ANÁLISE DE DADOS RELACIONADO AO DESEMPENHO DOS ESTUDANTES

Dayhane Cristine da Silva Sosa dos Santos[1]

Rosineide Fernando da Paz[2]

Marina Mitie Gishifu Osio[3]

Resumo: Uma das formas do sistema de ensino público ser monitorado é por meio das avaliações que os estudantes realizam no período escolar. Essas avaliações são compostas por questões específicas de algumas disciplinas e por outras questões que informam sobre características das escolas e de todos os elementos que podem impactar no desempenho dos estudantes, por exemplos, nível socioeconômico dos estudantes, localização da escola, formação acadêmica dos professores, dentre outros. Todas essas informações são armazenadas numa base de dados e são disponibilizadas para serem analisadas, por exemplo, por métodos estatísticos apropriados. Assim, o objetivo deste trabalho é analisar estatisticamente dados educacionais da cidade de Bragança Paulista e verificar o impacto de algumas variáveis no baixo desempenho dos estudantes, nas avaliações de matemática. Como contribuição do trabalho pode ser citado a indicação aos gestores, preocupados com o índice de desempenho dos estudantes, onde há maior necessidade de atenção das políticas públicas, de modo que estes possam repensar as ações da comunidade escolar para melhorar a qualidade do ensino da matemática. Dentre os fatores investigados neste trabalho, observamos que o fato de o professor corrigir as tarefas dos alunos e o número de computadores que o aluno possui em casa, influenciam positivamente no seu desempenho. Além disso, o número de reprovações contribui com o baixo rendimento do estudante.

Palavras-chaves: Dados educacionais, modelo estatístico, avaliação.

STUDENT PERFORMANCE-RELATED DATA ANALYSIS

Abstract: A way to supervise the public education system is through the assessments that periodically students perform in the schools. These assessments are composed of specific issues of some school subjects and other issues that inform about characteristics of schools and all elements that may impact student performance, for example, students' socioeconomic level, school location, academic background of the teachers, among others. All such information is stored in a database and are available for analysis, for example, by appropriate

[1] Estudante – Licenciatura em Matemática – Instituto Federal de Educação Ciência e Tecnologia de São Paulo – Campus Bragança Paulista. E-mail: dayhane.sosa@gmail.com

[2] Professora adjunta da Universidade Federal do Ceará – Campus de Russas – Russas – Ceará. E-mail: rfdapaz@ufc.br

[3] Professora do Instituto Federal de Educação Ciência e Tecnologia de São Paulo – Bragança Paulista, Licenciatura em Matemática. E-mail: marina@ifsp.edu.br

statistical methods. Thus, the objective of this study is to analyze statistically the educational data of the city of Bragança Paulista and to verify the impact of some variables in the low performance of the students, in the mathematical assessments. As contribution of this work we can cite the indicating to the public managers about the impact of some characteristics about the reality of the student in their performance index, it can show where there is greater need of attention of the public policies to improve the quality of the teaching of mathematics. Among the factors investigated in this study, we observed that the fact of the teacher to corrects the student's tasks and the number of computers that the student has at home, positively influence in their performance. In addition, the number of disapprovals contributes to the student's poor performance.

Keywords: Educational data, statistical model, assessments.

INTRODUÇÃO

O Índice de Desenvolvimento da Educação Básica (IDEB) é um dos instrumentos utilizados para monitorar o sistema educacional das escolas. O IDEB utiliza a média de desempenho dos estudantes nas provas que são aplicadas periodicamente pelo Sistema de Avaliação da Educação Básica (SAEB). Juntamente com as provas desse sistema, os estudantes respondem questionários socioeconômico (Questionário do Aluno), cujas informações são armazenadas numa base de dados que após devidas análises permitem o monitoramento das políticas públicas educacionais e podem indicar variáveis que afetam o baixo rendimento nas provas. Compreendendo quais variáveis impactam no desempenho dos estudantes, os gestores podem realizar ações efetivas cujo objetivo é a melhoria do tão preocupante indicador de que a escola é ruim.

Muitas vezes, os professores são responsabilizados pelo baixo índice de desempenho dos alunos, por estarem em contato direto com os estudantes, mas nem sempre a culpa é do professor, podendo existir outros fatores que levam a esses baixos índices, como por exemplo, o fato de os pais ou responsáveis pelo aluno conversarem ou não com o estudante sobre o que acontece na escola, dentre outras. Outros fatores podem ser identificados por meio de uma análise de dados educacionais que ajuda na compreensão do relacionamento existente entre as diversas variáveis que podem influenciar o desenvolvimento da aprendizagem dos alunos. Com isso, é possível nortear as ações gerenciais de planejamento, manutenção e melhoria da qualidade dos resultados das avaliações e, consequentemente, a melhoria do IDEB onde todos são beneficiados.

O Questionário do Aluno busca obter informações com características pessoais e familiares, tais como nível educacional dos responsáveis, renda familiar, idade de ingresso no

sistema escolar, entre outras. A base de dados resultante das notas e do Questionário do Aluno é organizada de forma hierárquica, e um método estatístico muitas vezes recomendado para analisar dados com essa estrutura é o modelo de regressão misto ou multinível. Maiores detalhes sobre esse tipo de modelo podem ser vistos em Goldstein (2011), Manghi (2012) e em Osio (2013).

Para analisar dados que estão disponibilizados em um único nível, por exemplo, características relacionadas apenas aos alunos, uma alternativa são os modelos de regressão linear simples ou múltiplo. Para analisar se características da escola, em que os alunos estudam, que estão em outro nível, utilizamos modelos de dois níveis. Neste trabalho, utilizamos três tipos de modelos para tentar verificar se alguns fatores específicos influenciam no desempenho do aluno na prova de matemática. Para isso utilizamos os dados do SAEB de 2013 da cidade de Bragança Paulista. Com base nos resultados obtidos nesse trabalho, observamos que o fato do professor corrigir as tarefas dos alunos e número de computadores que o aluno possui em casa, influenciam positivamente no desempenho dos mesmos. Também observamos que o número de reprovações influencia negativamente nesse desempenho. Com a nossa análise, podemos concluir, também, que o estado de conservação da sala de aula, a frequência com que o aluno vai à biblioteca e o fato de os pais ou responsáveis conversarem ou não com os estudantes, não influenciam no seu desempenho.

