UNIVERSIDADE FEDERAL DE JUIZ DE FORA

INSTITUTO DE CIÊNCIAS EXATAS

DEPARTAMENTO DE ESTATÍSTICA

Jéssica de Almeida Fernandes

Avaliação do efeito da escola pública de MG no desempenho dos alunos no ENEM

Juiz de Fora

2018

Jéssica de Almeida Fernandes

Avaliação do efeito da escola pública de MG no desempenho dos alunos no ENEM

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a colação do grau em Bacharel em Estatística.

Orientador: Augusto Carvalho Souza

Doutor em Economia - CEDEPLAR/UFMG

Juiz de Fora

2018

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

de Almeida Fernandes, Jéssica.
Avaliação do Efeito da Escola Pública de MG no desempenho dos alunos no ENEM / Jéssica de Almeida Fernandes. -- 2018.
39 p.

Orientador: Augusto Carvalho Souza
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, ICE/Engenharia, 2018.

1. Modelos Hierárquicos. 2. Modelos Multiníveis. 3. ENEM. I. Carvalho Souza, Augusto, orient. II. Título.

Jéssica de Almeida Fernandes

Avaliação do efeito da escola pública de MG no desempenho dos alunos no ENEM

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a colação do grau em Bacharel em Estatística.

Orientador: Augusto Carvalho Souza

Doutor em Economia - CEDEPLAR/UFMG

Aprovado em 04 de Julho de 2018

BANCA EXAMINADORA

_____

Augusto Carvalho Souza

Doutor em Economia – CEDEPLAR/UFMG

_____

Ângela Mello Coelho

Doutora em Estatística e Experimentação Agronômica – ESALQ/USP

_____

Ronaldo Rocha Bastos

Ph. D in Regional Planning - Liverpool University

Para Ivonete, Iran e Nayara.

Com todo o meu amor.

Agradecimentos

Em primeiro lugar agradeço a Deus, que me deu força e perseverança até aqui, que não me desamparou em nenhum momento de minha vida e fez Sua presença visível em todas as situações. Nunca estive só. À minha mãe, minha base, a quem palavras não são suficientes para agradecer. Todo meu amor é seu. Ao meu pai por todo o apoio e motivação dado, pelos conselhos e ombro amigo. À minha esposa, Nayara, que me consolou a cada tombo e riu cada riso meu, que comemorou todas as minhas vitórias e com quem espero desfrutar de todos os minutos de minha vida e além. Às minhas amigas, que por horas me ouvem falar e me motivam em cada desânimo vivido. Em especial à Thalita, companheira de profissão e de vida que é um anjo da guarda para mim. À Mariana por todas as horas de conversas, todas as mil mensagens e por ser um porto seguro sempre que tudo deu errado. À Laila, que mesmo na distância se faz presente em todas as etapas e tem uma parte marcada em mim. À Isabela, por toda a luta com Física 3, pelas risadas e pela maravilhosa formatura. Nunca teria conseguido sem vocês, de verdade, muito obrigada. À minha família que sempre me inspirou e me fez buscar mais, principalmente minhas avós Terezinha e Amélia, e meu padrinho Airton. Obrigada por acreditarem em mim.

Agradeço também imensamente aos companheiros da Guiando, uma empresa maravilhosa e diferenciada sem a qual eu jamais me desenvolveria tanto a ponto de estar aqui. Obrigada por todos os conselhos técnicos, comportamentais e lições de vida.

Ao professor Augusto por aceitar o desafio de me orientar e perdoar os meus leves atrasos. Por todo conhecimento compartilhado e por toda paciência.

À Universidade Federal de Juiz de Fora (UFJF) pelo acolhimento nesses anos e em especial por fornecer equipamentos e estrutura adequados para um bom aprendizado.

Aos professores do departamento de Estatística, em especial aos professores Ângela e Marcel, pelos ensinamentos, dedicação e paciência ao longo dos anos.

