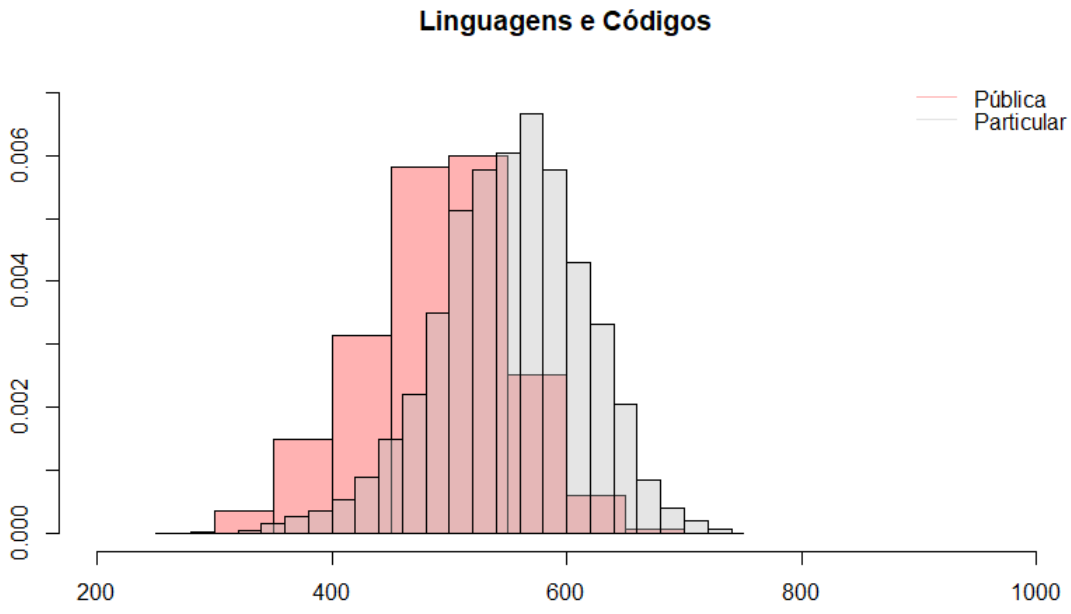


FIGURA 5.4: Histogramas sobrepostos: Linguagens e Códigos. Fonte: O trabalho.

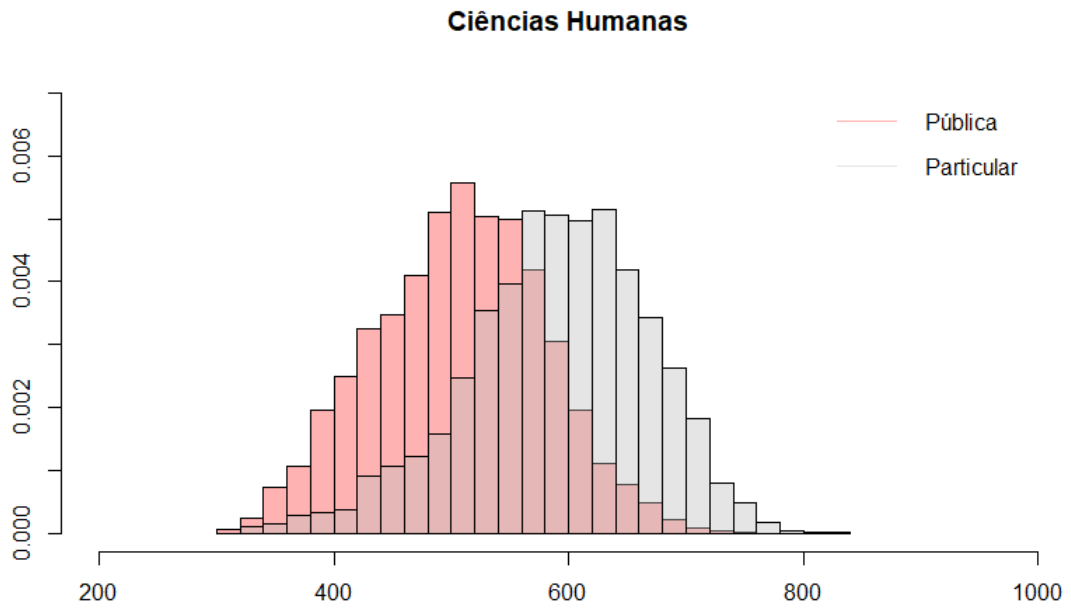


FIGURA 5.5: Histogramas sobrepostos: Ciências Humanas. Fonte: O trabalho.