Na iniciação científica, desenvolvida no ano de 2015, foram analisados os dados obtidos a partir da base de dados que contém informações do SAEB de 2013, aplicados aos estudantes do nono ano do ensino fundamental da rede pública. Apresentamos neste trabalho as informações relevantes do estudo realizado.

O trabalho está organizado da seguinte forma: na Seção 2 apresentamos uma discussão sobre os dados utilizados para avaliar o desempenho dos alunos e sua análise descritiva. A Seção 3 apresenta uma análise realizada por meio de um modelo de regressão linear simples. Na Seção 4 apresentamos os resultados obtidos por meio de um modelo de regressão linear múltiplo. A Seção 5 é dedicada a discussão dos resultados obtidos por meio de um modelo misto. Finalmente, na Seção 6 apresentamos a conclusão do trabalho.

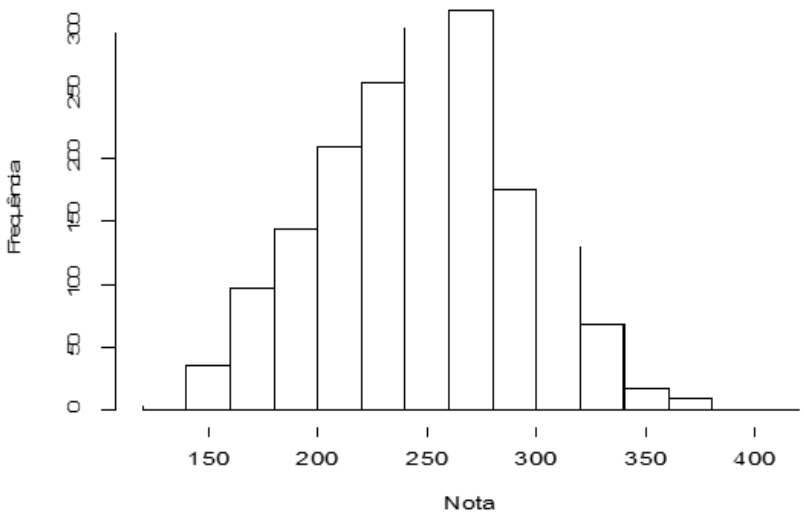
DADOS E ANÁLISE DESCRITIVA

A avaliação nacional do rendimento escolar conhecida como Prova Brasil foi criada e incorporada ao SAEB em 2005 com o objetivo de avaliar a qualidade do ensino ministrado

nas escolas das redes públicas e fornece indicadores contextuais sobre o trabalho das escolas. A Prova Brasil é aplicada a cada dois anos nas escolas que tenham no mínimo 20 alunos matriculados e os estudantes que participam da avaliação são os que estão no final de cada ciclo. Os dados analisados nessa seção estão disponíveis na página do Instituto Nacional de Estudos e Pesquisa e referem-se aos dados do SAEB da avaliação nacional do rendimento escolar - Prova Brasil, do ano de 2013. Foram utilizados os dados dos 1773 estudantes do nono ano do ensino fundamental, distribuídos nas 17 escolas da região do município de Bragança Paulista, estado de São Paulo.

Na análise descritiva dos dados, o histograma foi o primeiro gráfico construído, o qual norteou a escolha da distribuição para modelar os dados. A Figura 1 mostra o histograma das notas dos alunos o qual sugere que os dados vêm de uma distribuição normal.

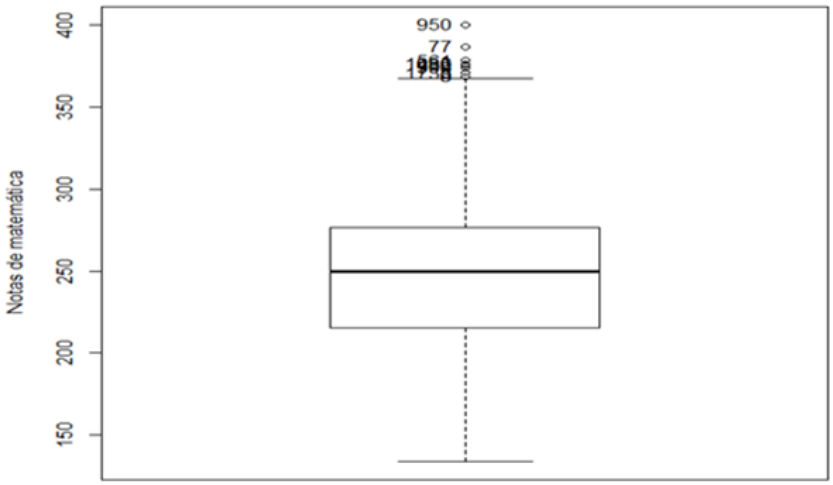
Figura 1 – Histograma das notas dos alunos na prova de matemática



Fonte: A pesquisa.

Na Figura 2 é apresentado o *box plot*, gráfico de caixa, das notas de matemática na prova dos alunos onde se podem observar alguns valores acima do limite superior, são os pontos exteriores ou valores atípicos, segundo Bussab e Morettin (2010). Com relação aos dados, esses valores correspondem aos estudantes que obtiveram nota muito diferente dos demais e, neste caso, notas maiores que as comumente observadas. Estes pontos podem ser bastante influentes nas análises realizadas.

Figura 2 – *Box plot* das notas de matemática



Fonte: A pesquisa.

A Tabela 1 apresenta o valor aproximado de algumas medidas de posição (quartis e média) e uma medida de dispersão (desvio padrão) referente ao conjunto de dados de notas de Bragança Paulista – SP.

