Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Desempenho no Enem dos concluintes do Ensino Médio em escolas da AMB:

Uma Abordagem Multinível

Roberto de Souza Marques Buffone

Orientadora: Profª. Maria Teresa Leão Costa

Brasília

2019

Roberto de Souza Marques Buffone

Desempenho no Enem dos concluintes do Ensino Médio em escolas da AMB:
Uma Abordagem Multinível

Trabalho de Conclusão de Curso apresentado no Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção de título de Bacharel em Estatística.

Orientadora Profª. Maria Teresa Leão Costa

Brasília

2019

Para (e por) Deus.
Para (e por) minha família.
Para (e por) minhas mães Marias.
Mãe, pai, irmão e dinda, nós conseguimos!

Agradecimentos

Agradeço primeiramente a Deus, educador do universo. Sem Sua mão, não existiria a caminhada até aqui traçada. Agradeço a Maria Santíssima, minha advogada e protetora.

Agradeço a toda a minha família, que me deu suporte nos momentos de maior necessidade. Em especial, minha mãe Maria do Carmo de Souza Marques que sempre me incentivou em todas as minhas escolhas e meu pai Giacomo Buffone, hoje não mais presente em corpo, mas certamente contente em espírito por minha conquista. Agradeço a meu irmão Marcone de Souza Marques, que sempre surgiu em minha vida como um segundo pai. Agradeço a minha dinda Bianca de Souza Marques, que me acolheu em sua casa no início da jornada, me tratando como um filho. Agradeço a meu primo e grande amigo Rodrigo de Souza Silva, que me deu as primeiras aulas de lógica de programação e a minha cunhada Érika Lopes de Carvalho de Souza Marques, que sempre me deu os melhores conselhos possíveis.

Agradeço às pessoas que moldaram a minha paixão pelas ciências exatas, sabendo que sem elas eu não chegaria aqui. Em especial, ao Professor Hilnei Macedo da Silva, professor de matemática do meu ensino médio. Saúdo a todo o corpo docente do Colégio Estadual Teotônio Marques Dourado Filho, da minha querida Morro do Chapéu.

À minha orientadora, Professora Maria Teresa Leão Costa, que me ajudou em todo o período acadêmico, culminando com este trabalho de conclusão, sempre com toda a paciência possível. Agradeço aos professores da banca avaliadora, Dr^{a}. Juliana Betini Fachini Gomes e Dr. Leandro Tavares Correia, que compartilharam suas experiências para uma melhor execução deste trabalho. Agradeço também a todos professores do Departamento de Estatística da Universidade de Brasília, fornecendo seu conhecimento de forma ímpar a todos que ali os procuravam.

Agradeço a meus amigos que estiveram comigo durante toda a caminhada da graduação e aqueles que infelizmente, por consequências da vida, ficaram pelo caminho. Aos amigos que não estiveram fisicamente presentes durante minha formação mas que sempre emanaram energias positivas, torcendo por minha conquista.

Por fim, agradeço ao INEP pela disponibilidade das informações que neste trabalho estão contidas e ao SAS, que além de fornecer um *software* capaz de realizar minhas análises, proveu um referencial teórico imensamente útil.

A todos, muito obrigado!

"Pois Eu, o Senhor, teu Deus, Eu te seguro pela mão e te digo:
Nada temas, Eu venho em teu auxílio."
(Isaías 41, 13)

Resumo

Motivado pela importância do âmbito educacional no desenvolvimento do ser-humano e da sociedade e quem ele vive, viu-se a necessidade de um estudo que levasse em consideração fatores potenciais para alteração do desempenho de um estudante.

Considerando uma modelagem linear multinível, técnica utilizada quando as unidades observacionais estão interligadas de alguma forma, foi analisado o desempenho médio das 5 competências do Exame Nacional do Ensino Médio (Enem) - Linguagens, Códigos e suas tecnologias, Ciências da Natureza e suas tecnologias, Ciências humanas e suas tecnologias, Matemática e suas tecnologias e Redação - dos concluintes do ensino médio em escolas da Área Metropolitana de Brasília, para o ano de 2018. Buscou-se encontrar fatores ligados à condição socioeconômica do estudante que podem vir a interferir no desempenho do aluno, bem como aspectos associados às escolas, entidade que explica em média 45% da variabilidade do desempenho dos discentes, justificando assim o uso da técnica escolhida. Para auxiliar o estudo, o Censo Escolar realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) foi utilizado como base secundária, trazendo uma maior quantidade de informações para o estudo. Durante o processo de obtenção de índices tanto de infraestrutura quanto socioeconômicos foram utilizadas técnicas de Análise Fatorial, considerando variáveis sabidamente importantes para esses quesitos.

Fatores relacionados ao estudante como sua idade, sexo, cor ou raça, nível socioeconômico e ocupação de trabalho dos responsáveis são itens chave para a estimativa de desempenho obtida. Quanto a variáveis ligadas a escolas, fatores como a localidade da instituição, a dependência administrativa, a média de alunos por turma, média de horas-aula por dia e proporção de docentes com ensino superior importam para a nota obtida pelo aluno no exame. Esse fatores em conjunto constroem o modelo obtido no estudo.

Palavras-chave: Desempenho educacional; Modelos Lineares Multiníveis; Exame Nacional do Ensino Médio; Área Metropolitana de Brasília.

Sumário

Lista de ilustrações

Lista de tabelas

1 Introdução

Os impactos dos investimentos na educação são de difícil compreensão dado que, não existe uma interferência apenas naquele que estuda, mas também em todo o ambiente que o permeia (BARROS; MENDONÇA, 1998). Para a avaliação do retorno que estes investimentos trazem, diversos exames são utilizados como fonte de informação, tentando também identificar em que nível educacional essas aplicações tem maior eficiência.

Perante à importância do âmbito educacional no desenvolvimento do ser-humano e da sociedade em que ele vive, viu-se necessário estudos em grande escala da educação básica por meio de provas a nível nacional, tais como a Prova Brasil e o Enem. Esses testes subsidiam estudos que auxiliam o entendimento de fatores que impactam no desempenho cognitivo dos alunos, que podem ser categorizados em quatro grandes eixos: projetos pedagógicos, a estrutura escolar, a família e características do próprio aluno (SOARES; ALVES, 2013). Analisar os fatores com maior potencial para alterar o desempenho escolar dentro de cada categoria é um dos objetivos principais das avaliações em larga escala como o Enem (LAROS; MARCIANO; ANDRADE, 2012).

O Exame Nacional do Ensino Médio (Enem), prova aplicada de forma majoritária para concluintes do Ensino Médio, subsidia estudos e políticas públicas desde 1998. Inicialmente, o Enem tinha o único objetivo de observar o desempenho destes alunos e assim, analisar o investimento realizado nesse grupo focal. Atualmente, além de grande componente para observar o retorno obtido com os investimentos aplicados, o Enem serve também, em grande escala, como porta de entrada para a Educação Superior no Brasil.

Haja vista que, o presente sistema de educação possui uma verticalidade nítida, ou seja, uma pessoa surge como detentora do conhecimento e essa distribui sua sabedoria para níveis mais baixos (FREIRE, 1979), em que os receptores estão de certa forma agregados entre si, deve-se considerar essa questão ao analisar problemas no âmbito educacional. Com isso, fazendo uso de técnicas de análise multinível, que agregam às variáveis do aluno, informações quanto ao espaço em que estes estudam, busca-se conhecer o quanto essa organização educacional interfere no resultado pessoal de cada aluno.

Num país de dimensões continentais, com desigualdades sociais presentes entre as regiões (MEDEIROS; OLIVEIRA, Aug. 2014), decidiu-se trabalhar com uma área mais restrita. Foi selecionado portanto o Distrito Federal e suas proximidades.

Os municípios pertencentes ao entorno do território da capital da União possuem uma grande relação socioeconômica com o Distrito Federal. Mesmo que não seja considerada oficialmente uma região metropolitana, por conta dos locais participantes dessa relação não se encontrarem no mesmo estado, o mutualismo entre esses municípios torna inviável

que não se dê o tratamento de agrupamento dado comumente à qualquer outra região metropolitana (Codeplan, 2014). Uma característica que auxilia a comprovação dessa relação é a presença de uma considerável migração pendular que ocorre diariamente entre a região do entorno e o DF (QUEIROZ, 2016). Desta forma, para estudar fatores relacionados ao desempenho de estudantes do DF no Enem é importante considerar as escolas localizadas não só no Distrito Federal, mas também em seu entorno.

Define-se então a Área Metropolitana de Brasília, composta por 12 municípios goianos, sendo eles: Águas Lindas de Goiás, Alexânia, Cidade Ocidental, Cocalzinho de Goiás, Cristalina, Formosa, Luziânia, Novo Gama, Padre Bernardo, Planaltina de Goiás, Santo Antônio do Descoberto e Valparaíso de Goiás, além do município brasiliense, que pode ser repartido em 32 Regiões Administrativas (RA).

A ideia principal deste estudo é identificar fatores que influenciam nas notas obtidas pelos alunos concluintes do Ensino Médio em escolas da Área Metropolitana de Brasília no Enem do ano de 2018 e para isso será desenvolvido um Modelo Multinível.

2 Objetivos

2.1 Objetivo Geral

Desenvolver um Modelo Linear Multinível buscando fatores que influenciem as notas dos alunos da Área Metropolitana de Brasília concluintes do Ensino Médio no Enem em 2018.

2.2 Objetivos Específicos

- Estudo da técnica bem como sua aplicação utilizando *softwares* estatísticos como R e SAS.
- Análise de características dos alunos e suas escolas a fim de fomentar conjecturas já existentes.
- Estudo de fatores sociais, econômicos e demográficos que influenciem nos resultados obtidos na avaliação.

3 O Modelo Linear Multinível

Modelos multiníveis são utilizados para tratar de problemas em que existe mais de um nível de análise, uma espécie de hierarquia, na qual as unidades dentro do mesmo nível estão interligadas de alguma forma. Esses modelos podem ser aplicados em diversas áreas, como por exemplo na saúde, onde pode existir a hipótese de que, dentro de um hospital, pacientes atendidos pelo mesmo médico estejam interligados de certa maneira. Mesmo com aplicações em vários setores, o maior uso desse tipo de modelagem ocorre considerando dados educacionais, que é o ramo abordado por esse estudo.

No estudo apresentado, será considerada a análise do rendimento dos concluintes do Ensino Médio, supondo que alunos pertencentes a mesma escola estão interligados, devido à exposição ao mesmo "risco de aprendizado", fazendo com que seja necessária a aplicação de uma análise hierárquica, uma vez que a condição de independência entre as unidades observacionais é violada (HOX, 2010), conforme Figura 1.

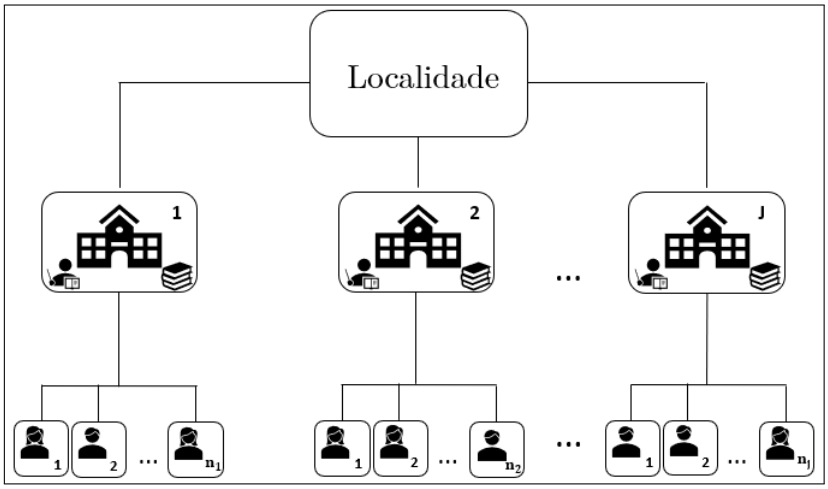


Figura 1 – Organização escola-aluno

3.1 O problema no uso do Modelo de Regressão Clássico

Se aplicado um modelo de Regressão Linear Clássico considerando p variáveis explicativas, da forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i, \tag{3.1}$$

em que:

(i) Y_i é a variável resposta observada para o i-ésimo aluno;

(ii) X_{1i} é o valor observado na primeira variável explicativa para o aluno i;

(iii) X_{pi} é o valor observado na p-ésima variável explicativa para o aluno i;

(iv) β_0 é o intercepto do modelo, representando o valor esperado da variável resposta (Y_i) quando todas as variáveis explicativas (X_i's) são iguais a zero;

(v) β_1 é o coeficiente de inclinação associado à primeira variável explicativa, representando o crescimento obtido em Y_i quando X_{1i} aumenta em uma unidade;

(vi) β_p é o coeficiente de inclinação associado à p-ésima variável explicativa, representando o crescimento obtido em Y_i quando X_{pi} aumenta em uma unidade;

(vii) ε_i é a componente de erro aleatório do modelo referente ao i-ésimo aluno.

Deve-se levar em consideração os seguintes pressupostos:

(i) $E(\varepsilon_i) = 0$;

(ii) $Var(\varepsilon_i) = \sigma^2$;

(iii) $Cov(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j$;

(iv) $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

Porém, para o caso em questão, a suposição (iii) de independência entre as observações é violada, uma vez que os elementos da unidade observacional (alunos) estão de certa forma ligados entre si. Segundo Laros e Marciano (2008), quanto maior for a correlação entre os indivíduos, maior a inadequação do modelo de regressão tradicional.

3.2 Descrição do Modelo Linear Multinível

Para que o problema apresentado anteriormente seja contornado, pode-se adotar uma Modelagem Linear Multinível considerando, para o problema apresentado, os alunos suscetíveis a concluir o Ensino Médio ao final do ano de 2018 como nível mais baixo de análise e as escolas de que são oriundos esses alunos como um nível mais acima, formando assim dois níveis hierárquicos.

Existem variáveis associadas a cada nível apresentado. Por exemplo, o sexo do aluno (X_1) ou a renda média familiar (X_2) são variáveis competentes ao nível do aluno. Já o número de alunos que cada escola possui (W) é uma variável medida no nível da escola. A variável resposta (Y) sempre é medida no nível mais baixo da análise e neste

caso será o desempenho médio obtido pelo concluinte do Ensino Médio no Enem de 2018. Considerando então essas variáveis, tem-se o seguinte modelo:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \varepsilon_{ij}, \tag{3.2}$$

em que o índice i vai de 1 até n_j para cada escola j, e representa o aluno dentro da j-ésima escola. Já o índice j varia de 1 até J, sendo que J representa o número de escolas no estudo.