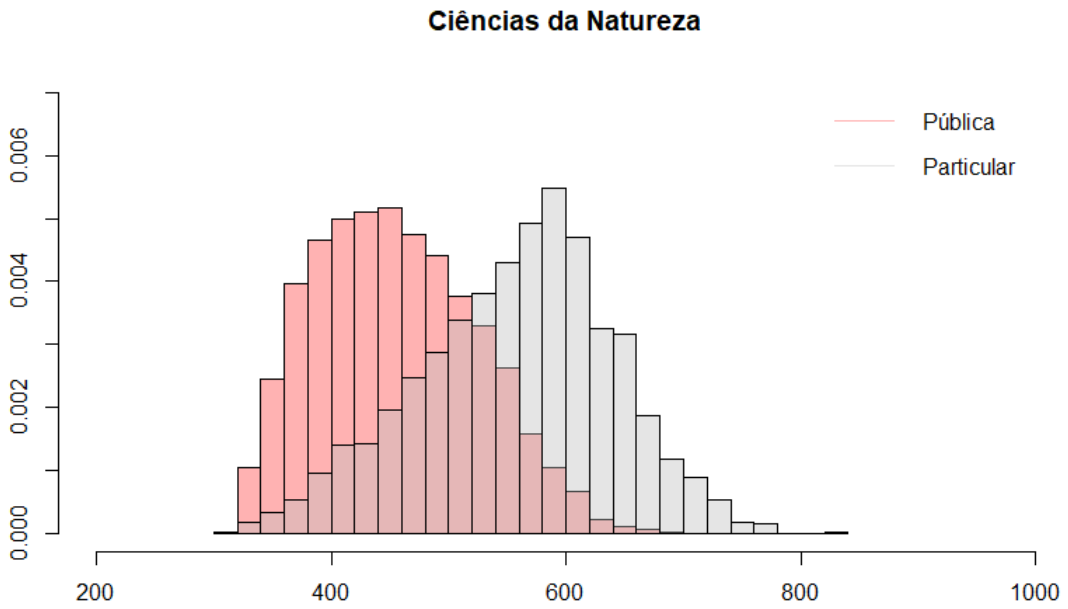


FIGURA 5.6: Histogramas sobrepostos: Ciências Naturais. Fonte: O trabalho.

Mudando de frente e indo para exatas, vamos olhar primeiro Ciências da Natureza (Figura 5.6) que tem médias de 557,6 e 455,8 para particular e pública respectivamente, gerando diferença de aproximadamente 18% a menos para pública. O ponto principal neste caso é que as medianas são 566,2 e 450,7 para particular e pública, porém aqui notamos algo ainda não visto: a mediana está abaixo da média para escola pública, ou seja, o maior número de candidatos tem notas inferiores a média. Além de não vermos isso para escola pública antes, para escola particular este fato não ocorre aqui também, sigamos para Matemática.

Em Matemática (Figura 5.7) temos médias 637,2 e 503,0 para particular e pública respectivamente, o que gera a diferença de aproximadamente 21% a menos para escola pública, a maior diferença entre áreas por tipo de escola. A mediana segue comportamento similar ao anterior, sendo 639,8 e 502,3 para particular e pública respectivamente, onde o maior número de candidatos tem score inferior a média.

A discrepância de score entre os tipos de escola cresceu na sequência de Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática; além do fato de que para as áreas exatas a diferença ainda está no fato de a maioria dos candidatos de escola pública ter score inferior a média, gerando um ponto de atenção quanto ao nosso Sistema de Ensino.