Tabela 1 – Medidas resumo

1º quartil	Mediana	3º quartil	Média	Desvio padrão
215,70	249,66	276,73	247,81	45,08

Fonte: A pesquisa.

ANÁLISE POR MEIO DO MODELO DE REGRESSÃO LINEAR SIMPLES

Para essa e as próximas análises, foram escolhidas quatro questões do Questionário do Aluno - SAEB 2013 - da cidade de Bragança Paulista como variáveis que podem explicar a nota do aluno, obtida na Prova Brasil – 2013 na cidade de Bragança Paulista. As questões escolhidas são as questões de números 13, 39, 46, 55, apresentadas na Tabela 2. Especificamente, nesta seção vamos denotar por Y a variável nota de um aluno e tentar verificar se o fato de os responsáveis conversarem com os estudantes (Acompanhamento pelos pais, 1 - sim ou 2 - não) influenciam ou não no seu desempenho, ou seja, vamos

considerar a questão de número 31 da Tabela 2, como uma variável explicativa no modelo estatístico de regressão linear cuja variável resposta é representada por Y.

Tabela 2 – Descrição das questões do Questionário do Aluno

Pergunta nº.	Descrição	Alternativas
13	Na sua casa tem computador?	a. Não tem
		b. Sim, um
		c. Sim, dois
		d. Sim, três
		e. Sim, quatro ou mais
31	Seus pais ou responsáveis conversam com você sobre o que acontece na escola?	a. Sim
		b. Não
39	Com qual frequência você costuma ir à biblioteca?	a. Sempre, ou quase sempre
		b. De vez em quando
		c. Nunca ou quase nunca
46	Quando você entrou na escola?	a. Na creche (0 a 3 anos)
		b. Na pré-escola (4 a 5 anos)
		c. Na primeira série (6 a 7 anos)
		d. Depois da primeira série
55	O (A) professor (a) corrige o dever de casa de Matemática?	a. Sempre, ou quase sempre
		b. De vez em quando
		c. Nunca ou quase nunca
		d. Não passa dever de casa

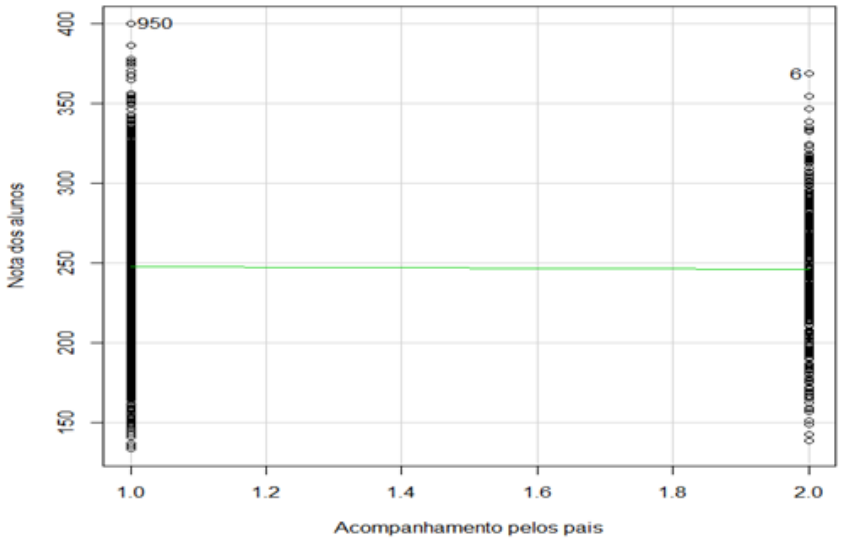
Fonte: A pesquisa.

Inicialmente, foi feita a suposição de que dada uma amostra aleatória, Y_1, …, Y_n, da variável aleatória Y, cada Y_i segue uma distribuição normal com média μ_i e variância σ^2 ($Y_i \sim N(\mu_i; \sigma^2)$), ou seja, a nota do aluno é considerada inicialmente como seguindo uma

distribuição normal com média $\mu_i \beta_0 + \beta_1 x_i$ e variância σ^2, em que x_i representa o i-ésimo valor observado da variável explicativa do modelo, para i = 1, ... , 1773.

A Figura 3 mostra o gráfico de dispersão com a reta ajustada, onde podemos observar que a média dos alunos que responderam sim (quando a variável assume o valor 1) não tem diferença significativa da média dos alunos que responderam não (quando a variável assume o valor 2), ou seja, não existem evidências de que a variável "Acompanhamento pelos pais" seja significativa no modelo ajustado.

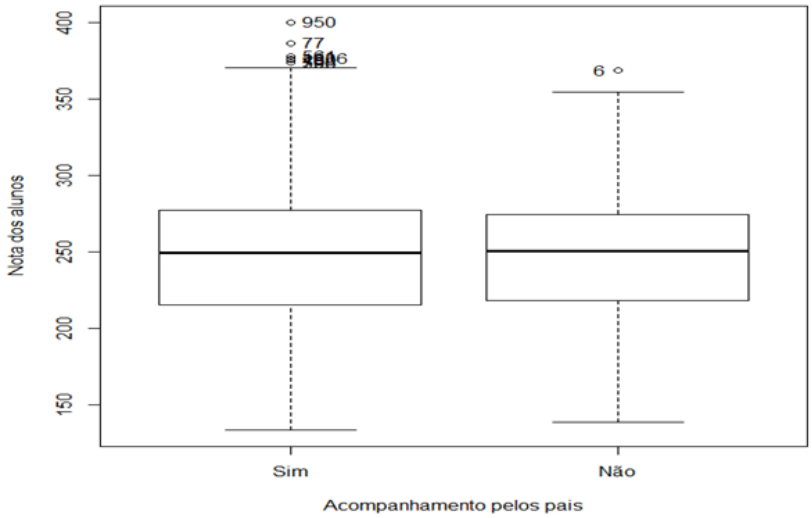
Figura 3 – Gráfico de dispersão da variável "Acompanhamento pelos pais" versus variável nota, com a reta ajustada pelo modelo de regressão linear simples



Fonte: A pesquisa.