Diferentemente da equação 3.1, na equação 3.2 os coeficientes da equação (β) são variáveis aleatórias que variam para cada escola. A interpretação desses valores se mantém praticamente igual à feita no Modelo de Regressão Clássico porém, agora o intercepto e as inclinações mudam conforme a unidade do nível mais elevado (escola) muda. Como suposições para o novo modelo, o termo ε_{ij} possui média 0 e variância σ_e^2.

A aleatoriedade dos coeficientes ocorre devido à sua origem, que parte das variáveis observadas no nível da escola, nesse exemplo representada pelo tamanho da escola (W). Portanto, tem-se:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \tag{3.3}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \tag{3.4}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}W_j + u_{2j}, \tag{3.5}$$

sendo:

(i) $u_{0j} \sim N(0, \sigma^2_{u_0})$

(ii) $u_{1j} \sim N(0, \sigma^2_{u_1})$

(iii) $u_{2j} \sim N(0, \sigma^2_{u_2})$

(iv) $Cov(u_{qj}, u_{lj}) = \sigma_{ql}, \quad \forall q \neq l.$

As parcelas u_{0j}, u_{1j} e u_{2j} são as componentes de erro associado ao intercepto e aos coeficientes do modelo e são comumente chamadas de erros do nível 2. Os termos γ são os novos coeficientes de regressão, agora ligados à variável W, medida no nível da escola.

Substituindo as equações 3.3, 3.4 e 3.5 na equação 3.2 é obtido o modelo completo de regressão multinível, também chamado de modelo saturado, dado por:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{01}W_j + \gamma_{11}X_{1ij}W_j + \gamma_{21}X_{2ij}W_j \\ & + u_{1j}X_{1ij} + u_{2j}X_{2ij} + u_{0j} + \varepsilon_{ij}, \end{aligned} \tag{3.6}$$

em que, na primeira linha $(\gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{01}W_j + \gamma_{11}X_{1ij}W_j + \gamma_{21}X_{2ij}W_j)$ tem-se a parte determinística do modelo e abaixo $(u_{1j}X_{1ij} + u_{2j}X_{2ij} + u_{0j} + \varepsilon_{ij})$ os fatores randômicos, com questões não captadas pelo modelo.

3.3 Estimação dos parâmetros

Nos Modelos Lineares Multiníveis três tipos de parâmetros podem ser estimados. Os efeitos fixos (γ_{0p}), os coeficientes de inclinação randômicos associados às variáveis do nível mais inferior (β_j) e os componentes de variância (σ_j) . Para obter essas estimativas o método mais comumente utilizado pela literatura estudada é o de Máxima Verossimilhança que tem como vantagem a produção de estimativas assintoticamente eficientes e consistentes. Duas funções podem ser usadas nesse método: A máxima verossimilhança completa (MVC) e a máxima verossimilhança restrita (MVR). Na primeira, são incluídos os coeficientes da regressão e os componentes da variância na função de verossimilhança de forma conjunta. Já na MVR, primeiro são estimados os componentes da variância e posteriormente os coeficientes de regressão.

A máxima verossimilhança completa não pode ser maximizada de forma analítica, por isso alguns processos iterativos são utilizados para encontrar o ponto de inflexão máximo da função.

3.4 Testes de hipóteses

Para cada parâmetro estimado via método de máxima verossimilhança, são geradas não só as estimativas mas também o erro padrão dessa estimativa. Essa medida pode ser utilizada para a criação de testes de significâncias da forma:

$$t = \frac{\text{estimativa} - \text{parâmetro}}{\text{erro padrão da estimativa}}, \tag{3.7}$$

em que t está associado à distribuição t-Student com os graus de liberdade definidos pelo método de Satterthwaite.

Por exemplo, testando se o *p-ésimo* efeito fixo é significante para o modelo, ou seja, considerando a hipótese nula $\hat{\gamma}_{0p} = 0$ tem-se então:

$$t = \frac{\hat{\gamma}_{0p}}{EP(\hat{\gamma}_{0p})}, \tag{3.8}$$

seguindo uma distribuição t-Student com os graus de liberdade aproximados pelo método de Satterthwaite.

3.5 Estratégia de análise

Seja p o número de variáveis explicativas incluídas no modelo no nível mais baixo e q o número de variáveis explicativas no nível mais alto, tem-se então o seguinte modelo multinível composto por dois níveis:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}W_{qj} + \gamma_{pq}X_{pij}W_{qj} + u_{pj}X_{pij} + u_{0j} + \varepsilon_{ij}. \tag{3.9}$$

Alguns passos podem ser seguidos para realizar a modelagem ideal quando trabalha-se com Modelos de Regressão Multinível. Recomenda-se iniciar a modelagem da forma mais simples possível(HOX, 2010), portanto:

Passo 1

Ajusta-se o modelo sem variáveis explicativas, apenas com o intercepto, da forma:

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij}, \tag{3.10}$$

sendo γ_{00} o intercepto do modelo, u_{0j} e ε_{ij} as componentes de aleatoriedade no nível dos grupos e no nível dos indivíduos, respectivamente.

Esse modelo é bastante útil para estimar o coeficiente de correlação intraclasse, que tem como objetivo verificar a proporção de variação que pode ser explicada pelo agrupamento aplicado (no caso em estudo, o quanto que a variação entre as escolas pode explicar o desempenho dos alunos). Esse coeficiente é dado por:

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_e}. \tag{3.11}$$

Passo 2

Analisa-se um modelo apenas com as variáveis explicativas referentes ao primeiro nível da análise, setando os componentes da variância dos coeficientes de inclinação em zero. O modelo descrito pode ser apresentado como:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + u_{0j} + \varepsilon_{ij}. \tag{3.12}$$

Na equação 3.12, X_{pij} é uma matriz com as p variáveis referentes ao primeiro nível da análise. Nesse passo, pode-se testar a significância de cada variável explicativa do nível inferior e mudanças quanto às variáveis inseridas podem ser feitas.

Passo 3

Agora, são adicionadas à equação 3.12 as variáveis no nível superior, ou seja, referentes a escola. Desta forma:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}W_{qj} + u_{0j} + \varepsilon_{ij}, \tag{3.13}$$

sendo W_{qj} uma matriz com as q variáveis referentes ao nível mais alto da análise.

Os modelos apresentados nos passos 2 e 3 podem ser chamados de modelos de componente de variância uma vez que estes decompõem a variância do intercepto em diferentes componentes, para cada nível hierárquico.

Passo 4

Analisa-se se quaisquer coeficientes de regressão do nível 1 tem uma componente significativa de variância entre os grupos do nível 2. Dado pela equação:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}W_{qj} + u_{pj}X_{pij} + u_{0j} + \varepsilon_{ij}, \tag{3.14}$$

em que u_{pj} são os resíduos do segundo nível dos coeficientes de inclinação das variáveis do nível micro (X_{pij}).

Passo 5

Por fim, serão adicionadas à equação 3.14 as interações entre as variáveis do nível macro e as variáveis que obtiveram uma variabilidade significativa, verificadas no passo anterior. Com isso, é obtido o modelo completo:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}W_{qj} + u_{pj}X_{pij} + \gamma_{pq}X_{pij}W_{qj} + u_{0j} + \varepsilon_{ij}. \tag{3.15}$$

3.6 Qualidade do Modelo

Algumas medidas podem ser utilizadas para aferir a qualidade do modelo e formular comparações entre modelos propostos durante o estudo. Uma dessas medidas é o *deviance*, valor baseado na função de verossimilhança (L), dado por:

$$d = -2 \times \ln(L). \tag{3.16}$$

De forma geral, quanto menor o valor de d melhor será o ajuste do modelo.

O *deviance* também pode ser utilizado para comparar dois modelos encaixados, que são casos onde um modelo mais específico pode ser derivado de um modelo mais geral apenas removendo alguns dos parâmetros do modelo mais completo. Sendo este o caso, a diferença entre os *deviances* possui uma distribuição Qui-quadrado com os graus

de liberdade dados pela diferença do número de parâmetros estimados. A equação 3.17 apresenta a estatística do teste:

$$D = d_{ms} - d_{mg}, \tag{3.17}$$

sendo que d_{mg} é o *deviance* calculado para o modelo geral e d_{ms} o *deviance* do modelo simples.

Caso os modelos não sejam encaixados, essa medida não pode ser utilizada e outros métodos devem ser aplicados. Um exemplo é o *Akaike Information Criterion* (AIC) que usa o *deviance* como base mas, de certa forma, penaliza de acordo com o número de parâmetros (k) estimados pelo modelo.

$$AIC = d + 2k. \tag{3.18}$$

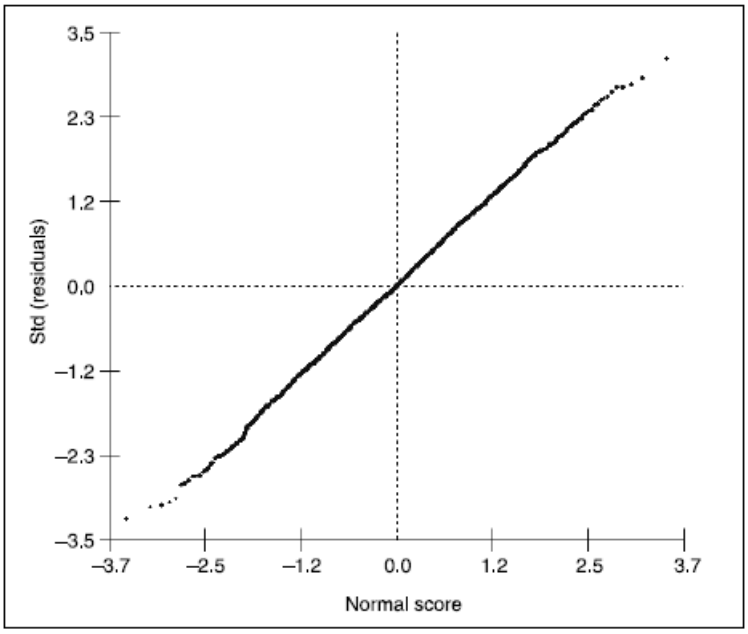
Para fazer comparações utilizando essas medidas, tem-se como pressuposto que o ajuste realizado parte do mesmo banco de dados, com o mesmo método de estimação. Assim como no *deviance*, os modelos com menor AIC têm um melhor ajuste, tendendo assim a ser o modelo selecionado.

3.7 Diagnóstico do Modelo

Após ajustar o modelo, é necessário verificar se as suposições de normalidade, linearidade e homocedasticidade são satisfeitas. Diferentemente do Modelo de Regressão Clássico, num Modelo Linear Multinível existem diversos resíduos, sendo um para cada componente de efeito aleatório do modelo. A partir desses resíduos é possível verificar as hipóteses assumidas tanto graficamente quanto numericamente, por meio de testes estatísticos.

Um gráfico muito utilizado para verificar a suposição de normalidade dos resíduos é o Gráfico de Probabilidade Normal, também conhecido por QQ-plot dado pela Figura 2.

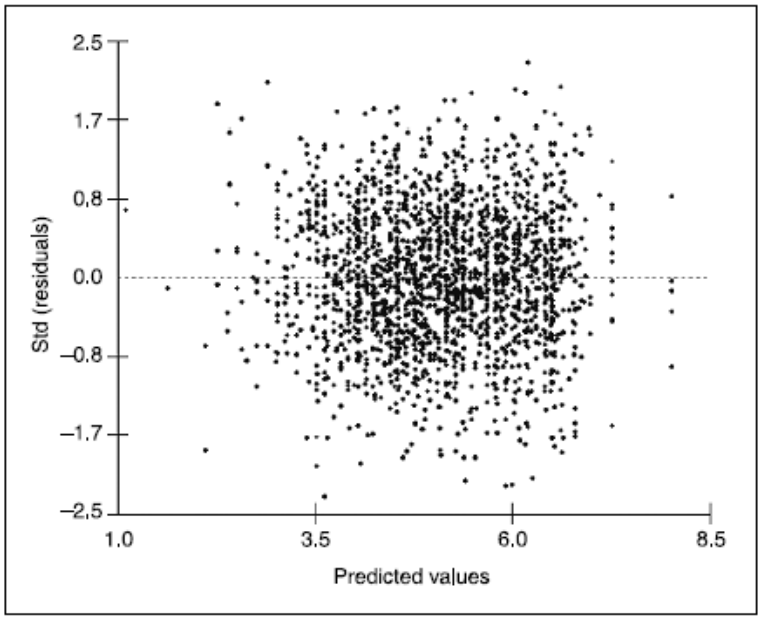
Neste gráfico, são plotados os quantis empíricos do resíduos *versus* os quantis teóricos da distribuição Normal. Se os pontos plotados formam uma linha diagonal cruzando o plano, significa que os resíduos estão bem ajustados à distribuição Normal. Além da visualização gráfica, testes como Shapiro-Wilk podem ser realizados para verificar a adequabilidade dos resíduos à distribuição Normal. Caso os resíduos não se adéquem à distribuição Normal, algumas transformações podem ser realizadas para que a análise prossiga.



Fonte: Hox (2002).

Figura 2 – Gráfico de Probabilidade Normal

A fim de verificar a homocedasticidade dos resíduos, condição que indica a variabilidade dos resíduos de forma constante, pode-se analisar o gráfico de resíduos padronizados *versus* valores preditos (Figura 3).



Fonte: Hox (2002).

Figura 3 – Resíduos padronizados × Valores Preditos

Os pontos apresentados no gráfico não devem conter nenhum padrão visível, estando espalhados aleatoriamente ao redor da reta vertical que traça o gráfico em $y = 0$.

Aqui também é possível a utilização de testes estatísticos para verificar a hipótese de homocedasticidade. Os teste de Hartley e Levene são bastante utilizados para verificar essa hipótese sendo o primeiro recomendado apenas para quando os resíduos se adéquam à distribuição Normal (KUTNER et al., 2005).

4 Metodologia

Neste capítulo são apresentadas as bases de dados utilizadas para obtenção de informações, bem como o processo de criação de indicadores para auxiliar o estudo.