Aos professores da banca por aceitarem nosso convite e engrandecerem ainda mais nosso trabalho. Muito obrigada!

“Diz-se que, mesmo antes de um rio cair no oceano ele treme de medo (...). E somente quando ele entra no oceano é que o medo desaparece. Porque apenas então o rio saberá que não se trata de desaparecer no oceano, mas tornar-se oceano.”

(Osho)

“É preciso força pra sonhar e perceber que a estrada vai além do que se vê.”

(Los Hermanos)

Resumo

O ENEM é o principal meio de ingresso ao ensino superior no Brasil, sendo critério integral ou parcial obrigatório para a entrada nas Universidades Públicas do país e servindo como avaliação de seleção para universidades no exterior. Entretanto, a distribuição das vagas ofertadas não é feita de forma igualitária, pois alunos que advém de ensino público têm direito a concorrer a vagas exclusivas por meio da Lei nº 12.711/2012 de 29 de agosto de 2012. Nela, está estabelecida a cota para alunos de escola pública, cuja motivação é a de que o ensino público é responsável pela defasagem na nota obtida no exame. Além disso, também obtivemos noção da diferença entre as proporções de alunos em ambos os tipos de ensino. Com o objetivo de averiguar essa informação, este trabalho pretende quantificar qual impacto que o tipo da escola, se pública ou privada, de ensino médio tem nas notas dos alunos de Minas Gerais no ENEM 2014. Para obter esse resultado utilizamos modelos hierárquicos de dois níveis e pudemos concluir que o tipo de escola é responsável por aumentar a nota do aluno em 44,24 pontos, caso ele tenha feito a maior parte do Ensino Médio em escola particular. Outros resultados obtidos foram relativos à responsabilidade das características de cada aluno, como classe de *status* socioeconômico e escolaridade da mãe do candidato.

Palavras-Chave: ENEM; Modelos Hierárquicos; Modelos Multiníveis.

Abstract

The ENEM is the main means of entering higher education in Brazil, being an integral or partial criterion for entry into the Public Universities of the country and serving as a selection evaluation for universities abroad. However, the distribution of the places offered is not done in an egalitarian way, since students who come from public education have the right to compete for exclusive places by means of Law no. 12.711 / 2012 of August 29, 2012. In it, the quota is established for public school students whose motivation is that public education is responsible for the gap in the grade obtained in the exam. In addition, we also obtained notion of the difference between the proportions of students in both types of education. In order to obtain this information, this work intends to quantify what impact the type of school, whether public or private, of high school has on the grades of the students of Minas Gerais in ENEM 2014. To obtain this result we use hierarchical models of two levels and we were able to conclude that the type of school is responsible for raising the student's grade by 44.24 points if he has done most of the high school in private school. Other results were related to the responsibility of the characteristics of each student, such as socioeconomic status class and schooling of the candidate's mother.

Keywords: ENEM; Hierarchical Modeling; Multilevel Modeling.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Ao longo da história, sempre foi importante avaliar os seres humanos em relação a determinados critérios. Segundo Soeiro e Aveline (1982), desde os primórdios da humanidade, em algumas tribos, os jovens só eram considerados adultos após terem sido aprovados em um teste sobre seus conhecimentos e costumes. Após o século XVIII, começaram a serem formadas as primeiras escolas modernas e as avaliações começaram a ser feitas de forma mais estruturada. Já no final do século XIX até parte do século XX, uma área de destaque foi a psicometria, caracterizada por testes padronizados para medir o desempenho e a inteligência das pessoas.

Atualmente, existem várias maneiras de verificarmos a aprendizagem de um aluno sobre um determinado tema ou conteúdo. Neste sentido, em 1998 o Ministério da Educação (MEC) criou o Exame Nacional do Ensino Médio (ENEM). Com essa prova pretendia-se avaliar os conhecimentos dos alunos sobre os conteúdos do ensino médio e também a criação de parâmetros para a auto-avaliação do participante e a criação de referências nacionais para o aperfeiçoamento do ensino.