Embora as maiores notas encontradas ocorressem quando os pais conversam com os alunos sobre o que acontece na escola, existe uma variabilidade maior neste campo, como mostra o *box plot* da figura 4.

Figura 4 – *Box plot* da variável "Nota" para alunos que tiveram ou não acompanhamento pelos pais



Fonte: A pesquisa.

Nos gráficos apresentados nas Figuras 3 e 4, pode-se perceber que a variável escolhida parece pouco significativa para prever o rendimento das notas dos alunos. Essas análises sugerem que não houve mudança significativa em relação à nota média de cada grupo.

As estimativas dos coeficientes da regressão linear simples e o p-valor para cada uma delas podem ser vistos na Tabela 3. Com base no teste hipótese $H_0: \beta_1 = 0$, podemos concluir que a variável "Acompanhamento pelos pais" não é significativa, pois o p-valor é de 0.461. Então, ao nível de 95% de confiança, não rejeitamos a hipótese nula em que $\beta_1 = 0$, ou seja, com base nesses resultados concluímos que o fato de os responsáveis conversarem com os alunos sobre o que ocorre na escola não influenciam no desempenho do aluno na prova de matemática.

Tabela 3 – Estimativas dos coeficientes do modelo de regressão simples

	Estimativa	Erro padrão	p-valor
β_0	250, 214	3, 430	$2e^{-16}$
β_1	-2, 022	2, 745	0, 461

Fonte: A pesquisa.

ANÁLISE POR MEIO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLO

Como dito na seção anterior, neste trabalho outras variáveis foram consideradas como possíveis fatores que podem explicar a nota do aluno na prova de matemática. Para isso, consideramos um modelo de regressão linear múltiplo. Para esse modelo, vamos supor, inicialmente, as variáveis número de computadores que o aluno possui em sua casa (NComp), a frequência semanal do aluno à biblioteca (FBib), idade do aluno ao ingressar na escola (IIEsc) e frequência de correção da tarefa pelo professor (FCTaref) como variáveis que podem explicar a variável nota do aluno (Y), essas variáveis estão detalhadas na Tabela 2. O modelo completo, com todas as variáveis explicativas consideradas inicialmente, pode ser representado como:

$$Y_i = \beta_0 + \beta_1 NComp_i + \beta_2 FBib_i + \beta_2 IIEsc_i + \beta_3 FCT_i + \varepsilon_i,\ \ \varepsilon_i \sim \mathrm{N}\,(0, \sigma^2),$$

para i=1, …,1773, em que Y_i é a nota do i-ésimo aluno e $NComp_i$, $FBib_i$, $IIEsc_i$ e $FCTaref_i$ representam os valores das variáveis explicativas na i-ésima observação. Outra forma de representar o modelo é:

$$Y_i \sim \mathrm{N}\,(\mu_i;\ \sigma^2),$$

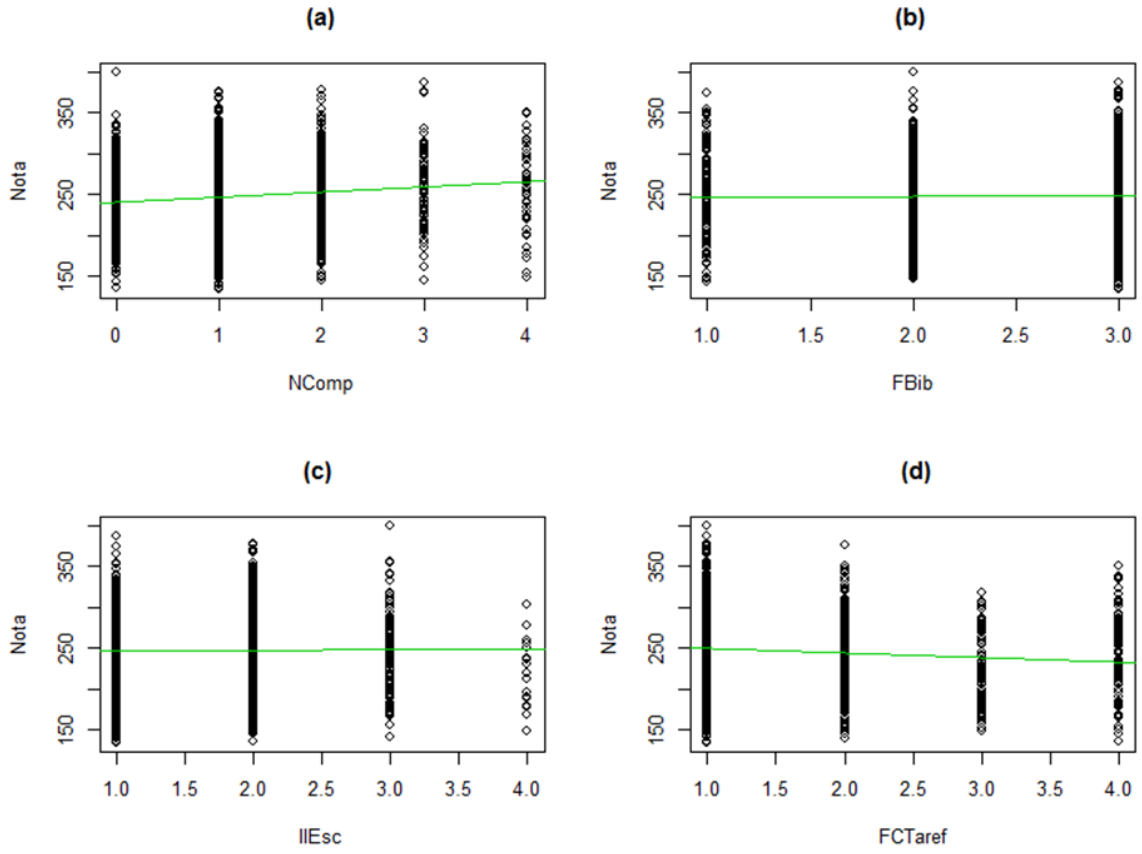
com $\mu_i = E[Y_i] = \beta_0 + \beta_1 NComp_i + \beta_2 FBib_i + \beta_3 IIEsc_i + \beta_4 FCTaref_i$, sendo a média da variável aleatória, para i=1,…,1773, em que β_0 é a média global da Nota.