4.1 Os dados

4.1.1 O Exame Nacional do Ensino Médio

Instituído pela portaria MEC nº 438, de 28 de maio de 1998, o Exame Nacional do Ensino Médio, popularmente conhecido por sua sigla, Enem, foi criado com o intuito principal de avaliação de políticas públicas no âmbito da educação básica, fornecendo também ao cidadão que deseja participar do teste, uma vez que este não possui caráter obrigatório, um parâmetro para auto-avaliação.

Planejado e operacionalizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), o Enem é uma avaliação aplicada de forma descentralizada, em diversos municípios. Qualquer cidadão pode realizar o teste, havendo uma restrição para jovens menores de 18 anos que não possuem chance de concluir o Ensino Médio ao final do ano de aplicação do Enem. Para estes, a inscrição deve ser feita aderindo a condição de "treineiro", sendo o resultado obtido útil apenas para a auto-avaliação do estudante.

O Exame passou por duas significativas reformulações: Uma em 2004 quando foi implementado o Programa Universidade para Todos (ProUni) que provê bolsas de estudos para alunos com um bom *score* no teste; e a segunda no ano de 2009 com a implementação do Sistema de Seleção Unificada (SISU), que insere alunos no Ensino Superior Público do país.

Atualmente, o Enem é realizado em dois fins de semana, sendo o primeiro contendo as provas de Linguagens, Códigos e suas Tecnologias, Redação e Ciências Humanas e suas Tecnologias. Já no segundo dia, são aplicadas as provas de Ciências da Natureza e suas Tecnologias e Matemática e suas Tecnologias. A variável resposta do estudo realizado consiste na média das notas de todas as competências citadas para cada indivíduo.

Serão considerados apenas os alunos do ensino regular de escolas federais, estaduais e privadas da Área Metropolitana de Brasília com conclusão do ensino médio prevista para o ano de 2018, presentes nos dois dias de prova, contabilizando assim 23.741 candidatos às vagas no Ensino Superior brasileiro.

Juntamente com o desempenho que cada inscrito no Exame obteve, são disponibi-

lizados os dados referentes à situação socioeconômica do estudante, coletados por meio de questionário realizado no momento da inscrição. Para fins de complementação das informações, principalmente quanto às escolas consideradas, o Censo Escolar de 2018 será utilizado como base secundária.

4.1.2 Censo Escolar

Considerado o principal instrumento de coleta de dados da educação básica do país, o Censo Escolar é coordenado pelo INEP, assim como o Enem, mas realizado de forma colaborativa com as secretarias municipais e estaduais de educação, havendo a participação de todas as escolas do Brasil, sendo estas públicas ou privadas (INEP, 2019).

O Censo Escolar tem caráter declaratório, sendo preenchido pelas escolas em duas etapas. A primeira consiste na coleta de informações referentes à própria escola juntamente com dados sobre professores, turmas, gestores e alunos. Neste etapa, características quanto a estrutura da escola que são de fundamental importância para o estudo aqui realizado, são coletadas. Já na segunda etapa, é informada a situação do aluno, considerando seu rendimento ao final do ano letivo.

Assim como o Enem, o Censo escolar serve como ferramenta para fomento à políticas públicas no âmbito educacional, servindo como fonte para uma série de indicadores referentes a educação básica do país, como o Índice de Desenvolvimento da Educação Básica (Ideb).

4.1.3 Localização das escolas

Em ambas as bases de dados que se trabalhou, o menor nível de agrupamento referente à localidade das escolas é o de município. Como o estudo em questão se refere à Área Metropolitana de Brasília, viu-se necessária a presença da informação da Região Administrativa que a escola se situa, considerando o Distrito Federal. Para isso, foi utilizado o pacote *"RSelenium"* do software R, que funciona como uma ferramenta de *web scraping*, acessando o navegador de internet e coletando dados da página solicitada de forma robotizada. Com a ferramenta, foram obtidos o endereço da escola (com a RA embutida na informação), e as coordenadas geográficas de cada entidade, podendo assim observar como a variável resposta está distribuída no espaço do Distrito Federal e entorno.

As Regiões Administrativas podem ser classificadas em 4 grupos distintos, de acordo com os padrões de rendimento familiar médio da localidade. Segundo a Codeplan, as classes são definidas conforme Tabela 1 que também apresenta um 5º grupo, com os municípios do entorno do Distrito Federal, já descritos anteriormente. Vale ressaltar que nem todas as Regiões Administrativas possuem escolas no estudo realizado, uma vez que foram considerados apenas colégios com suporte ao ensino médio regular.

Tabela 1 – Área Metropolitana de Brasília - Agrupamento de RA's

	Região Administrativa	**Renda Média**
Grupo 1 (alta renda)	Brasília Jardim Botânico Lago Norte Lago Sul Park Way Sudoeste Octogonal	R$ 15.622,00
Grupo 2 (média-alta renda)	Águas Claras Candangolândia Cruzeiro Gama Guará Núcleo Bandeirante Sobradinho Sobradinho II Taguatinga Vicente Pires	R$ 7.266,00
Grupo 3 (média-baixa renda)	Brazlândia Ceilândia Planaltina Riacho Fundo Riacho Fundo II SIA Samambaia Santa Maria São Sebastião	R$ 3.101,00
Grupo 4 (baixa renda)	Fercal Itapoã Paranoá Recanto das Emas SCIA-Estrutural Varjão	R$ 2.472,00
Entorno	Águas Lindas de Goiás Alexânia Cidade Ocidental Cocalzinho de Goiás Cristalina Formosa Luziânia Novo Gama Padre Bernardo Planaltina Santo Antônio do Descoberto Valparaíso de Goiás	-

Fonte: Codeplan, 2018.

4.2 Indicadores

A fim de trazer ao estudo uma maior quantidade de informações, buscou-se a obtenção de medidas que pudessem auxiliar a predição da nota obtida pelos alunos da AMB no Enem 2018. Com isso, alguns indicadores foram desenvolvidos utilizando como base algumas variáveis contidas nos conjuntos de dados utilizados. A ideia de construir índices para o estudo em questão parte justamente da necessidade de redução do espaço paramétrico, transformando k variáveis em apenas uma, trazendo uma melhor interpretação dos fatores que explicam o desempenho do aluno.

Nos estudos relativos ao âmbito educacional, deve-se considerar as características socioeconômicas dos alunos uma vez que esses fatores estão fortemente associados ao desempenho escolar (SOARES, 2004). Com isso, viu-se a necessidade da utilização de algum índice que levasse em consideração tais características, podendo agrupar a informação de diversas variáveis em apenas um valor, ponderando cada variável de acordo com a importância da mesma para a definição de diferentes níveis socioeconômicos. A técnica inicialmente utilizada foi a Análise Fatorial, método que busca descrever a relação de covariância entre diversas variáveis em função de termos não observáveis chamados de fatores (JOHNSON; WICHERN, 2007).

O modelo fatorial possui a seguinte forma geral:

$$X_i = \mu_i + l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \epsilon_i, \tag{4.1}$$

de modo que, X_i representa o valor da i-ésima variável e μ_i a sua média, F_j é o j-ésimo fator aleatório, l_{ij} são os *loadings*, valores que definem o grau de relacionamento linear entre X_i e F_j, e ϵ_i o erro aleatório do modelo.

Para estimar os *loadings* foi utilizado o Método de Componentes Principais e para uma melhor visualização os fatores foram rotacionados com a técnica Varimax. O número de fatores retidos no estudo pode ser selecionado segundo um *"Scree plot"*, gráfico dos autovalores (λ_i) obtidos na estimação dos *loadings*, em ordem decrescente. Quando a linha apresentada no gráfico começa a se manter constante, tem-se um indicação do número de fatores a se selecionar, sendo o restante destes descartado para o erro do modelo (ϵ_i).

Analisando questões socioeconômicas, foram utilizadas as variáveis descritas no Apêndice B, e três fatores foram retidos, de acordo com a análise do *Scree-Plot*. Quando observados os três fatores selecionados, não parece haver uma ligação entre as variáveis de maior importância para cada fator. Mesmo assim, os três fatores foram testados no modelo multinível separadamente, e nenhum deles foi significante para o modelo.

Sendo assim, buscou-se uma alternativa para a criação de um indicador de nível socioeconômico do estudante e para isso foi utilizado o Critério de Classificação Econômica

Brasil (CCEB).

O CCEB é um indicador desenvolvido pela Associação Brasileira de Empresas de Pesquisa (ABEP) para a definição de segmentação por poder aquisitivo. O índice original considera variáveis relacionadas ao poder aquisitivo da família, o grau de instrução do chefe da casa e o acesso aos serviços públicos (no caso, água encanada e pavimentação da rua de moradia).

A tabela de variáveis consideradas pelo índice, juntamente com seu sistema de pontuação e as alterações necessárias no CCEB original estão apresentadas no Apêndice D. A soma dessas pontuações fornece ao integrante da família sua classe social, de acordo com regras de corte que aqui não serão tratadas, uma vez que o *score* obtido tem um maior valor para o estudo do que a própria classe social.

Após estudos com essa técnica, viu-se que para o modelo em questão, o Critério de Classificação Econômica Brasil (CCEB), citado anteriormente, teve um melhor desempenho para fins do modelo desenvolvido, conforme seção 3.6.

Para a análise de componentes presentes na escolas, indicando um nível de infraestrutura da entidade, também foi utilizada a Análise Fatorial. Conforme análise dos autovalores, cinco fatores foram retidos para a análise. Agora, os fatores parecem possuir algum significado dentro do problema apresentado (Apêndice D). O primeiro fator entrega pesos maiores para itens referentes à estrutura básica das escolas, como possuir água encanada, banheiro, sala de diretoria e sala dos professores. O segundo fator indica a presença de equipamentos eletrônicos na escola, dando uma maior importância para a posse de impressora, copiadora, DVD e televisão. Já o terceiro fator traz um maior peso para características estruturais mais avançadas, como laboratório de ciências e quadra de esportes. O quarto fator, considera a presença de uma estrutura para estudantes portadores de necessidades especiais (PNE). A última dessas quantidades não trouxe uma informação significante quanto alguma característica de infraestrutura escolar. Nenhum dos fatores descritos foi significativo para o modelo, de acordo com o teste de razão de verossimilhanças.

4.3 Variáveis explicativas do modelo

Foram consideradas para o modelo multinível as variáveis apresentadas na Tabela 2, que contêm informações referentes aos estudantes (Nível 1), e na Tabela 3, com variáveis da escola (Nível 2). Tais variáveis foram utilizadas para predizer a média dentre as notas obtidas em cada competência do Enem 2018, variável resposta do estudo que será tratada como “desempenho”, para a simplicidade da análise.

Considerando a análise do comportamento da variável resposta nas diferentes

categorias das variáveis explicativas alguns níveis foram agregados, permitindo assim a redução no número de parâmetros do modelo a serem estimados.

Tabela 2 – Variáveis do estudante

Nome	Descrição	Tipo
NU_IDADE	Idade	Discreta
SEXO_M	Sexo: Masculino	Indicadora
COR_RACA_PPI	Cor/raça: Preta, parda ou indígena	Indicadora
OCUP_G4	Ocupação: Grupo 4*	Indicadora
OCUP_G5	Ocupação: Grupo 5*	Indicadora
CCEB	Critério de Classificação Econômica Brasil	Contínua

*Grupo ocupacional do responsável melhor empregado. Vide apêndice A.

Tabela 3 – Variáveis da escola

Nome	Descrição	Tipo
RA_GRUPO_2	RA: Grupo 2*	Indicadora
RA_GRUPO_3	RA: Grupo 3*	Indicadora
RA_GRUPO_4	RA: Grupo 4*	Indicadora
RA_ENTORNO	Entorno	Indicadora
DEP_FEDERAL	Dependência Administrativa: Federal	Indicadora
DEP_PRIVADA	Dependência Administrativa: Privada	Indicadora
IND_ATU_TER	Média de alunos por turma - 3º ano	Contínua
IND_HAD_TER	Média de horas-aula diária - 3º ano	Contínua
IND_TDI_TER	Taxa de Distorção Idade-Série - 3º Ano	Contínua
IND_DSU_EM	Proporção de Docentes com Curso Superior**	Contínua

*Grupos de Regiões Administrativas definidos na Seção 4.1.3.
**Docentes do Ensino Médio.

Os níveis que não aparecem nas variáveis indicadoras apresentadas nas tabelas são tratadas como valores de referência para a modelagem. Um exemplo é a variável "Sexo", que na Tabela 2 aparece apenas como "SEXO_M" (Sexo: Masculino). Na modelagem, essa variável indica o ganho (ou perda) que se obtém na nota quando o estudante é do sexo masculino em relação a alunos do sexo feminino.

5 Resultados

Para o estudo realizado, foram considerados alunos do ensino regular da Área Metropolitana de Brasília (AMB), concluintes do ensino médio no ano de 2018, de escolas Federais, Estaduais e Privadas com mais de 10 alunos inscritos no Enem. Além disso, removeu-se da base de dados estudantes que não tinham todas as informações necessárias presentes no banco de dados e por isso, consequentemente, candidatos que faltaram a algum dia de prova do Enem. No total, tem-se **23.741 estudantes** de **318 escolas** participantes do estudo.

Inicialmente, foi realizada uma análise exploratória da variável resposta do estudo. Nas seções seguintes observou-se as variáveis explicativas do modelo, referidas aos estudantes e às escolas, e sua relação com o desempenho obtido.

Vale ressaltar que outras variáveis foram analisadas durante o estudo, mas estas não se mostraram significantes para a definição do desempenho do aluno no exame.

5.1 Análise do desempenho

Desde o ano de 2009, o Exame Nacional do Ensino Médio tem o formato que se é conhecido hoje, uma prova com uma redação e 180 questões objetivas, igualmente divididas em 4 competências (Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática). A fim de observar a evolução da performance dos estudantes nessas questões, foi realizada uma análise histórica dessas notas, buscando entender o comportamento do desempenho dos alunos desde 2009[1]. O ano de 2014 foi eliminado da análise por questões de inconsistência nos resultados obtidos.