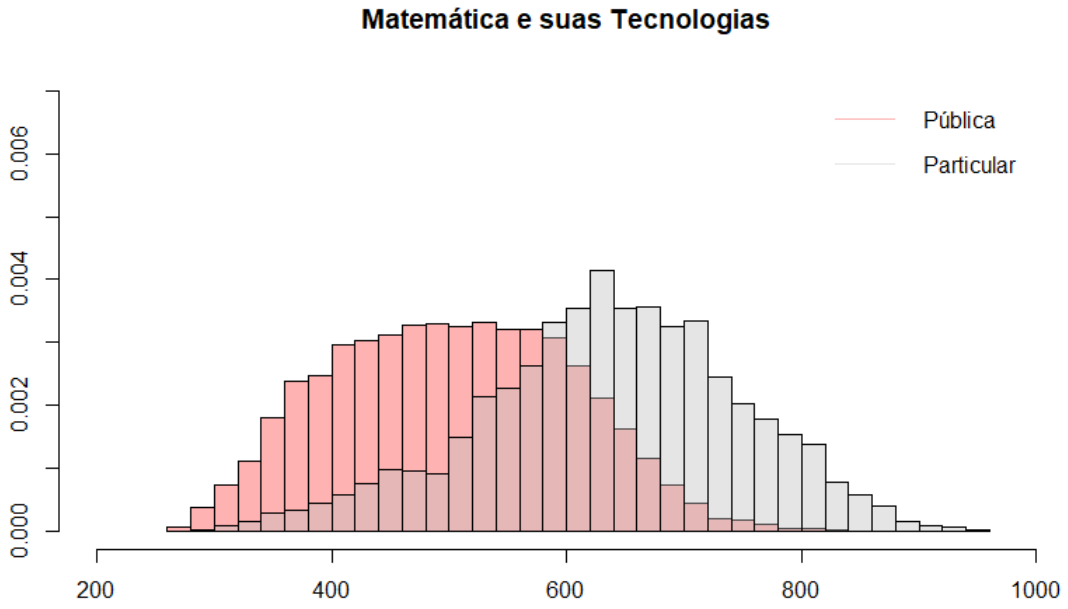


FIGURA 5.7: Histogramas sobrepostos: Matemática e suas Tecnologias. Fonte: O trabalho.

Capítulo 6

Conclusão

Enquanto resultados deste TCC, amostralmente inferimos que o tipo de escola é uma variável muito significativa quanto ao desempenho dos candidatos na realização do ENEM, sendo a vantagem para os oriundos de escolas particulares. A diferença é mais significativa em áreas de Ciências Exatas, como Matemática, Física e Química. Com esses resultados, a atenção deve ser direcionada para que se possa buscar mais informações acerca dessa diferença entre tipos de escola por áreas do conhecimento do ENEM, contribuindo conjuntamente com os trabalhos de Sampaio e Guimarães (2009) e Moraes e Beluzzo (2014).

Outro resultado é em relação a previsão de desempenho do estudante, realizada a partir de um modelo de regressão linear. As oportunidades aqui podem ser tanto a aplicação em dados de estudantes do Ensino Médio para prever seu desempenho nos vestibulares, quanto após a entrada na faculdade ou universidade, de forma a prever o desempenho do calouro ao longo do curso de graduação. Recentemente no Reino Unido foi desenvolvido um algoritmo para decidir as notas de estudantes para concorrerem às vagas das universidades, após o exame nacional ser cancelado devido à pandemia do Coronavírus[1], o que gerou revolta da comunidade e o governo precisou voltar atrás (BBC NEWS, 20 de outubro de 2020).

Enquanto contribuição para a minha formação este trabalho permitiu a descoberta e interação com um tema atual e de grande importância, o qual está além da grade curricular, é interdisciplinar e permite analisar e agir de forma construtiva na educação, tanto como professor quanto gestor.

Enquanto trabalho futuro, almeja-se analisar e modelar o desempenho de estudantes de outras regiões e correlacionar com variáveis socioeconômicas presentes nos dados do ENEM.

[1]Pandemia global causada pelo vírus SARS COV-2, nomeado também como COVID-19, com início na China em 2019 e surto internacional em Fevereiro de 2020.

Referências Bibliográficas

ABLEDU, Godfred Kwame. *Multiple Regression Analysis of Assessment of Academic Performance of Students in the Ghanaian Polytechnics*. Research on Humanities and Social Sciences, v. 2, n. 9, p. 15-25, 2012.

BASTOS, Edson Vinicius Pontes; GUIMARÃES, Julio Cesar Ferro; SEVERO, Eliana Andréa. *Modelo de regressão linear para análise de investimentos em uma empresa do ramo petrolífero*. Revista Produção e Desenvolvimento, v. 1, n. 1, p. 77-88, 2015.