. O diagrama de dispersão para cada uma das variáveis, NComp, FBib, IIEsc e FCTaref, versus a variável Nota é apresentado na Figura 5, com a reta ajustada usando o modelo linear simples. Com a construção dos diagramas de dispersão é possível ilustrar o quanto os valores observados se desviam da média e quais das variáveis podem explicar o desempenho dos alunos. Nessa figura podemos observar que as variáveis “idade em que ingressou na escola” e “frequência à biblioteca” parecem não influenciar no desempenho do aluno. No entanto, a nota do aluno parece diminuir conforme o professor diminui a frequência da correção das tarefas.

A Figura 5 mostra os gráficos (a), (b), (c) e (d) que contêm as retas ajustadas por meio do modelo de regressão linear simples para as variáveis Ncomp, Fbib, IIEsc e FCTaref, respectivamente. Nessa figura, o gráfico (a) fornece evidências de que o fato do aluno frequentar a biblioteca não influencia no desempenho do aluno. Ou seja, pode-se notar que a média das notas não parece variar conforme a frequência do aluno à biblioteca. Com relação à idade de ingresso na escola (IIEsc), o gráfico (c), dessa figura, não fornece evidências de que o fato dos alunos terem acesso à escola com idade precoce influência no seu desempenho,

pois as médias das notas dos grupos de alunos que ingressaram na escola mais cedo parece diferenciar de forma muito insignificante da média de grupos de alunos que ingressaram na escola com mais idade, dando evidências de que não faz diferença na nota da prova de matemática o aluno ter entrado mais cedo ou mais tarde na escola. Já o fato do professor corrigir o dever de casa da disciplina de matemática aparentemente tem relação com o desempenho do aluno, pois pode ser notado no gráfico (d) que parece existir diferença significativa entre a média das notas dos alunos que tiveram suas tarefas corrigidas pelo professor durante a vida escolar e a média das notas dos alunos que não tiveram suas tarefas corrigidas. Além disso, as maiores notas podem ser observadas quando o valor dessa variável explicativa é 1, ou seja, entre os alunos que sempre tiveram suas tarefas corrigidas pelo professor. O gráfico (a) mostra uma pequena inclinação da reta, que pode significar uma influência do número de computadores que o aluno tem em sua residência na nota da prova de matemática. Essa possível dependência será discutida mais adiante.

Figura 5 – Diagrama de dispersão das variáveis NComp, FBib, IIEsc e FCTaref versus a variável Nota, com restas ajustadas pelo modelo linear simples



Fonte: A pesquisa.

A Tabela 4 mostra os valores dos coeficientes de correlação Pearson (veja, por exemplo, BUSSAB & MORETTIN, 2010) calculados para medir a correlação entre as variáveis consideradas neste trabalho. O coeficiente de correlação de Pearson é uma medida que varia entre -1 e 1, em que zero indica que não existe correlação entre as variáveis, 1 significa que as variáveis são totalmente correlacionadas positivamente e -1 quando são totalmente correlacionadas negativamente. Pode ser notado nesta tabela que não existe evidência para forte correlação entre as variáveis explicativas. Ainda considerando os resultados apresentados na Tabela 4, pode-se concluir que as variáveis NComp e FCTaref possuem correlação com a variável Nota, como já havia sido observado na Figura 4 (a) e (d). Ou seja, entre Nota e NComp há uma correlação positiva e entre Nota e FCTaref uma correlação negativa, ambas consideráveis. Essa tabela também mostra que pelo coeficiente de correlação de Pearson existe pouco correlação entre as variáveis explicativas.

Tabela 4 – Matriz de Correlação entre as variáveis

	Nota	NComp	FBib	IIEsc	FCTaref
Nota	1				
NComp	0, 120	1			
FBib	0.004	0.009	1		
IIEsc	0.010	-0.078	-0.025	1	
FCTaref	-0, 101	0, 035	0.067	-0.031	1

Fonte: A pesquisa.

Para analisar quais das variáveis explicativas, consideradas inicialmente, influenciam no desempenho dos alunos na prova de matemática, foi feita uma seleção de variáveis por meio da comparação dos modelos obtidos pela inclusão ou exclusão de variáveis explicativas. Os modelos considerados são apresentados na Tabela 5, em que cada modelo é um modelo de regressão onde foram consideradas as variáveis explicativas descrita nessa tabela. Para comparar esses modelos foi usado o critério AIC, cujos valores estão ordenados na tabela em ordem crescente.

Sabemos que o critério AIC penaliza modelos com muitas variáveis sendo que valores menores de AIC são preferíveis. Note que não existe diferença significativa entre os três primeiros modelos apresentados na Tabela 5, pois os valores do AIC para esses três modelos

estão muito próximos. No entanto, usando o critério da parcimônia (que consiste em dar preferência a modelos com menos parâmetros) o modelo final escolhido é aquele que considera as variáveis explicativas Ncomp e FCTaref. Portanto o primeiro modelo mostrado na Tabela 5 foi escolhido como sendo o modelo que melhor descreve os dados observados da variável aleatória Nota, entre os modelos considerados. Para o ajuste dos modelos lineares simples e múltiplo, usamos o pacote "lme4" (Bates et al., 2015) do software R.