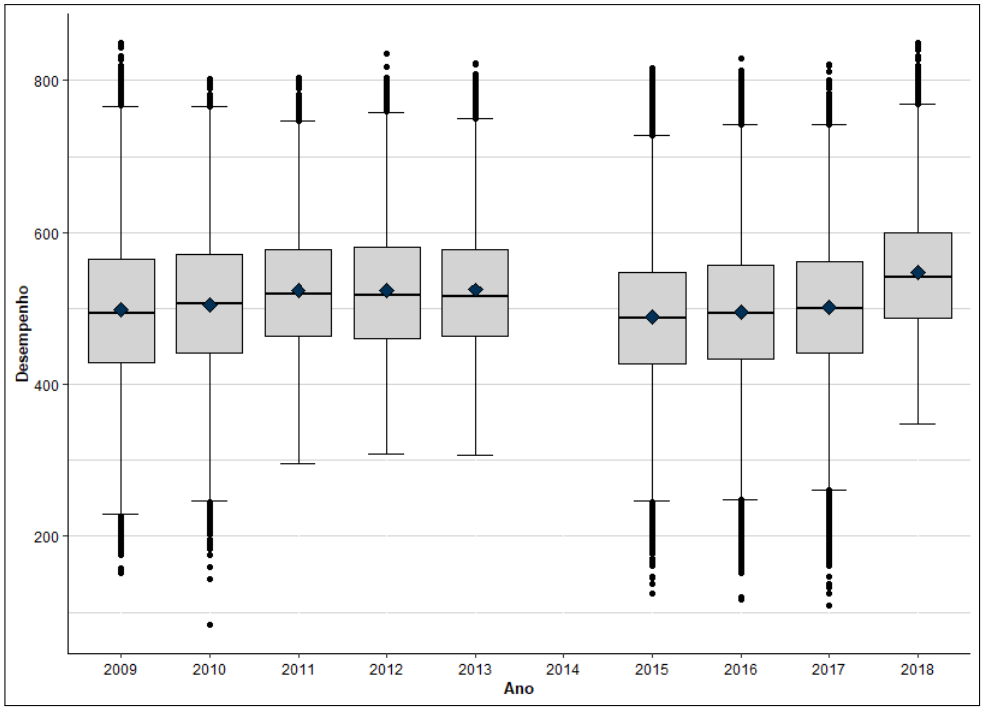
Nota-se uma queda do desempenho no ano de 2015 mas após isso, a nota média dos alunos vem em constante crescimento. Percebe-se também um aumento no número de inscritos no teste na região destacada até o ano de 2016, havendo após esse período uma pequena queda nesse contingente. Dentre os anos analisados, os alunos que realizaram o teste em 2018 obtiveram a melhor média de desempenho nas 5 competências do Enem.

[1] Para a série temporal, não foi considerada a exclusão de estudantes advindos de escolas municipais e escolas com menos de 10 alunos inscritos no Enem.

Tabela 4 – Análise do desempenho entre 2009 e 2018

Ano	Média	Desvio Padrão	Coeficiente de Variação	Número de Alunos
2009	497,68	102,1	20,51%	14.183
2010	504,88	96,1	19,03%	15.998
2011	523,22	80,8	15,44%	19.829
2012	523,25	83,2	15,90%	22.787
2013	524,28	82,6	15,76%	25.138
2014	-	-	-	-
2015	488,78	94,4	19,31%	29.961
2016	495,75	95,9	19,35%	30.663
2017	501,07	92,4	18,43%	29.002
2018	547,61	80,2	14,65%	25.912

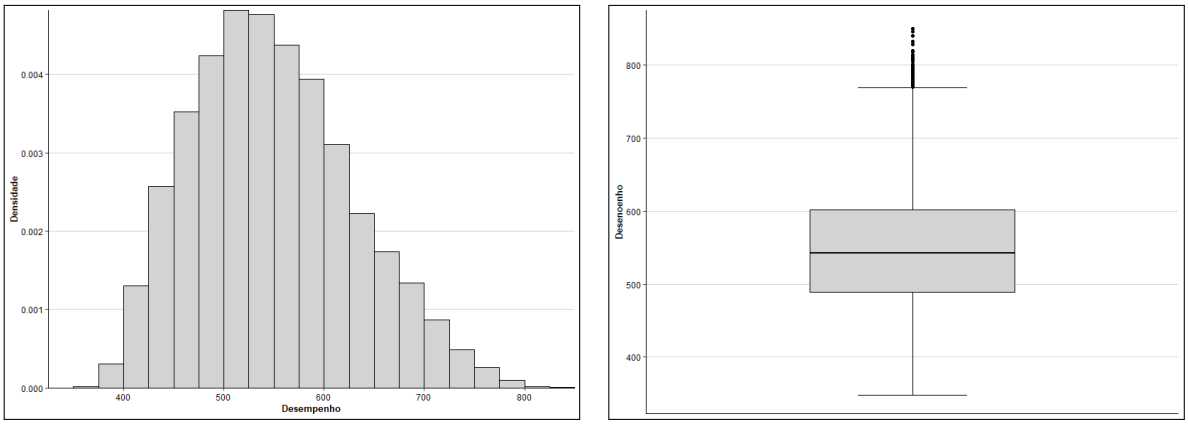
Fonte: Enem/INEP.



Fonte: Enem/INEP.

Figura 4 – Distribuição do desempenho entre 2009 e 2018

Focando no ano de 2018, período utilizado para a modelagem, tem-se a seguinte distribuição do desempenho médio dos estudantes nas 5 competências do Enem:



(a) Histograma do desempenho (b) Box plot do desempenho

Figura 5 – Distribuição do desempenho - 2018

Tabela 5 – Desempenho do estudante - 2018

	Desempenho
Média	548,99
Mediana	541,94
Desvio padrão	80,12
Coeficiente de Variação	14,59%
Mínimo	347,64
Máximo	849,66

Nota-se uma distribuição assimétrica à direita, indicando um maior número de alunos com desempenho mais inferior e alguns estudantes destoantes da grande massa obtendo maiores notas.

5.2 Perfil dos estudantes

A seleção de variáveis para o modelo partiu de uma estatística exploratória buscando a variabilidade da variável resposta segundo determinada categoria. Além disso, foram realizados testes de comparação de médias a fim de identificar variáveis que traziam uma diferença significativa para o desempenho do aluno.

Com isso, iniciou-se a análise pelas variáveis do primeiro nível, referentes aos estudantes, começando pelo sexo. Na Tabela 6, é possível ver uma quantidade de mulheres um pouco maior (57,8%) porém, com um desempenho médio inferior ao dos homens (557,4 contra 542,8).

Tabela 6 – Distribuição do desempenho do estudante por sexo - 2018

	Sexo	
	Masculino	Feminino
Média	557,4	542,8
Mediana	552,0	534,9
Desvio Padrão	81,5	78,5
N	10.019 (42,2%)	13.772 (57,8%)

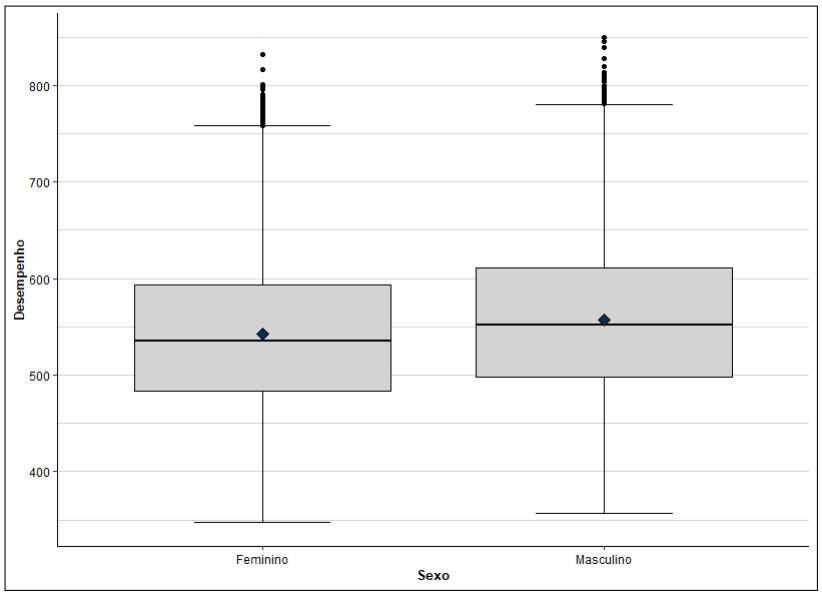


Figura 6 – Desempenho do estudante por sexo - 2018

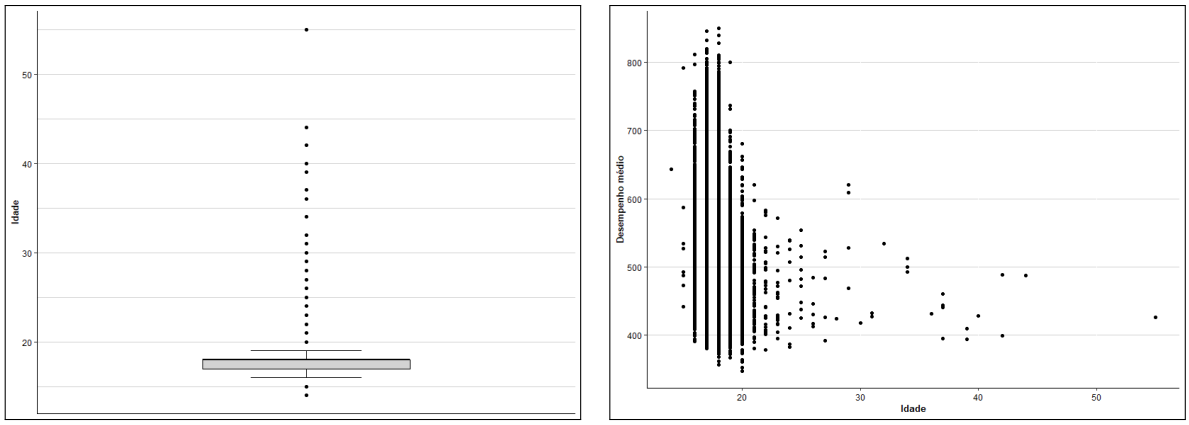
Outro fator que pode vir a interessar na análise é a idade do aluno. Uma vez que o estudo realizado tem como foco os estudantes concluintes do ensino médio, alunos mais velhos podem representar pessoas que reprovaram algum ano em sua escola, estando assim, fora da idade recomendada para o 3º ano do ensino médio.

Tabela 7 – Distribuição da Idade

	Idade
Média	17,67
Mediana	18
Desvio padrão	1,03
Coeficiente de Variação	17,22%
Mínimo	14
Máximo	55

Conforme a Tabela 7, foram encontrados no estudo alunos com idade entre 14 e 55 anos. A média desta variável é de 17,7 anos, apresentando assim um indício de que, mesmo aparecendo no estudo alunos com idade avançada, estes são exceção ao centro da distribuição. Com a Figura 7a, nota-se que grande parte dos alunos se agrega com idade abaixo dos vinte anos.

Quando a variável é analisada juntamente com o desempenho obtido, é perceptível uma queda na nota obtida quando a idade é mais avançada.



(a) Box plot da idade dos estudantes - 2018 (b) Desempenho por idade - 2018

Figura 7 – Idade dos estudantes - 2018

Uma outra variável considerada no estudo foi a raça/cor do estudante. Estudantes pretos, pardos ou indígenas foram agregados numa mesma classe chamada de PPI e foi analisado seu resultado comparado à pessoas "Não PPI" (Brancos e amarelos). Observa-se na Tabela 8 e no gráfico 8 um melhor resultado para alunos desta última classe.

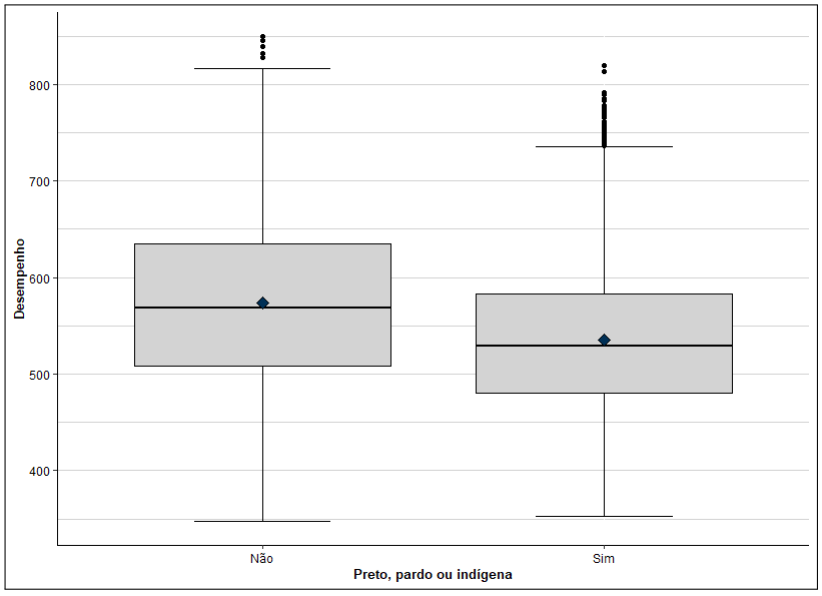


Figura 8 – Desempenho do estudante por raça/cor - 2018

Tabela 8 – Distribuição do desempenho do estudante por raça/cor - 2018

	Raça/cor	
	PPI	Não PPI
Média	534,9	573,4
Mediana	529,34	568,8
Desvio Padrão	72,9	85,9
N	15.051 (63,4%)	8.690 (36,6%)

Em seguida, foi observada a ocupação do responsável melhor empregado, onde categorias de trabalho foram criadas conforme critério descrito no Apêncide A. Vale ressaltar que essa variável foi analisada juntamente com a escolaridade do responsável melhor instruído e a renda familiar e percebeu-se uma correlação forte, podendo trazer problemas de multicolinearidade para o modelo. Dentre as três variáveis, a que trazia ao modelo um menor valor de AIC era a renda familiar porém, analisou-se a qualidade da informação e, uma vez que é mais fácil um aluno de nível médio informar com precisão a profissão do responsável do que a renda da sua família, foi considerada a variável de ocupação. Com isso, a distribuição dessa informação é dada a seguir:

Tabela 9 – Desempenho do estudante por ocupação do responsável - 2018

	Ocupação		
	Grupo 1, 2 ou 3	Grupo 4	Grupo 5
Média	516.0	566.5	626.5
Mediana	511.7	565.3	633.3
Desvio Padrão	63.5	74.3	82.8
N	12.039 (50,7%)	8.505 (35,8%)	3.197 (13,5%)

