UNIVERSIDADE FEDERAL DO PARANÁ

Adriane Machado (GRR20149152)

Cinthia Zamin Cavassola (GRR20149075)

Luiza Hoffelder da Costa (GRR20149107)

INFLUÊNCIA DE CARACTERÍSTICAS RELACIONADAS À ESCOLA, DEMOGRÁFICAS E SOCIAIS NO DESEMPENHO MÉDIO DOS ALUNOS NO ENSINO MÉDIO EM PORTUGAL NA DISCIPLINA DE MATEMÁTICA

Curitiba, 2017

Adriane Machado
Cinthia Zamin Cavassola
Luiza Hoffelder da Costa

RELATÓRIO REFERENTE À ANÁLISE DE DADOS UTILIZANDO MÉTODOS ESTATÍSTICOS DE MODELOS DE REGRESSÃO LINEARES PARA REALIZAR UM ESTUDO SOBRE A RELAÇÃO DE VÁRIAVEIS DIVERSAS COM DESEMPENHO EM MATEMATICA DOS ALUNOS NO ENSINO MÉDIO EM PORTUGAL

Relatório técnico apresentado como atividade avaliativa na disciplina de Modelos Lineares Generalizados, no Curso de Graduação em Estatística, na Universidade Federal do Paraná.

Professor Cesar Augusto Taconeli

Curitiba, 2017

RESUMO

Os dados apresentam os resultados de estudantes de ensino médio em duas escolas de Portugal. Os atributos dos dados incluem as notas dos estudantes e aspectos dos alunos social e demograficamente relevantes, assim como aspectos das escolas objeto do estudo. Os dados foram coletados usando relatórios escolares e questionários. A base de dados foi formada a partir da performance escolar em matemática. A base de dados foi modelada por meio de técnicas de regressão linear. A nota final, G3, possui uma correlação forte com as notas G1 e G2: isso ocorre pois G3 é a nota do último ano do ensino médio, enquanto G1 e G2 correspondem às notas do primeiro e segundo ano, respectivamente. Para a análise, foi feita a média das notas G1, G2 e G3 como variável dependente.

PALAVRAS-CHAVE: PERFORMANCE; ENSINO MÉDIO; PORTUGAL; MATEMÁTICA; REGRESSÃO LINEAR

SUMÁRIO

1 INTRODUÇÃO

O presente trabalho tem por objetivo analisar o desemprenho na disciplina de matemática de alunos no ensino médio através de métodos estatísticos de análise de regressão linear.

A variável resposta é média das notas dos alunos nos três anos do ensino médio na disciplina de matemática. Para isso, dados referentes a características demográficas, sociais e relacionadas à escola serão utilizados.

A critério do pesquisador foram escolhidas características, também chamadas como covariáveis, que a priori poderiam ter alguma influência sobre a variável resposta; no caso há 33 covariáveis para explicar o desenvolvimento médio de alunos nos três anos do ensino médio na disciplina de matemática.

Foram analisados os dados de duas escolas portuguesas, obtidos via relatórios e questionários. Uma vez que o primeiro continha informações escassas (ou seja, apenas as notas e o número de ausências estavam disponíveis), foi complementado com o último, o que permitiu a coleção de vários atributos demográficos, sociais e relacionados à escola (por exemplo, idade do aluno, consumo de álcool, educação da mãe).

O objetivo final é identificar as principais variáveis que afetam o sucesso/falha educacional.

2 MATERIAL E MÉTODOS

2.1 MATERIAL

O banco de dados refere-se a informações coletadas em duas escolas de Portugal, nos anos de 2005 a 2006. Foram coletadas informações de 395 alunos, distribuídas nas seguintes covariáveis:

school: Escola - Binária ('GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
sex: Gênero - Binária ('F' - feminino ou 'M' - masculino)
age: Idade - Numérica (variação de 15 a 22)
address: Endereço - Binária ('U' - urbano e 'R'-rural)
famsize: Tamanho da família - Binária ('LE3' - menor ou igual a 3 ou 'GT3' - maior que 3)
Pstatus: Estado de coabitação dos pais - Binária ('T' - convivência ou 'A' - separado)
Medu: Escolaridade da Mãe - Categórica ('0' - nenhum, '1' - até 4º ano, '2' - 5º a 9º ano, '3' - ensino médio ou '4' - ensino superior)
Fedu: Escolaridade do Pai - Categórica ('0' - nenhum, '1' - até 4º ano, '2' - 5º a