Tabela 5 – Estimativas dos AICs para os modelos considerados

Modelo	AIC
NComp + FCTaref	18497,88
NComp + IIEsc + FCTaref	18499,40
NComp + FBib + FCTaref	18499,70
NComp + FBib + IIEsc + FCTaref	18501,21
NComp	18515,73
FCTaref	18523.21
NComp + IIEsc	18517,07
NComp + FBib	18517,72
NComp + FBib + IIEsc	18519,05
FBib + FCTaref	18525,01
IIEsc + FCTaref	18525,13
FBib + IIEsc + FCTaref	18526,92
IIEsc	18541,05
FBib	18541,19
FBib + IIEsc	18543,02

Fonte: A pesquisa.

Considerando o modelo escolhido com base nos resultados apresentados na Tabela 5, foram obtidas as estimativas para os coeficientes da regressão, o erro padrão para cada valor estimado e o p-valor, cujos resultados estão apresentados na Tabela 6 com base nos valores obtidos para o p-valor, apresentados nessa tabela, pode-se concluir que as variáveis Ncomp e FCTaref são significativas para explicar a nota dos alunos na prova de matemática, ao nível de 95% de confiança.

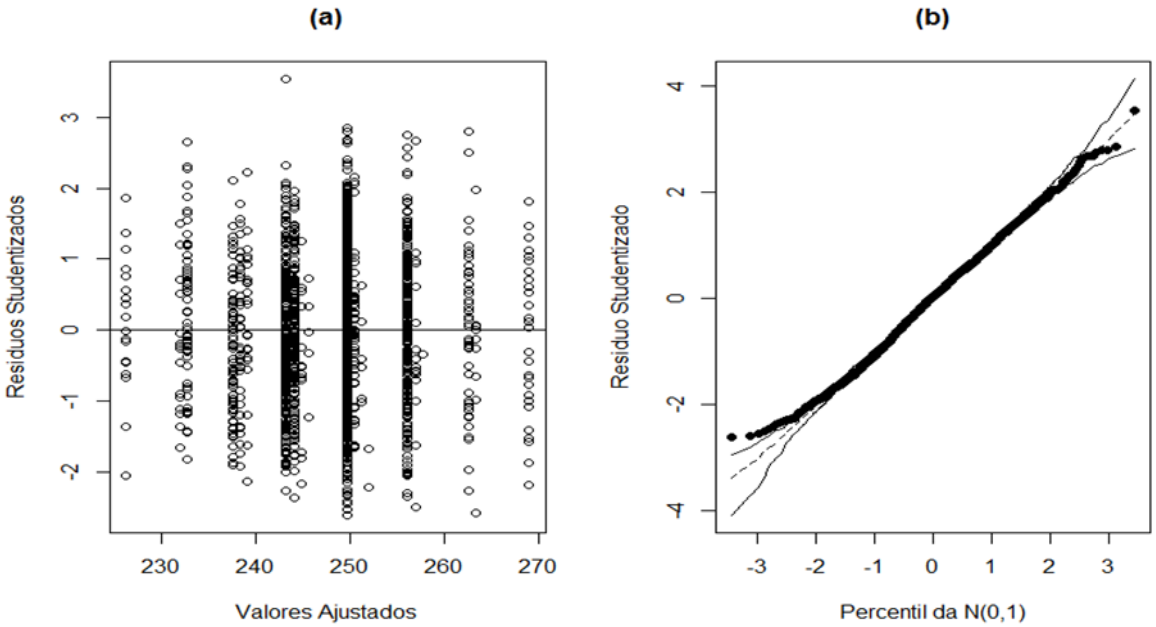
Tabela 6 – Estimativas dos coeficientes do modelo de regressão, erro padrão e p-valor do modelo escolhido utilizando o critério AIC

	Estimativa	Erro padrão	p-valor
β_0	242.456	3.290	2e-16
β_1	6.431	1.226	1.76e-07
β_2	-5.653	1.266	8.53e-06

Fonte: A pesquisa.

A Figura 6 (b) mostra o gráfico de probabilidades com o envelope simulado (PAULA, 2004) onde pode ser visto que quase todos os pontos se encontram dentro da banda de confiança, dando evidência de que a suposição de normalidade para os erros também está satisfeita. Ou seja, esses gráficos não apresentam indícios de que a distribuição normal seja inadequada para explicar as notas dos alunos na prova de matemática.

Figura 6 – Gráfico dos resíduos studentizados versus valores ajustados (a) e gráfico de probabilidade com envelope simulado (b)



Fonte: A pesquisa.

Das análises feitas até aqui, podemos concluir que a baixa frequência com que o professor corrige a tarefa de casa influencia negativamente no desempenho do aluno na prova

de matemática. Enquanto que quanto maior o número de computadores que existe na residência do aluno melhor é o seu rendimento na prova.

ANÁLISE POR MEIO DO MODELO DE REGRESSÃO LINEAR MISTO

O modelo linear misto é, de certa forma, similar ao modelo de regressão linear. Esse modelo estima o efeito de uma ou mais variáveis explicativas sobre a variável resposta (Y). Para o ajuste do modelo linear misto, também usamos o pacote "lme4" (Bates et al., 2015) do software R. As saídas do R para esse pacote fornecem uma lista de valores das variáveis explicativas, estimativas e intervalos de confiança dos seus efeitos, p-valores para cada efeito e uma medida de bondade de ajuste. O modelo especificado aqui possui dois níveis, nesse modelo o aluno e tratado como a unidade do nível 1, identificado pelo subscrito i, e a escola como unidade do nível 2, identificada pelo subscrito j. Considera-se a existência de J escolas, j = 1,2, ... ,J cada uma delas com n_j alunos, i=1,2,... n_j.