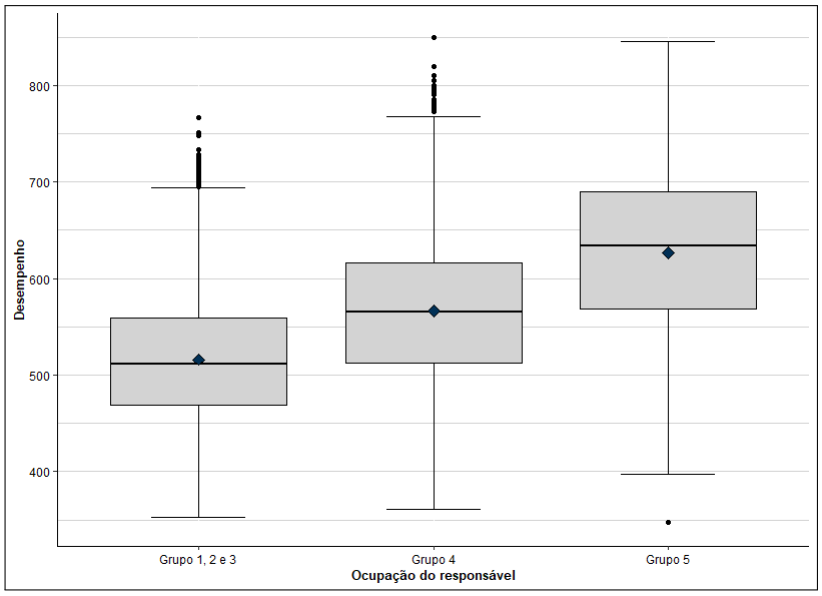
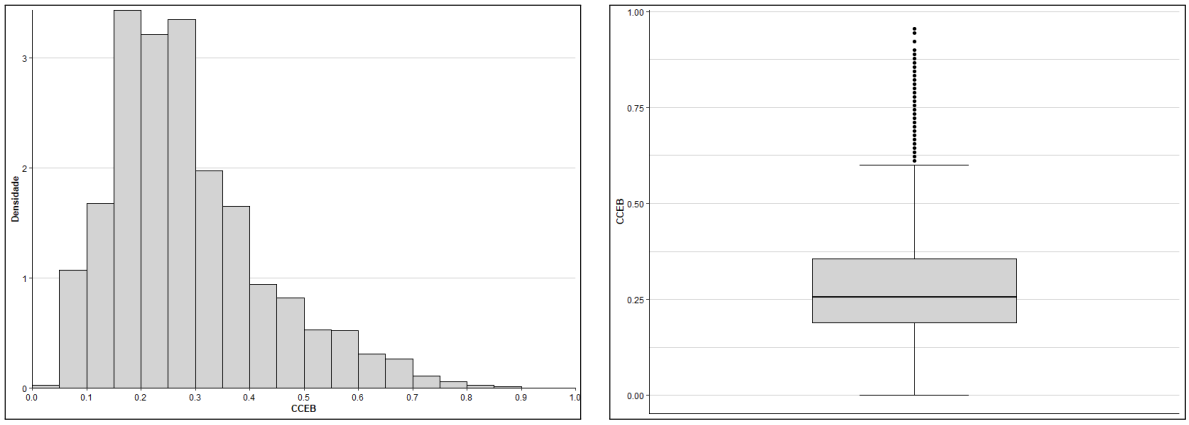


Figura 9 – Desempenho do estudante por ocupação do responsável - 2018

Alunos com responsáveis participantes de um grupo ocupacional que necessita de uma maior formação (Grupo 5, onde estão inclusos médicos, engenheiros, etc.) tem uma distribuição da nota mais elevada do que estudantes com responsáveis em condições de trabalho que exigem maior escolaridade, conforme Figura 9.

Por último, foi observada a distribuição do indicador socieconômico criado segundo regra da ABEP, conforme seção 4.2. Assim, segue a distribuição do CCEB modificado, quesito que foi utilizado no estudo como uma variável explicativa para o modelo proposto.

Conforme a Figura 10, percebe-se uma maior frequência entre 0,15 e 0,3 pontos. A distribuição é assimétrica à direita, indicando uma maior concentração de estudantes com índice socioeconômico mais baixo e alguns discentes com essa taxa mais elevada, saindo do padrão observado.



(a) Histograma do CCEB - 2018 (b) Box plot do CCEB - 2018

Figura 10 – Distribuição do CCEB - 2018

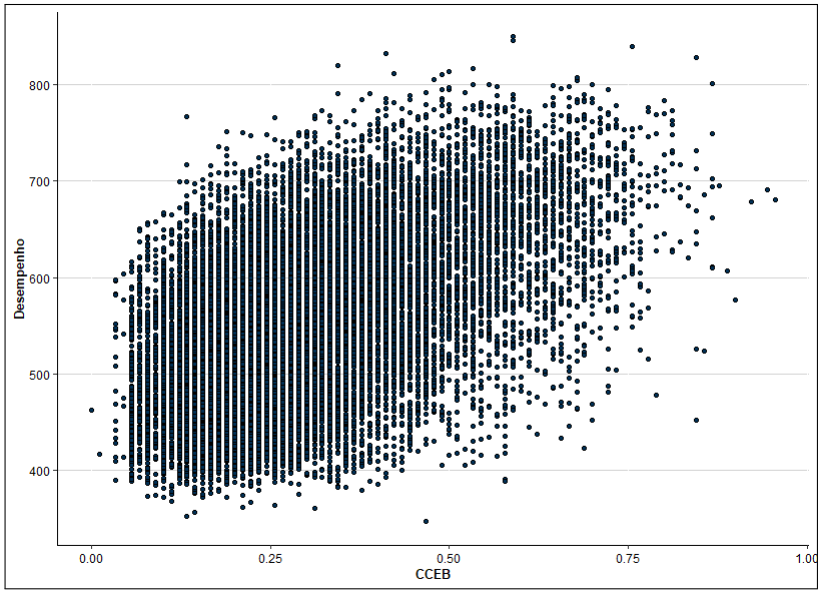


Figura 11 – Desempenho do estudante por CCEB - 2018

Quando analisado conjuntamente com o desempenho obtido, percebe um crescimento da nota quando o CCEB cresce, indicado que, de forma geral, alunos com uma melhor condição socieconômica tem um melhor desempenho no Enem.

5.3 Perfil das escolas

O segundo nível da análise contem às variáveis referentes à escola do estudante, em hipótese, fator considerado importante para o desempenho do aluno, a ser confirmado pela modelagem.

A localização da escola é o primeiro fator aqui observado. Nos municípios da AMB, a distribuição de escolas e alunos é apresentada na Tabela 10.

Tabela 10 – Distribuição das escolas e alunos na AMB - 2018

	Escolas		**Alunos**	
Cidade	N	(%)	N	(%)
Águas Lindas de Goiás	23	7,23	733	3,09
Alexânia	3	0,94	96	0,40
Brasília	200	62,89	19.108	80,49
Cidade Ocidental	4	1,26	217	0,91
Cocalzinho de Goiás	2	0,63	87	0,37
Cristalina	6	1,89	234	0,99
Formosa	17	5,35	637	2,68
Luziânia	23	7,23	1.015	4,28
Novo Gama	7	2,20	219	0,92
Padre Bernardo	2	0,63	83	0,35
Planaltina de Goiás	9	2,83	366	1,54
Santo Antônio do Descoberto	4	1,26	199	0,84
Valparaíso de Goiás	18	5,66	747	3,15
Total	318	100	24.077	100

A cidade com maior frequência de escolas e alunos é Brasília com 200 escolas alocando 80,49% dos estudantes. Os municípios seguintes são Luziânia e Águas Lindas de Goiás, com apenas 23 escolas.

Considerando agora como "desempenho da escola" a média das notas dos estudantes que participaram do Enem e concluiriam em 2018 o ensino médio naquela entidade, foi obtido um comportamento onde, no geral, escolas da zona mais central da AMB possuíam maiores médias (como pode ser visto na Figura 12, onde os pontos mais escuros denotam maiores notas) e escolas mais periféricas obtiveram menores notas, em média.

Quando o mapa é aproximado para a região do Distrito Federal, conforme Figura 13, nota-se uma coloração mais escura na zona mais próxima ao Plano Piloto, centro político de Brasília, indicando assim melhores notas nessa região.

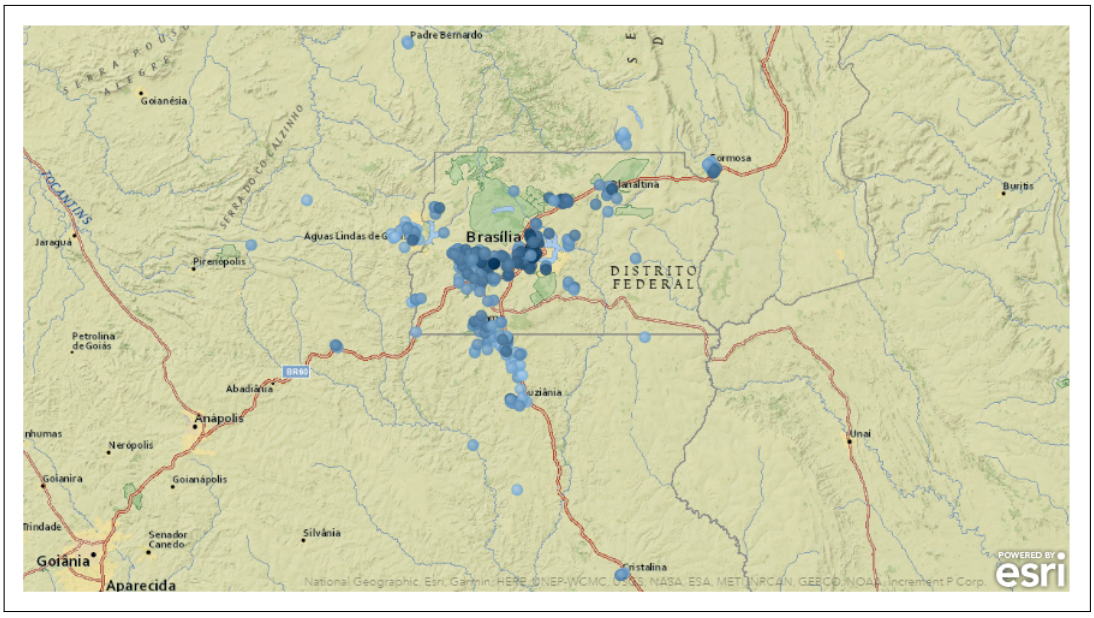


Figura 12 – Mapa de desempenho da escola - AMB - 2018

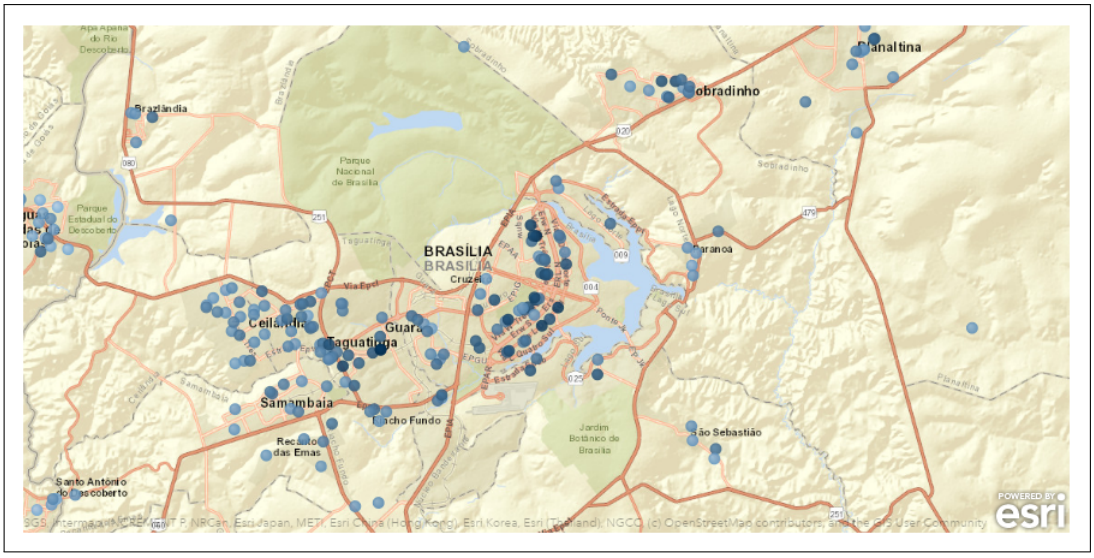


Figura 13 – Mapa de desempenho da escola - DF - 2018

Para o modelo, as localidades foram agrupadas em 5 classes, sendo 4 destas compostas por Regiões Administrativas do DF e a 5ª os muncípios do entorno, conforme seção 4.1.3. Analisando as notas perante essa classificação, tem-se um decaimento na distribuição do desempenho dada a renda média do local (variável usada para o agrupamento). No grupo 4, onde municípios mais afastados do centro são alocados, foram observadas menores notas, confirmando assim a ideia levantada pelos mapas anteriores. O mesmo acontece com as cidades no entorno do Distrito Federal.