A partir do ano de 2009 o ENEM sofreu diversas mudanças em sua composição, entre elas: deixou de ser uma prova com 63 questões multidisciplinares, passou a ser estruturado por competências em quatro áreas do conhecimento – Linguagens, Códigos e suas tecnologias, Matemática e suas tecnologias, Ciências Humanas e suas tecnologias e Ciências da Natureza e suas tecnologias – e a prova passou a ser realizada em dois dias diferentes. Além disso, em 2011 o ENEM começou a ser utilizado como mecanismo único, alternativo ou complementar para o acesso à Educação Superior no Brasil, seja ela pública ou privada.

Além de ser a principal forma de entrada dos alunos em instituições públicas, o ENEM ainda é usado para acesso a programas do governo como Financiamento Estudantil (FIES) e bolsas de estudo em universidades privadas através do Programa Universidade para Todos (ProUni). Outra utilização do programa é para o acesso a Universidades de Portugal com as quais, segundo o site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio

Teixeira (INEP), já foram firmados 29 acordos com universidades portuguesas para aceite do ENEM como forma de ingresso dos alunos. Com tantas opções de utilização, o ENEM é um dos focos do processo de ensino/aprendizagem em escolas de ensino médio no Brasil, sejam elas de rede pública municipal, estadual, federal ou da rede privada.

Entretanto, a distribuição das vagas não é feita de forma igualitária. Isso acontece por conta da lei nº 12.711/2012 de 29 de agosto de 2012, conhecida como "a lei das cotas". A partir dessa lei, todas as universidades devem reservar 50% de suas vagas para candidatos oriundos de escolas públicas. Segundo o sítio do MEC (htttp:// www.mec.gov.br), o objetivo dessa ação afirmativa é corrigir a desigualdade de pontuação entre os alunos dos dois tipos de rede de ensino. De acordo com Menezes Filho (2007), alunos de escolas particulares têm um desempenho melhor que os de escola pública mesmo levando em consideração variáveis de confusão *(confounders)* de cunho familiar e pessoal.

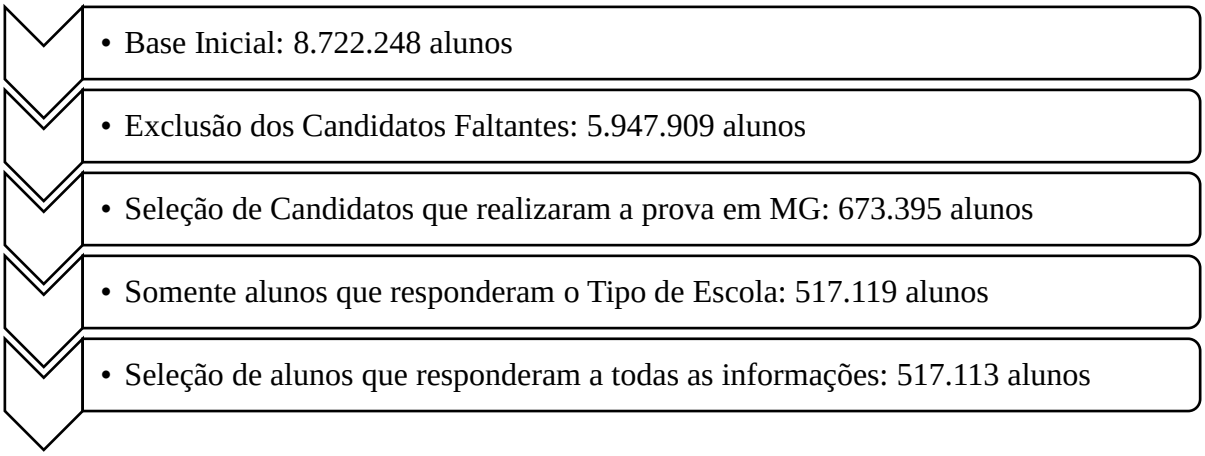
Os dados do ENEM de 2014 foram disponibilizados pelo sítio do INEP.[1] Nesse sítio é possível fazer download de todas as informações das edições do exame contendo os microdados, um dicionário das variáveis, as provas do ano de edição com seus gabaritos e um manual sobre o Enem.