9º ano, '3' - ensino médio ou '4' - ensino superior)
Mjob: Profissão da Mãe - Nominal ('teacher' - professora, 'health' - saúde, 'services' - serviços , 'at_home' - do lar ou 'other' - outro)
Fjob: Profissão da Pai - Nominal ('teacher' - professor, 'health' - saúde, 'services' - serviços , 'at_home' - do lar ou 'other' - outro)
reason: Razão para estudar na escola - Nominal ('home' - perto de casa, 'reputation' - reputação, 'course' - curso de preferência ou 'other' - outro)
guardian: Guardião - Nominal ('mother' - mãe, 'father' - pai ou 'other' - outro)
traveltime: Tempo até chegar na escola - Numérica (1 - <15 min., 2 - 15 a 30 min., 3 - 30 min a 1 hora ou 4 - > 1 hora)
studytime: Horas de estudo semanal - Numérica (1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, or 4 - > 10 horas)
failures: Número de reprovações nos anos anteriores - Numérica (n - 1 a 3 ou 4 - 3 ou mais)
schoolsup: Ajuda da escola - Binário ('yes' - sim ou 'no' - não)
famsup: Ajuda da família - Binário ('yes' - sim ou 'no' - não)
paid: Aulas extras particulares - Binário ('yes' - sim ou 'no' - não)
activities: Atividades extracurriculares - Binário ('yes' - sim ou 'no' - não)
nursery: Frequentaram pré-escola - Binário ('yes' - sim ou 'no' - não)
higher: Pretende cursar ensino superior - Binário ('yes' - sim ou 'no' - não)
internet: Acesso à internet em casa - Binário ('yes' - sim ou 'no' - não)
romantic: Possui um relacionamento - Binário ('yes' - sim ou 'no' - não)
famrel: Qualidade das relações familiares - Numérica (de 1 - muito ruim até 5 - excelente)
freetime: Tempos livres - Numérica (de 1 - pouco até 5 - muito)
goout: Sair com os amigos - Numérica (de 1 - pouco até 5 - muito)
Dalc: Consumo de álcool durante a semana - Numérica (de 1 - pouco até 5 - muito)
Walc: Consumo de álcool final de semana - Numérica (de 1 - pouco até 5 - muito)
health: Estado da saúde - Numérica (de 1 - pouco até 5 - muito)
absences: Número de faltas - Numérica (de 0 até 93)
G1: Nota primeiro ano - Numérica (de 0 até 20)
G2: Nota segundo ano - Numérica (de 0 até 20)
G3: Nota terceiro ano - Numérica (de 0 até 20)

Através do uso de regressão linear é possível analisar e quantificar as relações entre a variável resposta e as variáveis explicativas. Como se tem mais que duas variáveis explicativas, será utilizada a regressão linear múltipla, a qual o modelo geral se encontra abaixo:

$$\mathbf{Y} = \beta 0 + \beta 1x1 + \beta 1x2 + \cdots + \beta px + \varepsilon$$

Para se efetivar a técnica da regressão linear múltipla, foram excluídos da base de dados alunos que por algum motivo não especificado apresentaram

média igual a zero ao final de algum dos anos do ensino médio, já que não foi fornecida a informação se o aluno realmente teve sua média igual a zero ou se a razão foi uma desistência ou outro motivo desconhecido.

O atributo G3 (nota no terceiro ano do ensino médio) possui uma correlação forte com os atributos G1 e G2 (respectivamente, notas no primeiro e no segundo ano), dado que se trata do mesmo indivíduo. Com o objetivo de reduzir o erro causado por essa correlação, criamos uma variável que se refere à média das notas dos três anos, a qual foi denominada G4.

3 MODELAGEM ESTATÍSTICA

Inicialmente foi criado um modelo contendo todas as covariáveis para verificar se as características incluídas no estudo teriam de fato efeito significativo no desempenho dos alunos; foram descartadas as covariáveis não significativas, detectadas via baixo p-valor de contribuição para a explicação do modelo, visualizada no resumo do modelo amplo.