BBC NEWS. *'Algoritmo roubou meu futuro': solução para 'Enem britânico' na pandemia causa escândalo*. 2020. Disponível em <https://noticias.uol.com.br/ultimas-noticias/bbc/2020/08/20/algoritmo-roubou-meu-futuro-solucao-para-enem-britanico-na-pandemia-causa-escandalo.htm?cmpid=copiaecola>. Acesso em 19 nov. 2020.

BRASIL. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Relatório pedagógico: Enem 2011-2012. 2015.

BRASIL. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Escalas de Proficiência 1998/2008. 2018.

CAMILO, Cássio Oliveira. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. 2009. 29 p. Universidade Federal de Goiás, Goiânia, 2009.

EMPLASA, Empresa Paulista de Planejamento Metropolitano. *Sobre Região Metropolitana de Ribeirão Preto - RMRP*. 2019. Disponível em <https://emplasa.sp.gov.br/RMRP>. Acesso em 19 nov. 2020.

FARIA, Bruno Fernando Pinheiro. *Teste F na regressão linear múltipla para dados temporais em correlação serial*. 2011. Tese de Doutorado.

GADHAVI, Mahesh; PATEL, Chirag. *Student final grade prediction based on linear regression*. Indian J. Comput. Sci. Eng., v. 8, n. 3, p. 274-279, 2017.

GOLDSCHMIDT, Ronaldo Ribeiro. *Uma Introdução à Inteligência Computacional: fundamentos, ferramentas e aplicações*. Rio de Janeiro Brasil: IST-Rio, v. 1, p. 32,

2010.

HAN, Jiawei; PEI, Jian. KAMBER, Micheline. *Data mining: concepts and techniques*. Elsevier, 2011.

HELENA, Maria. *Tutorial ? Ajuste e Interpretação de Regressão Linear com R*. 2019. Disponível em <https://medium.com/data-hackers/tutorial-ajuste-e-interpreta%C3%A7%C3%A3o-de-regress%C3%A3o-linear-com-r-5b23c4ddb72>. Acesso em 19 nov. 2020.

HOFFMANN, Rodolfo. *Análise de regressão: uma introdução à econometria*. Piracicaba: ESALQ/USP, 2015. 393 p.

HUANG, Shaobo; FANG, Ning. *Regression models of predicting student academic performance in an engineering dynamics course*. In: American Society for Engineering Education. American Society for Engineering Education, 2010.

IBRAHIM, Zaidah; RUSLI, Daliela. *Predicting students? academic performance: comparing artificial neural network, decision tree and linear regression*. In: 21st Annual SAS Malaysia Forum, 5th September. 2007.

IESB, Centro Universitário. *SAS ENTERPRISE MINER 1 - Introdução ao Ambiente da Mineração de Dados e do Machine Learning*. 2020. Disponível em <https://www.iesb.br/qualificacao/curso/sas-enterprise-miner-1—introducao-ao-ambiente-da-mineracao-de-dados-e-do-machine-learning>. Acesso em 19 nov. 2020.

INEP. *Histórico ENEM*. 2019. Disponível em <http://portal.inep.gov.br/enem/historico>. Acesso em 17 nov. 2020.

INEP. *Matriz de referência ENEM*. 2015. Disponível em <http://portal.inep.gov.br/matriz-de-referencia>. Acesso em 17 nov. 2020.

INEP. *Diário oficial da União*. 2009. Disponível em <http://download.inep.gov.br/educacao_basica/enem/legislacao/2009/portaria_enem2009_2.pdf>. Acesso em 17 nov. 2020.

INEP. *2º dia Caderno 5 - Amarelo*. 2012. Disponível em <http://download.inep.gov.br/educacao_basica/enem/provas/2012/caderno_enem2012_dom_amarelo.pdf>. Acesso em 17 nov. 2020.

LOH, Stanley. *BI na Era do Big Data para Cientistas de Dados*. Porto Alegre: Amazon, 2014.

MAROCO, João. *Análise Estatística com o SPSS Statistics.*: 7ª edição. ReportNumber, Lda, 2018.

MASIERO, Miguel Slomp. *Seleção de variáveis para predição utilizando regressão linear em processos logísticos de distribuição*. Universidade Federal do Rio Grande

do Sul, 2011.