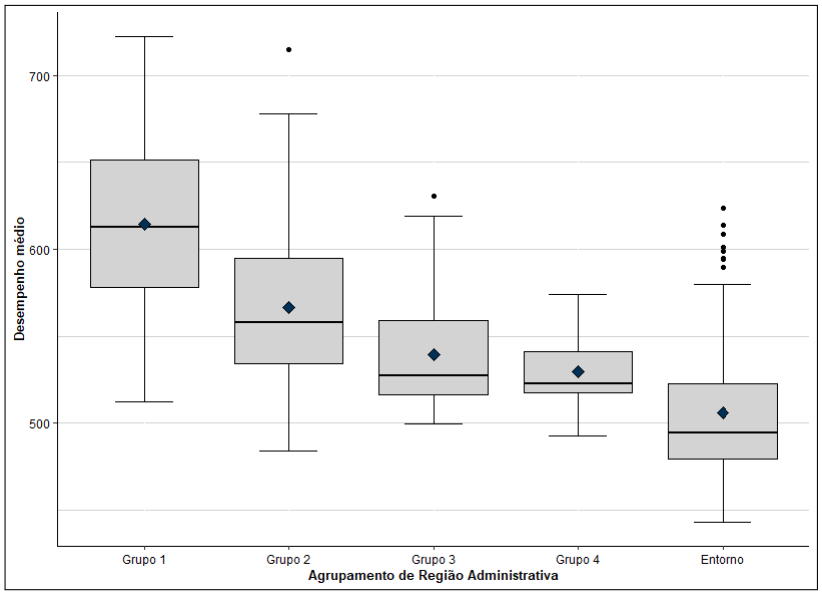


Figura 14 – Desempenho da escola por agrupamento de Localidade - 2018

Tabela 11 – Desempenho da escola por agrupamento de Região Administrativa - 2018

	Agrupamento de Região Administrativa				
	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Entorno
Média	614,3	566,5	539,5	529,7	505,8
Mediana	612,5	558,0	527,3	522,6	494,6
Desvio Padrão	53,3	44,1	30,5	25,9	39,5
N	45 (14%)	82 (26%)	64 (20%)	9 (3%)	118 (37%)

Tratando da dependência administrativa das escolas, o rendimento médio obtido em escolas privadas é melhor do que nos outros tipos. Um fato curioso que aparece nos dados e fica explícito na Figura 15 é a baixa variabilidade das notas médias em escolas estaduais, indicando que tais entidades mantém um padrão do desempenho. Infelizmente, essa equidade do ensino está nivelada por baixo, onde 95% das notas médias para escolas estaduais estão abaixo de 546 pontos.

Tabela 12 – Desempenho da escola por dependência administrativa - 2018

	Dependência Administrativa		
	Federal	Estadual	Privada
Média	527,32	433,2	495,6
Mediana	579,1	506,2	582,2
Desvio Padrão	31,4	28,3	46,6
N	13 (4%)	170 (53%)	135 (42%)

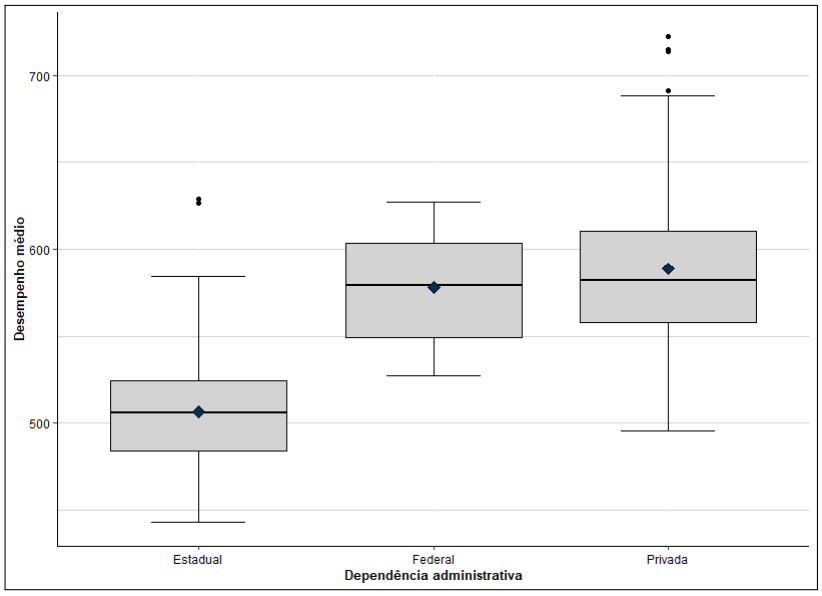


Figura 15 – Desempenho da escola por Dependência Administrativa - 2018

Alguns indicadores que medem atributos da escola foram analisados, buscando assim uma relação entre estes e o desempenho médio daquela escola. Inicialmente, observa-se algumas medidas descritivas de tais indicadores (Tabela 13), sendo eles: Média de alunos por turma, Média de horas-aula diária e Taxa de distorção idade-série (TDI)[2], todos relativos ao 3º ano do ensino médio, além da proporção de docentes do ensino médio com curso superior em seu currículo.

As escolas participantes do estudo possuem em suas turmas, em média, 31,5 alunos tendo 5,25 horas-aula por dia. Aproximadamente 92% dos professores do ensino médio possuem curso superior e a Taxa Média de Distorção Idade-série é de 17%.

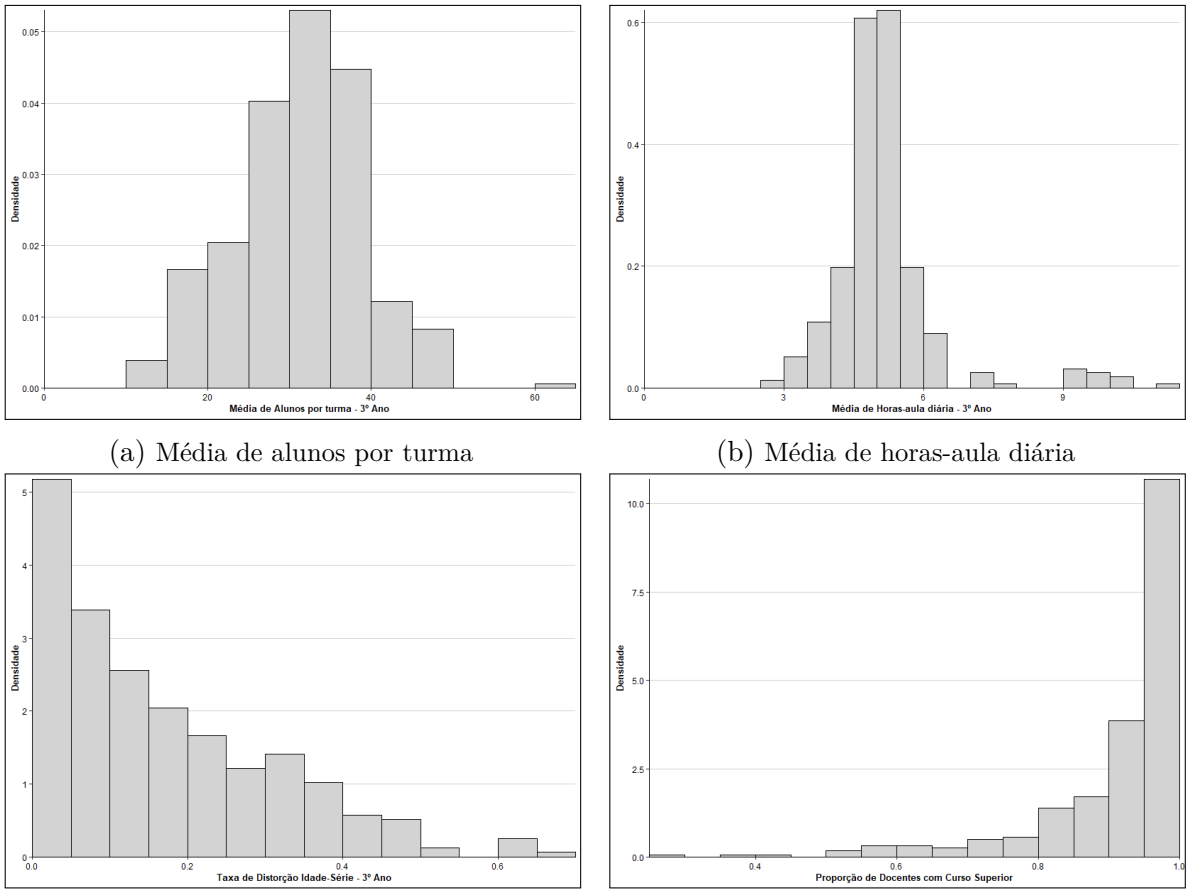
Quando analisada a Figura 16, nota-se que algumas poucas escolas possuem uma taxa de Distorção Idade-série maior que 50%, valor muito alto e fora do padrão apresentado. Quanto a proporção de professores com curso superior, a grande maioria das entidades tem ao menos 90% do seu corpo docente graduado com o ensino superior. São vistas como exceções, escolas com menos de 50% de educadores com curso superior.

Tabela 13 – Indicadores educacionais - 2018

	Média	**Mediana**	**Desvio Padrão**
Média de alunos por turma - 3º ano	31,56	32,3	7,88
Média de horas-aula diária - 3º ano	5,25	5,1	1,21
Taxa de Distorção Idade-Série - 3º ano	0,17	0,13	0,15
Proporção de docentes com curso superior	0,92	0,964	0,11

Se analisado o desempenho médio das escolas de acordo com os indicadores propostos

[2] A Taxa de distorção idade-série se refere à proporção de alunos que possuem ao menos dois anos a mais do que a idade indicada para aquela série (no caso, o 3º ano do ensino médio).



(a) Média de alunos por turma

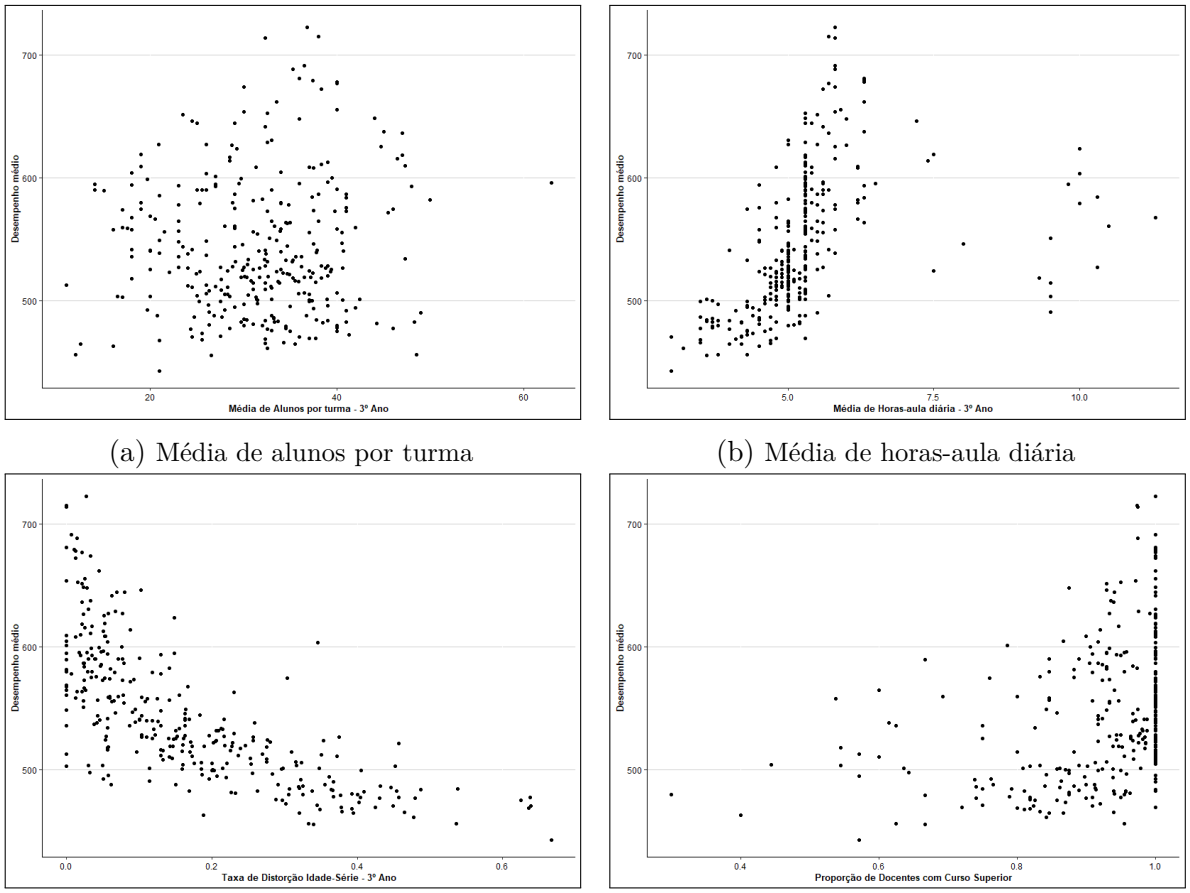
(b) Média de horas-aula diária

(c) Taxa de Distorção Idade-série

(d) Proporção de docentes com curso superior

Figura 16 – Distribuição dos Indicadores Educacionais - 2018

(Figura 17), nota-se uma tendência de crescimento desse desempenho quando a média de horas-aula e a proporção de docentes com curso superior aumenta. Já para uma maior TDI, a nota obtida pelos alunos daquela escola tende a cair.



(a) Média de alunos por turma

(b) Média de horas-aula diária

(c) Taxa de Distorção Idade-série

(d) Proporção de docentes com curso superior

Figura 17 – Desempenho de acordo com os Indicadores Educacionais - 2018

5.4 Modelagem

Conforme passos apresentados na seção 3.5 aqui inicia-se a modelagem multinível dos dados, tendo como variável resposta o desempenho médio dos alunos nas cinco competências do Enem de 2018. O modelo é composto por dois níveis, estudantes e escolas.

O primeiro modelo, aqui chamado de M0, é analisado sem o uso de variáveis explicativas, se utilizando apenas do intercepto. Com isso, tem-se:

Tabela 14 – Modelo sem variáveis explicativas - Modelo nulo (M0)

Variáveis explicativas	**Modelo Nulo (M0)**		
Efeito fixo	**Estimativa**	**Erro Padrão**	**P-valor**
Intercepto	544,46	3,10	<,0001
Efeito Aleatório - Nível 2			
Variância do Intercepto	2.948,41	241,36	<,0001
Efeito Aleatório - Nível 1			
Variância do resíduo	3.571,34	33,00	<,0001
Correlação intraclasse	45,22%		
Deviance	262.775,5		
Número de parâmetros	3		

Com esse modelo, é obtida uma primeira estimativa quanto o desempenho dos alunos que é de 544.46 pontos. Como estimativa de variância entre as escolas foi obtido um valor de 2.948,41 e entre os alunos 3.571,34. Esses valores são extremamente úteis para o cálculo do Coeficiente de Correlação Intraclasse (ICC), que no problema estudo foi de 45,22%. Isso significa que mais de 45% da variação do desempenho do aluno está contida na escola que ele estuda, justificando assim a utilização do método escolhido.

Partindo agora para o segundo modelo (M1), onde são incluídas as variáveis explicativas do primeiro nível, referentes ao aluno, foi obtido um *deviance* de 261.408,0 significativamente menor, pelo teste de razão de verossimilhanças, do que o obtido no modelo M0.