Na sequência iniciou-se um processo de seleção de variáveis (atividade fundamental da modelagem estatística), onde é possível escolher e verificar quais variáveis influenciam significativamente no desempenho dos alunos. O algoritmo escolhido para esta tarefa foi o “stepwise”, algoritmo que faz uma inclusão e exclusão de variáveis no modelo utilizando como medida de seleção o AIC, conhecido como Critério de Informação de Akaike: menores valores de AIC indicam uma melhor contribuição do conjunto de covariáveis para a explicação do comportamento da variável resposta. A equação da reta ajustada foi:

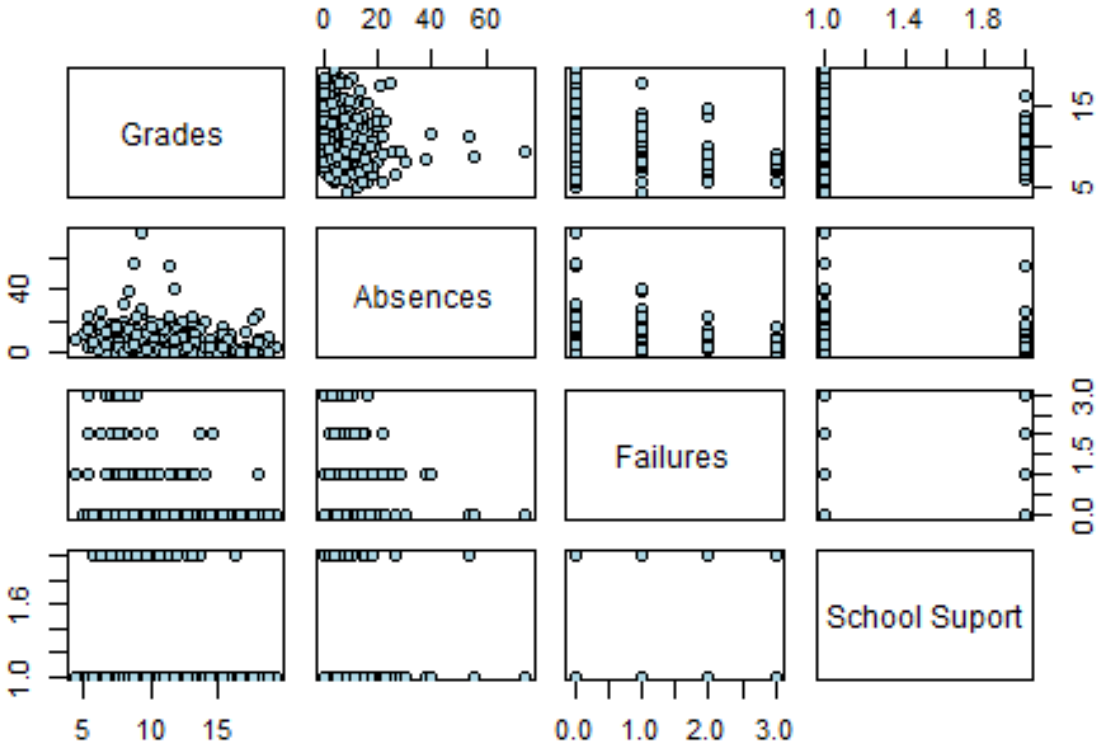
$$\hat{Y} = 12.37547 - \ 1.28359x_1 \ - 2.15943x_2 - 0.05392x_3$$

Onde: $x_1 = failures$
$x_2 = schoolsup$
$x_3 = absences$

As variáveis contidas no modelo são independentes e não possuem interação entre si, portando o modelo é uma regressão linear múltipla de característica aditiva. As covariáveis que apresentaram mais significância para o modelo foram o número de reprovações em anos anteriores, se o aluno obtinha alguma forma de o apoio da escola e o número de faltas.