Ao responder à questão, cujo objetivo é avaliar as condições de funcionamento da escola, os alunos puderam avaliar se na escola existem recursos para auxiliar no ensino aprendizagem, por exemplo, se na escola existe ou não televisão (tv) e vídeo cassete e as condições de uso. As três categorias distintas da resposta sobre os materiais disponíveis são razoáveis ou ruins. Portanto, as variáveis consideradas, neste exemplo, são nomeadas por $Nurep_{ij}$ e $Sala_j$, que representam o número de reprovações que aluno i apresenta em seu histórico escolar e que estuda na escola j e condições de uso da sala que existe na escola j, respectivamente. Denotamos por Y_{ij} o rendimento escolar em Matemática do aluno i que estuda na j-ésima escola. Assim, para modelar os dados, usamos o modelo misto especificado aqui pelas seguintes equações:

$$
\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j} Nurep_{ij} + \varepsilon_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} Sala_j + b_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} Sala_j + b_{1j} \qquad (1) \\
\varepsilon_{ij} &\sim N(0, \sigma_e^2)
\end{aligned}
$$

$\mathbf{b_j} \sim N_2(\mathbf{0}, \mathbf{D})$, para para i = 1, 2, ..., n_j, j = 1, 2, ..., 17, J = 17, com $n_1 + n_2 + \ldots + n_{17} = 1773$, em que $bj = (b_{0j}, b_{1j})$ e $N_2(\mathbf{0},\mathbf{D})$ representa a distribuição normal bivariada com vetor de médias **0**=(0,0) e matriz de variâncias e covariâncias $\mathbf{D} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$.

Tabela 7 – Estimativas dos coeficientes do modelo de regressão e intervalos de confiança

Parâmetro	γ_{00}	γ_{01}	γ_{10}	γ_{11}	σ	τ_{00}	τ_{11}
Estimativa	261.7	6.5	-28,5	2.3	42,7	16.20	16.85
Intervalo de confiança	(234,6; 288,7)	(- 4,9; 18,0)	(- 47,1; -6,8)	(-6,6; 9,9)	(41,3; 44,0)	(0,0; 67,0)	(0,0; 43,2)

Fonte: A pesquisa.

Os resultados obtidos por meio dessa análise estão apresentados na Tabela 7, onde podemos observar as estimativas para cada parâmetro do modelo com os respectivos intervalos de confiança obtidos pelo método *bootstrap*. Com base nos resultados apresentados nessa tabela, podemos perceber que existem evidências de que o número de reprovações obtida pelo aluno ($Nurep_{ij}$) influencia negativamente no seu desempenho, pois o parâmetro γ_{10} é negativo e parece ser significativo no modelo uma vez que o intervalo de confiança para esse parâmetro não contém o valor zero, (-29,4; -2,8). Neste modelo o efeito da escola é capturado por meio da inclusão da variável $\mathbf{b_j}$, para j=1, …,17, cujas estimativas dos parâmetros podem ser vistas na Tabela 5.1 ($\tau_{00,}$ τ_{10}), aqui consideramos um modelo que não leva em conta a correlação. Com base nessa tabela, a interação entre as variáveis Salaj e o número de reprovações ($Nurep_{ij}$) do aluno i da escola j, parece não ser significativo no modelo, assim como a variável $Sala_j$ que representa o estado de conservação da sala de aula. Devido à evidencia de essa variável não ser significativa, novos modelos foram testados.

Testamos um modelo sem a interação entre a as variáveis Salaj e $Nurep_{ij}$, no entanto a variável estado da sala de aula continuou não sendo significativa no modelo. Além disso, fizemos uma comparação entre três modelos: o Modelo 1 é o modelo completo apresentado pelas equações (1), o Modelo 2 é o modelo que não leva em conta a variável Sala e o Modelo 3 não considera a variável Sala e a variável Nurep é incluída apenas no efeito fixo. Os resultados dessa comparação estão apresentados na Tabela 8 onde podemos ver que não existem diferença significativa no ajuste desses modelos, segundo os critérios AIC, BIC e *deviance*. Assim, seguindo o critério da parcimônia, optamos pelo Modelo 3 por conter um número menor de parâmetros a serem estimados. As estimativas para os parâmetros do Modelo 3 estão apresentadas na Tabela 9, onde podemos ver que a variável Nurep continua significativa no modelo, como é de se esperar.

Tabela 8 – Modelos propostos e comparação dos modelos

Modelo 3: $E[Y_{ij}] = \gamma_{00} + \gamma_{10} Nurep_{ij} + b_{0j}$				
Modelo 2: $E[Y_{ij}] = \gamma_{00} + \gamma_{10} Nurep_{ij} + b_{1j}Nurep_{ij} + b_{0j}$				
Modelo 1: $E[Y_{ij}] = \gamma_{00} + \gamma_{01}Sala_j + \gamma_{10} Nurep_{ij} + \gamma_{11} Nurep_{ij} Sala_j + b_{1j}Nurep_{ij} + b_{0j}$				
Critério	Df	AIC	BIC	*deviance*
Modelo 3	4	18376	-9183.9	18368
Modelo 2	6	18379	-9183.2	18367
Modelo 1	7	18373	-9179.7	18359

Fonte: A pesquisa.

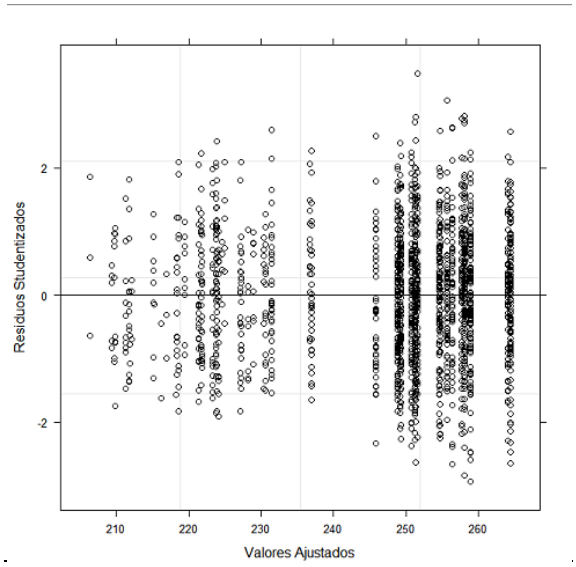
Tabela 9 – Estimativas dos coeficientes do modelo de regressão e intervalos de confiança

Parâmetro	γ_{00}	γ_{01}	σ	τ_{00}
Estimativa	276.7	-22,9	42,8	6,5
Intervalo de confiança	(270,2; 282,9)	(-26,6; -19,7)	(41,3; 44,0)	(2,5; 9,9)

Fonte: A pesquisa.