Tabela 15 – Modelo com variáveis explicativas do aluno (M1)

Variáveis explicativas	**Modelo com variáveis dos alunos (M1)**		
Efeito fixo	**Estimativa**	**Erro Padrão**	**P-valor**
Intercepto	682,70	7,38	<,0001
Idade	-9,09	0,38	<,0001
Sexo: Masculino	8,48	0,78	<,0001
Cor/raça: PPI	-6,17	0,84	<,0001
Ocupação: Grupo 4	11,79	1,00	<,0001
Ocupação: Grupo 5	11,36	1,67	<,0001
CCEB	58,59	4,19	<,0001
Efeito Aleatório - Nível 2			
Variância do Intercepto	1.881,35	159,98	<,0001
Efeito Aleatório - Nível 1			
Variância do resíduo	3.388,65	31,32	<,0001
Deviance		261.408,0	
Número de parâmetros		9	

Segundo às estimativas obtidas, a cada ano mais velho a nota esperada decai 9,09 pontos. Além disso, quando o concorrente é do sexo masculino, espera-se, em média, uma nota 8,5 pontos maior do que a do sexo oposto. Para pessoas pretas, pardas ou indígenas a situação esperada é de que a nota decaia em média 6,2 pontos em relação a brancos e amarelos. Outro fator que interfere no resultado obtido é a ocupação do responsável pelo estudante. Estudantes com pais com ocupações que exigem maior escolaridade tendem a ter um melhor resultado na prova, um aumento de 11,8 pontos para o grupo 4 e 11,4 pontos para o grupo 5, em relação a alunos com responsáveis empregados nos grupos 1, 2 ou 3. O índice socioeconômico também interfere no desempenho, onde cada décimo aumentado no índice implica numa esperança de 5,8 pontos a mais.

Quanto às estimativas para a variância, houve uma queda de 5,1% na variância do resíduo e de 36,2% para a variância do intercepto comparando com o modelo M0.

O próximo passo consiste na introdução das variáveis de escola (nível 2) no modelo

(M2). A adição dessas novas informações é significante, segundo o teste de razão de verossimilhanças aplicado.

Tabela 16 – Modelo com variáveis explicativas do aluno e escola (M2)

Variáveis explicativas	**Modelo com variáveis dos alunos e escolas (M2)**		
Efeito fixo	**Estimativa**	**Erro Padrão**	**P-valor**
Intercepto	615,23	19,99	<,0001
Idade	-8,93	0,38	<,0001
Sexo: Masculino	8,47	0,78	<,0001
Cor/raça: PPI	-5,87	0,84	<,0001
Ocupação: Grupo 4	10,98	1,00	<,0001
Ocupação: Grupo 5	10,40	1,68	<,0001
CCEB	54,99	4,22	<,0001
RA: Grupo 2	-35,39	4,39	<,0001
RA: Grupo 3	-40,10	4,766	<,0001
RA: Grupo 4	-32,19	8,58	0,0002
RA: Entorno	-47,89	5,29	<,0001
Dep. Adm.: Federal	34,91	8,18	<,0001
Dep. Adm.: Privada	37,17	4,06	<,0001
Média de Alunos por turma	0,69	0,18	0,0002
Média de Horas-aula diária	5,79	1,52	0,0002
Taxa de Distorção Idade-série	-59,07	15,49	0,0002
Docentes com Curso Superior	45,75	16,24	0,0051
Efeito Aleatório - Nível 2			
Variância do Intercepto	463,78	44,39	<,0001
Efeito Aleatório - Nível 1			
Variância do resíduo	3.384,97	31,40	<,0001
Deviance	258.977,6		
Número de parâmetros	19		

Conforme a Tabela 16, as variáveis de alunos seguem o mesmo padrão apresentado no modelo M1 com pequenas variações no valor das estimativas, mas sem nenhuma alteração nas conclusões feitas anteriormente. Quanto as ultimas variáveis adicionadas, que trazem informações sobre as escolas tem-se que o local em que a entidade é situada tem grande importância para a nota estimada, sendo que escolas fora do grupo 1 de localidades perdem ao menos 32 pontos, em média. O maior decaimento da nota esperada ocorre quando a escola está localizada no entorno, com uma queda esperada na média (intercepto apresentado) de 47,9 pontos. Quanto a dependência administrativa, espera-se que alunos advindos escolas federais ou privadas tenham um desempenho melhor do que alunos de escolas públicas estaduais. O aumento para escolas federais é de 34,9 pontos enquanto que para escolas privadas é de 37,2 no escore final, tudo isso levando em consideração a comparação com escolas cuja dependência administrativa é estadual.

Alguns indicadores escolares também foram importantes para predizer a nota obtida

pelos estudantes no Enem. De acordo com o modelo, quando a média de horas-aula diária para o 3º ano do ensino médio cresce em uma unidade, a nota obtida tem uma expectativa de crescimento de 5,8 pontos. Outros fatores que elevam a nota inferida é a porcentagem de docentes com curso superior no currículo (elevando 4,6 pontos a cada décimo acrescido) e a média de alunos por turma, fato curioso, uma vez que empiricamente, espera-se que turma com um menor número de alunos tenham um melhor aproveitamento. Por outro lado, quando a escola possui uma taxa de distorção idade-série alta, ou seja, uma grande porcentagem de alunos com ao menos dois anos a mais do que deveriam ter para o 3º ano indica uma queda no desempenho esperado.

Seguindo os passos anteriormente indicados, agora o modelo estudado (M3) leva em consideração as componentes de efeito aleatório, indicando se quaisquer coeficientes de regressão do nível dos alunos (nível 1) tem uma componente significativa de variância entre as escolas.

Pelo teste de razão de verossimilhanças, o modelo M3 se adéqua de melhor forma aos dados, existindo assim uma variância nos coeficientes inerente às escolas. Isso significa que a inclinação das retas de regressão para as variáveis idade e CCEB variam entre as escolas estudas. Com a adição desses componentes aleatórios, a variável "Taxa de Distorção Idade-série" deixou de ser significante para o problema, sendo assim retirada do modelo.

Vale lembrar que num modelo de regressão tradicional o valor da estimativa obtida para o parâmetro referente a variável representa o acréscimo no desempenho que aquela variável traz. Já numa estrutura de modelo multinível, este coeficiente representa uma média do crescimento obtido a cada alteração nas variáveis randômicas do modelo (Idade e Índice CCEB).

Quanto a estrutura das componentes de efeito aleatório, esta gera uma matriz de variância e covariância, estas ultimas não sendo apresentadas na Tabela 17 por questões de simplicidade. Tais covariâncias estão expostas no Apêndice E.

O último passo da modelagem considera além das componentes aleatórias, a interação entre variáveis de estudantes e da escola. Este modelo com interação não foi considerado ao final do estudo, uma vez que o aumento em sua complexidade (tanto computacional quanto de interpretação) é considerável e a melhora nas medidas da qualidade do modelo, como o AIC (Tabela 18), não foi tão grande. Com isso, tem-se como modelo final o M3.

No modelo selecionado, diversos fatores são indicados como importantes para a predição do desempenho no Enem. Iniciando pela idade, onde a cada ano completado tem-se a esperança de uma queda média de 11,9 pontos no desempenho do aluno, indicando assim que alunos mais novos e, por consequência, que estão na idade escolar correta para o 3º ano do Ensino Médio tem maior chance de obter um melhor desempenho. O sexo

Tabela 17 – Modelo com variáveis explicativas do aluno e escola, com efeito aleatório (M3)

Variáveis explicativas	**Modelo com efeito aleatório (M3)**		
Efeito fixo	**Estimativa**	**Erro Padrão**	**P-valor**
Intercepto	639,83	20,74	<,0001
Idade	-11,95	0,65	<,0001
Sexo: Masculino	9,16	0,78	<,0001
Cor/raça: PPI	-5,56	0,84	<,0001
Ocupação: Grupo 4	10,79	1,00	<,0001
Ocupação: Grupo 5	9,84	1,68	<,0001
CCEB	48,83	4,92	<,0001
RA: Grupo 2	-32,53	4,39	<,0001
RA: Grupo 3	-36,61	4,66	<,0001
RA: Grupo 4	-32,23	8,03	<,0001
RA: Entorno	-48,74	5,03	<,0001
Dep. Adm.: Federal	27,45	7,76	0,0005
Dep. Adm.: Privada	47,39	3,24	<,0001
Média de Alunos por turma	0,64	0,18	0,0004
Média de Horas-aula diária	8,47	1,31	<,0001
Docentes com Curso Superior	48,30	15,54	0,0021
Efeito Aleatório - Nível 2			
Variância do Intercepto	14.539	2.777,89	<,0001
Variância - Idade	43,56	8,14	<,0001
Variância - CCEB	1.106,80	373,54	0,0015
Efeito Aleatório - Nível 1			
Variância do resíduo	3.330,24	31,14	<,0001
Deviance	258.791,9		
Número de parâmetros	23		

do estudante também interfere na nota obtida, esperando que alunos do sexo masculino tenham uma melhor desempenho (9,2 pontos a mais) do que mulheres. Alunos pretos, pardos ou indígenas também possuem uma menor esperança de desempenho, tendo sua nota esperada 5,6 pontos menor do que alunos brancos ou amarelos. A ocupação dos pais é outro fator importante para predizer o desempenho dos alunos, uma vez que estudantes com seu responsável em empregos que exigem maior formação tendem a ter um desempenho de até 10,8 pontos acima em comparação a estudantes com pais em piores situações empregatícias. O índice socioeconômico aparece por último dentre as variáveis do primeiro nível mas também possui grande importância. A cada décimo acrescido no CCEB, espera-se em média 4,9 pontos a mais no desempenho obtido.

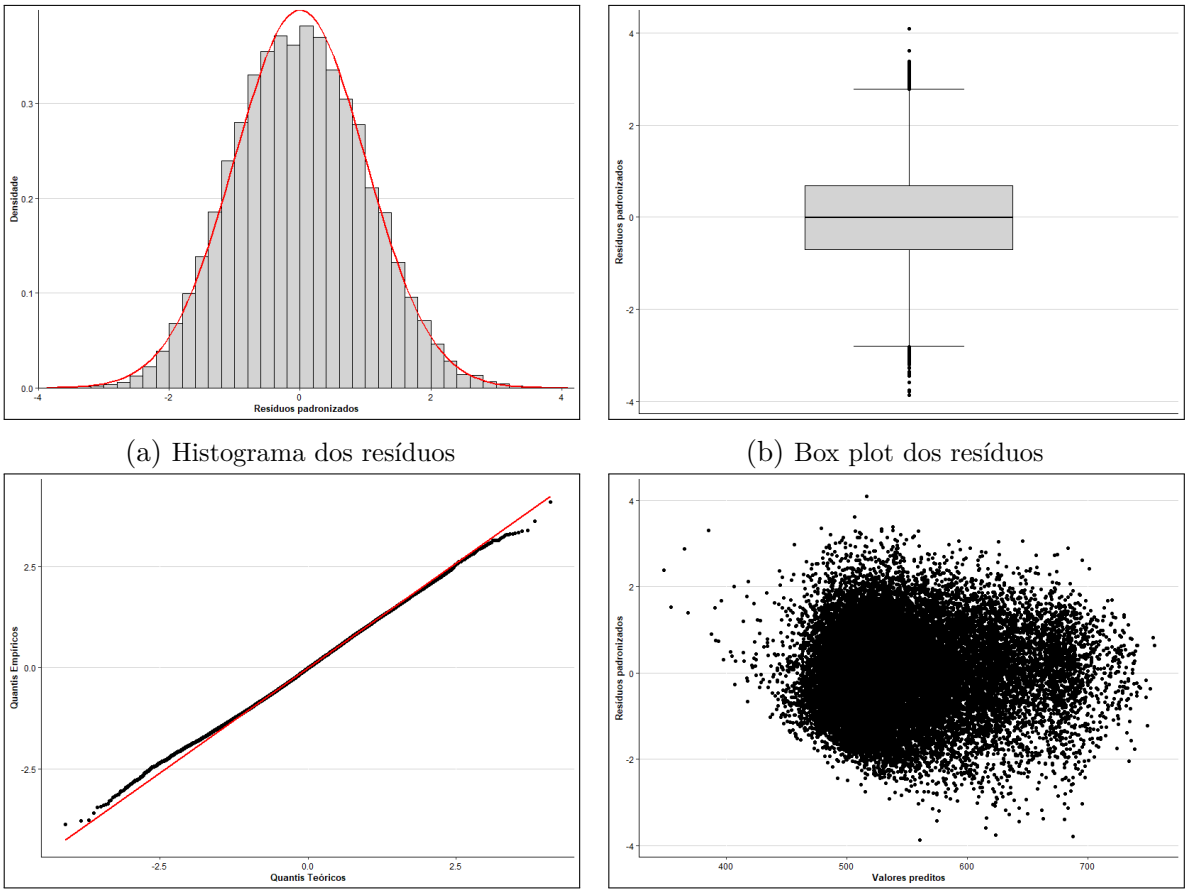
Variáveis relativas à escola também foram consideradas no último modelo, havendo assim significância em seu uso para predizer o desempenho obtido. A Região Administrativa da escola é um grande fator para o auxílio da estimativa de nota média obtida por um aluno. Estudantes de escolas fora das regiões administrativas de maior renda (Brasília, Jardim Botânico, Lago Norte, Lago Sul, Park Way, Sudoeste e Octogonal) tendem a ter

uma nota muito mais baixa do que os alunos de áreas de alta renda. Quando o estudante vem de uma escola do entorno tem-se a maior queda esperada, sendo ela de 48,7 pontos. A dependência administrativa da escola também interfere no desempenho do discente, onde estudantes advindos de escolas privadas tendem a ter uma nota 47,4 pontos maior do que alunos de escolas estaduais. Escolas com maior média de horas-aula, média de alunos por turma e docentes com curso superior também elevam a expectativa da nota de seus estudantes.