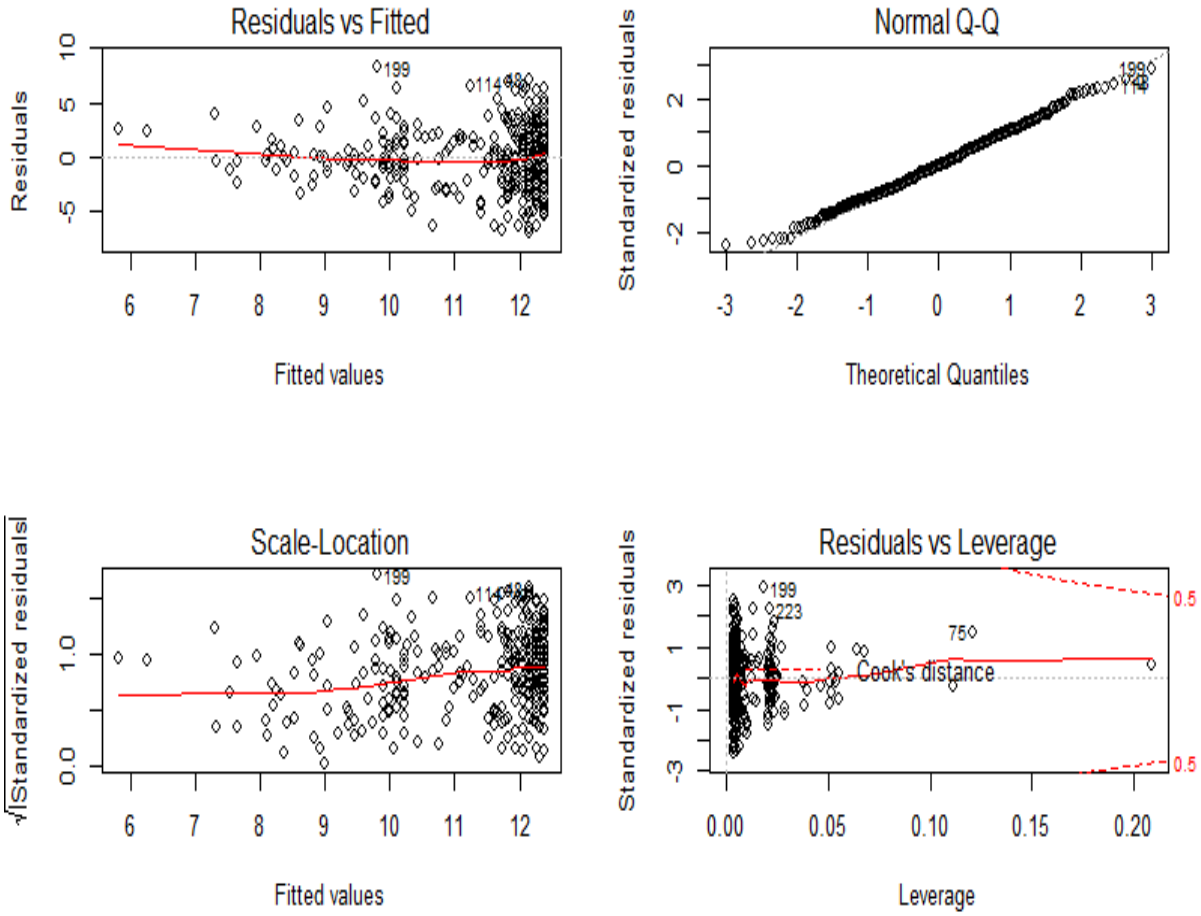
4 RESULTADOS E DISCUSSÃO

Definido o modelo, o interesse subsequente é avaliar a influência das variáveis “absences”, “failures” e “schoolsup” sobre a variável resposta. Para tal, foi feita uma análise exploratória visando avaliar essas relações. Na figura 1 são apresentados gráficos que mostram a relação das variáveis escolhidas para o modelo combinadas duas a duas.



Pode ser detectada uma tendência no número de faltas, o que pode ser explicado pela importância da presença dos alunos para compreensão do conteúdo.

Após a obtenção do modelo final ajustado e de uma breve interpretação do modelo, segue um conjunto de gráficos de diagnóstico de adequação do modelo, para verificar se todos os pressupostos então sendo atendidos e se há dados influentes.



Verifica-se tanto um preenchimento razoável dos pressupostos de normalidade (via qqplot) quanto de padrão aleatório dos resíduos (outros gráficos), sem pontos excessivamente influentes, indicando que o modelo atende aos pressupostos da regressão linear múltipla para sua validade.

5 CONCLUSÃO

Conclui-se que a regressão linear múltipla foi uma ferramenta útil para a obtenção de um modelo estatístico que pudesse selecionar os fatores preponderantes de influência na performance escolar em Matemática no ensino médio das duas escolas portuguesas analisadas e quantificar o grau de influência de cada covariável na variável resposta, via coeficientes do modelo. Verificou-se que os passos de verificação de pressupostos e seleção de covariáveis envolvidos no ajuste de um modelo de regressão não só validam a escolha da técnica, como refinam o modelo e o deixam corretamente alinhado com a realidade, contribuindo eficazmente ao estudo do fenômeno.