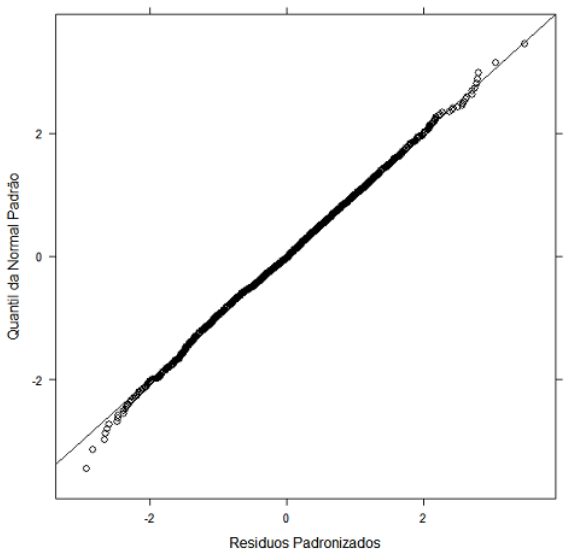
Para o modelo escolhido foi feita uma análise de resíduos, utilizando um conjunto de técnicas que além de serem utilizadas para verificar as suposições assumidas para o modelo, servem para analisar a qualidade do ajuste e auxiliam, por meios de gráficos, na busca de pontos discrepantes. O gráfico dos resíduos condicionais apresentado na Figura 7 fornece evidências de que o modelo está bem ajustado, pois o gráfico de resíduos não apresenta nenhuma estrutura, dando evidência de homoscedasticidade, ou seja, confirmamos a homogeneidade em relação as observações. Nesse gráfico podemos verificar a existência de alguns pontos fora do intervalo (-2,2), mas essas observações não comprometem a verificação da suposição. Quanto à normalidade dos erros, pode ser verificado na Figura 8 que não há fortes evidências do afastamento da suposição de normalidade.

Figura 7 – Gráfico dos resíduos studentizados versus os valores ajustados para o Modelo 3



Fonte: A pesquisa.

Figura 8 – Gráfico quantil-quantil, *qqplot*, para o Modelo 3



Fonte: A pesquisa.

CONCLUSÃO

Neste trabalho, utilizando modelo de regressão linear simples, múltiplo e misto para analisar alguns dos fatores que podem estar relacionados com o rendimento do aluno na prova de matemática do SAEB. Os dados utilizados estão disponíveis na página do Instituto Nacional de Estudos e Pesquisa e referem-se aos dados do sistema de avaliação da educação

básica, avaliação nacional do rendimento escolar - prova Brasil, do ano de 2013 da cidade de Bragança Paulista. Com o modelo de regressão linear simples verificamos que o fato de os pais conversarem com o aluno sobre o que ocorreu na escola, não influencia no rendimento dessa prova.

Com base nas estimativas dos coeficientes da regressão múltipla, pode-se observar que o fato de o professor corrigir cada vez menos a tarefa dos alunos influencia negativamente no desempenho desses alunos na prova, enquanto que, quanto mais computadores o aluno tem em casa, melhor é seu desempenho na prova. A possibilidade do número de computadores que tem na residência do aluno influenciar em sua nota, não parece ter fundamento. Então, devemos considerar a hipótese desta correlação ser um caso de correlação espúria, caso em que é encontrada uma correlação significativa, mas que não é uma autêntica dependência. Nesse caso, deve-se estudar com mais detalhe outros fatores que possam afetar ambas as variáveis simultaneamente, nota do aluno e número de computadores, e que possam explicar a correlação existente entre essas variáveis. Uma possível justificativa para provável correlação é a situação econômica do aluno, pois o número de computadores que o aluno tem em casa pode ser um indicativo de que este aluno tem uma situação econômica melhor do que aqueles que não têm, ou têm poucos computadores em casa, neste caso a situação econômica seria significativa para explicar a nota e não o número de computadores em si, caracterizando uma correlação espúria. Outra explicação para a significância dessa variável pode ser o fato de que quanto mais computadores o aluno tem em sua residência, mais esse aluno está exposto à informação, pois este pode buscar informações na internet a qualquer momento, uma vez que, mesmo que a família seja grande, esse aluno terá a oportunidade de usar o computador em sua casa. Neste último caso, não se teria um caso de correlação espúria, uma vez que a correlação seria verdadeira. Portanto uma análise mais aprofundada dessa correlação deve ser feita para investigar essa possível dependência de forma mais efetiva. Esse tema pode ser tratado em trabalhos futuros.

Com base no modelo de regressão linear misto, concluímos que o número de reprovações afeta o desempenho na prova, enquanto o estado de conservação da sala de aula parece não ser significativo para explicar o desempenho.

REFERÊNCIAS

BATES, Douglas et al. lme4: Linear mixed-effects models using Eigen and S4. **R package version**, v. 1, n. 7, p. 1-23, 2014.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica**. Saraiva, 2010.

HOX, Joop J.; MOERBEEK, Mirjam; VAN DE SCHOOT, Rens. **Multilevel analysis**: Techniques and applications. Routledge, 2010.

MANGHI, Roberto Ferreira. **Modelos elípticos multiníveis**. Tese de Doutorado. Universidade de São Paulo.

MARSCHNER I. 2014. **Fitting Generalized Linear Models**. Repository CRAN.

OSIO, Marina Mitie Gishifu. **Análise de modelos de regressão multiníveis simétricos**. 2013. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação.

PAULA, Gilberto Alvarenga. **Modelos de regressão**: com apoio computacional. São Paulo: IME-USP, 2004.

Recebido em 15 nov 2018; Aceito após revisão em 10 mar 2019.