Tabela 18 – Informações doS modelos

Modelo	***Deviance***	**AIC**	**Número de parâmetros**
M0	262.775,5	262.781,5	3
M1	261.408,0	261.426,0	9
M2	258.977,6	259.015,6	19
M3	258.791,9	258.837,9	23
M4	258.738,0	258.790,0	26

Com o modelo selecionado, foram obtidos os resíduos studentizados com distribuição apresentada na Figura 18. Seguindo um dos pressupostos apresentados na seção 3.7, os resíduos possuem uma distribuição normal, centrada no 0. Pela grande quantidade de observações, o box plot (Figura) apresenta alguns outliers tanto superiores quanto inferiores porém, estes não interferem tanto na hipótese de normalidade, uma vez que os resíduos discrepantes fogem do centro da distribuição para baixo e para cima, com uma intensidade muito semelhante, equilibrando assim a distribuição e fazendo com que ela não tenha grande assimetria. O QQ plot (Figura 18d) indica uma pequena distorção da normalidade nas caudas de distribuição confirmando a ideia apresentada no box plot porém, esse pequeno distanciamento não invalida a suposição de distribuição gaussiana nos resíduos.



(a) Histograma dos resíduos

(b) Box plot dos resíduos

(c) QQ plot dos resíduos

(d) Resíduos × Valores preditos

Figura 18 – Distribuição dos resíduos studentizados do modelo M3

6 Conclusão

Ao início do estudo foi definido o objetivo principal de desenvolver um Modelo Linear Multinível buscando fatores que influenciem no desempenho médio dos alunos concluintes do Ensino Médio na Área Metropolitana de Brasília em todas as competências do Enem em 2018.

Tais fatores ficaram claramente definidos pelo modelo. Questões sociodemográficas como sexo, idade, cor/raça e o quesito econômico, analisados por meio da ocupação do responsável pelo estudante e pelo Critério de Classificação Econômica Brasil (CCEB), foram variáveis determinantes para a análise do desempenho dos alunos. Espera-se um decréscimo da nota quando os alunos tem um menor índice socioeconômico, são mais velhos, possuem sexo feminino, são negros ou indígenas, e seus responsáveis possuem piores condições de trabalho, indicando assim uma trilha a se seguir na construção de políticas públicas.

Já para as instituições estudadas, existe um melhor desempenho de alunos advindos de escolas da região de maior renda de Brasília, escolas privadas e escolas com maior média de alunos por turma, média de horas-aula diária e proporção de docentes com curso superior.

O estudo aqui descrito serve como ponto inicial para diversas outras análises que possam contribuir com a evolução do sistema educacional do Brasil, como por exemplo análises específicas de cada competências do Enem ou aplicação da metodologia em outros testes massivos de desempenho educacional, como o SAEB e até mesmo o ENADE.

Apêndice

A Grupo Ocupacional

No questionário socioeconômico do Enem é realizada a seguinte pergunta: “A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da(o) sua(seu) mãe(pai) ou da mulher(homem) responsável por você. (Se ela(e) não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela(e)).”. A partir daí, são apresentados 5 grupos de ocupações para que o estudante escolha a opção adequada à sua realidade. Os grupos são descritos a seguir:

- **Grupo 1:** Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista.
- **Grupo 2:** Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria.
- **Grupo 3:** Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista.
- **Grupo 4:** Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria.
- **Grupo 5:** Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de empresas com mais de 10 empregados.

De acordo com a distribuição da nota média dos estudantes nas 5 competências do Enem, os 3 grupos iniciais foram agrupados para a análise.

B Análise fatorial - Indicadores Socioeconômicos

Tabela 19 – Fatores rotacionados - Variáveis socioeconômicas*

Variável	**Fator 1**	**Fator 2**	**Fator 3**
Em sua residência trabalha empregado doméstico?	0.68133	0.00969	-0.0515
Na sua residência tem banheiro?	0.72411	0.38755	0.02898
Na sua residência tem quartos para dormir?	0.52248	0.42956	0.06946
Na sua residência tem carro?	0.5769	0.46531	0.09509
Na sua residência tem motocicleta?	-0.03168	-0.01588	0.67525
Na sua residência tem geladeira?	0.54148	0.02938	0.24131
Na sua residência tem freezer?	0.44966	0.27334	0.34079
Na sua residência tem máquina de lavar roupa?	0.10591	0.60946	0.32165
Na sua residência tem máquina de secar roupa?	0.20564	0.11154	0.52721
Na sua residência tem forno micro-ondas?	0.15307	0.52777	0.24801
Na sua residência tem máquina de lavar louça?	0.56321	-0.01534	0.10828
Na sua residência tem aspirador de pó?	0.50799	0.37246	0.03944
Na sua residência tem televisão em cores?	0.64684	0.39452	0.07719
Na sua residência tem aparelho de DVD?	0.31205	0.16104	0.20187
Na sua residência tem TV por assinatura?	0.39691	0.49381	-0.06454
Na sua residência tem telefone celular?	0.24735	0.53974	0.2079
Na sua residência tem telefone fixo?	0.2039	0.59366	-0.20841
Na sua residência tem computador?	0.49724	0.51749	0.05872
Na sua residência tem acesso à Internet?	-0.06725	0.72573	0.01261

*Os valores acima de 0.5 foram marcados na tabela.

C Análise fatorial - Indicadores de Infraestrutura

Tabela 20 – Fatores rotacionados - Variáveis de Infraestrutura*

Variável**	**Fator 1**	**Fator 2**	**Fator 3**	**Fator 4**	**Fator 5**
Água filtrada	0.56243	0.00867	-0.14136	0.2058	-0.00095
Banheiro	0.71272	0.12188	-0.03175	0.15025	-0.01913
Banheiro PNE	0.12271	0.06666	-0.00166	0.74026	0.08338
Biblioteca ou Sala de leitura	0.20037	0.00213	0.65725	0.06137	0.23912
Cozinha	0.42877	0.10359	-0.10827	-0.38468	0.37606
Dependências PNE	0.08552	-0.00339	0.12143	0.73393	0.23345
Copiadora	0.20254	0.59647	0.08828	0.15154	0.01343
DVD	0.01132	0.73392	-0.01654	-0.03102	0.03567
Impressora	-0.10273	0.528	0.06546	-0.06349	-0.06867
Projetor Multimídia	0.13281	0.43741	-0.10443	0.21276	0.08961
Parabólica	-0.04916	0.44923	0.13542	-0.2587	0.24411
TV	0.00456	0.58216	-0.03488	0.13156	0.01643
Rede de esgotamento	-0.03254	-0.02954	0.13939	0.01824	-0.09263
Internet	0.10451	0.16263	0.19153	0.35552	-0.0744
Labortório de ciências	0.0588	0.06101	0.32981	0.49214	0.48269
Laboratório de Informática	0.12588	0.33546	0.46814	0.00443	0.11914
Coleta de lixo periódica	-0.12713	-0.02604	0.52873	0.13805	-0.134
Reciclagem de lixo	-0.03392	-0.00366	-0.12701	0.08872	0.52326
Funcionamento em prédio escolar	0.02855	0.17481	-0.17774	-0.09721	0.36658
Quadra de esportes	0.27663	0.09763	0.2874	0.29587	0.40923
Refeitório	0.01135	-0.0713	0.20358	0.17627	0.62009
Sala de diretoria	0.72495	0.00184	0.3159	0.00619	0.11206
Sala dos professores	0.73217	-0.03518	0.33332	0.00825	0.00054
Secretaria	0.33013	0.071	0.67495	0.08362	-0.01093

*Os valores acima de 0.5 foram marcados na tabela.
**As variáveis indicam se a escola possui o item indicado.

D Critério de Classificação Econômica Brasil

Para o Critério de Classificação Econômica Brasil (CCEB) foram consideradas as seguintes informações:

Tabela 21 – CCEB - Poder Aquisitivo

	Quantidade				
	0	**1**	**2**	**3**	**4+**
Banheiros	0	3	7	10	14
Empregados domésticos*	0	3	7	10	13
Automóveis	0	3	5	8	11
Microcomputador	0	3	6	8	11
Lava louça	0	3	6	6	6
Geladeira	0	2	3	5	5
Freezer	0	2	4	6	6
Lava roupa	0	2	4	6	6
DVD*	0	1	3	4	6
Micro-ondas	0	2	4	4	4
Motocicleta	0	1	3	3	3
Secadora de roupa	0	2	2	2	2

Fonte: ABEP.

Tabela 22 – CCEB - Escolaridade da pessoa de referência

Escolaridade	**Pontuação**
Fundametal I incompleto	0
Fundamental II incompleto	1
Médio incompleto	2
Superior incompleto	4
Superior completo	7

Fonte: ABEP.

Tabela 23 – CCEB - Acesso à serviços públicos

	Sim	**Não**
Água encanada*	0	4
Rua pavimentada*	0	2

Fonte: ABEP.

Para o estudo apresentado, as variáveis demarcadas com um asterisco precisaram de uma adaptação, de acordo com as variáveis disponibilizadas no banco de dados trabalhado. Tais adaptações estão descritas abaixo:

- O indicador utiliza o número de empregados domésticos na residência, enquanto que o Enem disponibiliza uma variável referente ao número de dias da semana em que se tem empregado doméstico na residência. Além disso, os níveis apresentados pelo

questionário socioeconômico do Enem estão agrupados, sendo os grupos definidos por “Não possui”, “Um ou dois dias por semana”, “Três ou quatro dias por semana” ou “Pelo menos cinco dias por semana”. Com isso, os novos pesos foram estabelecidos como 0, 5, 11 e 11, respectivamente;

- A variável “DVD” está codificada no banco de dados socioeconômicos do Enem com os níveis “Sim” e “Não”, sem a indicação da quantidade de aparelhos. Dessa forma, foi dado um peso 0 para quem não possui aparelho de DVD e 4 para quem possui;
- Foi considerado o maior nível de escolaridade apresentado por um dos responsáveis do estudante, sendo este não necessariamente o nível do chefe da família, como indicado pelo CCEB;
- As variáveis referentes ao acesso à serviços públicos foram retiradas do indicador;

Com as alterações feitas, o *score* obtido por algum aluno poderia chegar a 90 pontos, sendo que no índice original, essa pontuação atinge os 100 pontos. Para uma melhor interpretação da medida, as pontuações obtidas foram reescaladas, variando entre 0 e 1.

E Matriz de variância e covariância - Modelo M3

$$\mathbf{S} = \begin{bmatrix} 14.539 & -786,14 & 491,25 \\ -786,14 & 43,56 & -28,89 \\ 491,25 & -28,89 & 1.106,80 \end{bmatrix}$$

Sendo “Intercepto”, “Idade” e “CCEB” a ordem das variáveis que compõe a matriz.

Referências

ANDRADE, J. M.; LAROS, J. A. Fatores associados ao desempenho escolar: Estudo multinível com dados do SAEB/2001. *Psicologia: Teoria e Pesquisa*, v. 23, n. 1, p. 033–042, 2007.

ASSOCIAÇÃO BRASILEIRA DE EMPRESAS DE PESQUISA. Critério de Classificação Econômica Brasil. Acesso em: 11 de Novembro de 2019. Disponível em: <http://www.abep.org/criterioBr/01_cceb_2018.pdf>.

BARROS, R. P. de; MENDONÇA, R. Investimento em educação e desenvolvimento econômico. *A Economia Brasileira em Perspectiva - 1998*, Rio de Janeiro: IPEA, v. 2, p. 605–614, 1998.

BOTELHO, D. S. Análise do desempenho no ensino médio na área metropolitana de brasília: uma abordagem multinível. *Trabalho de Conclusão de Curso (Bacharelado em Estatística)*, Universidade de Brasília, Brasília, 2017.

COELHO, F. R. Seleção de modelos multiníveis para dados de avaliação educacional. *Dissertação(Mestrado em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística)*, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos - São Paulo, p. 171, 2017.

COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL - CODEPLAN. Delimitação do Espaço Metropolitano de Brasília (Área Metropolitana de Brasília), Brasília, 2014. Acesso em: 08 de novembro de 2019. Disponível em: <http://www.codeplan.df.gov.br/wp-content/uploads/2018/03/Delimita%C3%A7%C3%A3o-do-Espa%C3%A7o-Metropolitano-de-Bras%C3%ADlia-AMB.pdf>.

COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL - CODEPLAN. PDAD - Pesquisa Amostra por Amostra de Domicílios, Brasília, 2018. Acesso em: 08 de novembro de 2019. Disponível em: <http://www.codeplan.df.gov.br/wp-content/uploads/2019/03/PDAD_DF-Grupo-de-Renda-compactado.pdf>.

FREIRE, P. Educação e mudança. *Rio de Janeiro: Paz e Terra*, 1979.

HOX, J. J. *Multilevel analysis: Techniques and Applications*. 2. ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2010.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, Censo Escolar. Acesso em: 09 de maio de 2019. Disponível em: <http://inep.gov.br/web/guest/censo-escolar>.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, Enem. Acesso em: 09 de maio de 2019. Disponível em: <http://inep.gov.br/web/guest/enem>.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. 6. ed. New Jersey: Pearson Prentice Hall, 2007.

KUTNER, M. H. et al. *Applied Linear Statistical Models*. 5. ed. Boston: McGraw-Hill / Irwin, 2005.

LAROS, J. A.; MARCIANO, J. L.; ANDRADE, J. M. de. Fatores associados ao desempenho escolar em português: um estudo multinível por regiões. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 20, n. 77, p. 623–646, 2012.

LAROS, J. A.; MARCIANO, J. L. P. Análise multinível aplicada a dados do NELS:88. *Estudos em Avaliação Educacional*, v. 19, n. 40, p. 263–278, 2008.

MEDEIROS, M.; OLIVEIRA, L. F. B. de. Desigualdades regionais em educação: potencial de convergência. *Soc. estado*, Brasília, v. 29, n. 2, p. 561–585, Aug. 2014.

QUEIROZ, E. P. de. A migração intrametropolitana no distrito federal e entorno: o consequente fluxo pendular e o uso dos equipamentos urbanos de saúde e educação. *Anais*, p. 1–17, 2016.

RAUDENBUSH, S. W.; BRYK, A. S. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2. ed. Thousand Oaks, California: SAGE Publications, Inc, 2001. (Advanced Quantitative Techniques in the Social Sciences).

SOARES, J. F. O efeito da escola no desempenho cognitivo de seus alunos. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, v. 2, n. 2, p. 88–104, 2004.

SOARES, J. F.; ALVES, M. T. G. Desigualdades raciais no sistema brasileiro de educação básica. *Educação e Pesquisa*, São Paulo, v. 29, n. 1, p. 147–165, 2013.