MORAES, André Guerra Esteves de; BELLUZZO, Walter. *O diferencial de desempenho escolar entre escolas públicas e privadas no Brasil*. Nova economia, v. 24, n. 2, p. 409-430, 2014.

MORETTIN, Pedro Alberto; BUSSAB, Wilton Oliveira. *Estatística básica*. Saraiva Educação SA, 2017.

PITON-GONÇALVES, Jean; SOUZA, Erica Rachel; LAMONATO, Maiza. *Logaritmos, trigonometria, matrizes e determinantes: uma análise de itens do enem do ponto de vista curricular*. EMR-RS - ANO 18 - 2017 - número 18 - v.2 - pp. 69 a 86.

PITON-GONCALVES, Jean. *Testes adaptativos para o Enade: uma aplicação metodológica*. Revista Meta: Avaliação, v. 12, n. 36, p. 665-688, 2020.

PORTO, Fábio; ZIVIANI, Arthur. *Ciencia de Dados*. III Seminário de Grandes Desafios da Computação no Brasil, Rio de Janeiro, RJ, 2014.

RAJALAXMI, R. R. *et al. Regression Model for Predicting Engineering Students Academic Performance*. International Journal of Recent Technology and Engineering, p. 71-75. 2019.

RODRIGUES, Sandra Cristina Antunes. *Modelo de regressão linear e suas aplicações*. 2012. Tese de Doutorado. Universidade da Beira Interior.

SAMPAIO, Breno; GUIMARÃES, Juliana. *Diferenças de eficiência entre ensino público e privado no Brasil*. Economia Aplicada, v. 13, n. 1, p. 45-68, 2009.

SANTOS, Jucenilton Alves. *Reflexões sobre a evasão escolar: uma problemática na educação brasileira*. Revista Teias, v. 21, p. 260-270, 2020.

SÃO PAULO. *Plano Estadual de Educação - PEE/SP - Relatórios de Monitoramento*. 2020. Disponível em <https://www.fde.sp.gov.br/PagePublic/Interna.aspx?codigoMenu=324>. Acesso em 19 nov. 2020.

SILVA, Fernanda Andrea Fernandes; SANTIAGO, Mônica Lins; DOS SANTOS, Marcelo Câmara. *Análise de itens da prova de matemática e suas tecnologias do ENEM que envolvem o conceito de números racionais à luz dos seus significados e representações*. Revista Eletrônica de Educação Matemática, v. 8, p. 190-208, 2013.

SILVA OSSES, Aníbal Tomás. *Análise da predição da violência infantil por meio de árvores de decisão e regras de associação*. Universidade Federal de São Carlos. 2020.

SOULE, Patrick. *Predicting student success: A logistic regression analysis of data from multiple siu-c courses*. Carbondale. OpenSIUC, 2017.

TAVARES, Fernando Oliveira; PACHECO, Luís Miguel; BORGES, Jorge. *Fatores indicadores do preço de um quarto de hotel: Uma aplicação a uma amostra de hotéis portugueses*. Revista Espacios, volume 37, nº 26, p. 1-11. 2016.

TRAVITZKI, Rodrigo. *Avaliação da qualidade do Enem 2009 e 2011 com técnicas psicométricas*. Estudos em Avaliação Educacional, v. 28, n. 67, p. 256-288, 2017.

TRAVITZKI, Rodrigo. *ENEM: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar*. 2013. 320 p. Faculdade de Educação da Universidade de São Paulo, São Paulo, 2013.

UOL. *Profissões do Futuro: Ciência de Dados*. 2020. Disponível em <https://vestibular.brasilescola.uol.com.br/profissoes-futuro/ciencia-de-dados.htm>. Acesso em 19 nov. 2020.

VERNAGLIA, Teresa. *Meninas sem banheiro em casa têm notas 25% menores no Enem*. 2020. Disponível em <https://www.uol.com.br/universa/noticias/redacao/2020/07/14/lata-dagua-e-um-dos-desafios-da-pandemia-diz-executiva-de-saneamento.htm>. Acesso em 19 nov. 2020.

WIENBUSCH, Eloisa Maria. *Avaliação em larga escala: uma possibilidade para a melhoria da aprendizagem*. In: Anais do IX ANPED Sul, GT05: Estado e Política Educacional. Caxias do Sul/RS: ANPED Sul, 2012.