O foco deste trabalho será estudar o tamanho da influência que o tipo de escola em que o aluno cursou a maior parte do Ensino Médio na sua nota final no Enem 2014 e verificarmos o quanto essa influência impacta em suas possíveis aprovações no Ensino Superior.

Após a leitura e limpeza da base de dados, chegamos ao total de 517.113 alunos que realizaram a prova no estado de Minas Gerais, dentre os quase cinco milhões que haviam realizado o exame no Brasil. Os passos para a obtenção desse tamanho de população serão melhores descritos no Capítulo 2, mas podem ser observados pelo fluxograma de trabalho descrito na Figura 1.

[1] www.inep.gov.br/microdados

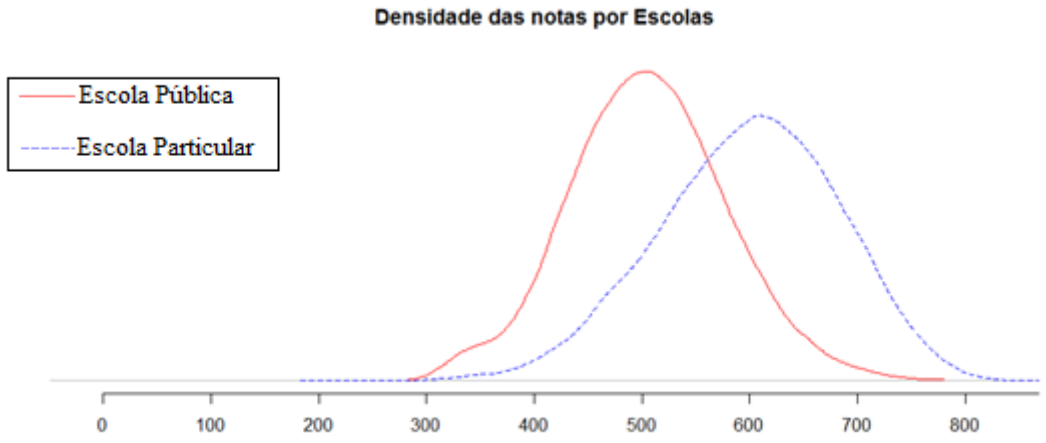
Figura 1 Fluxograma de Trabalho e obtenção da Base de Dados Final



Fonte: Do Autor

Separando os candidatos de Minas Gerais pelos dois grupos de escola em que cursaram a maior parte do ensino médio, pública ou privada, conseguimos perceber a diferença da distribuição das notas dos dois grupos de escolas, conforme mostra a Figura 2.

Figura 2 Densidade Estimada das Notas dos Alunos de MG por Tipo de Escola



Fonte: Do autor.

A partir da Figura 2 conseguimos notar que a distribuição que representa as notas dos alunos que tiveram a maior parte do ensino médio em escolas particulares está deslocada em relação à curva dos candidatos que tiveram seu ensino médio, predominantemente, em escolas públicas. O objetivo principal deste trabalho é estudar o tamanho dessa diferença e o que ela representa na vida desses estudantes. Para isto, usaremos principalmente, o modelo hierárquico ou multinível para estimar o efeito da escola pública na nota média dos alunos.

1.1 Modelos Hierárquicos

De acordo com Pinheiro (2005), os indivíduos são influenciados pelos grupos sociais em que eles vivem, e por sua vez esses indivíduos imprimem características e significados ao grupo. Isso significa dizer que não podemos olhar os indivíduos de forma separada, uma vez que a correlação entre indivíduos de um mesmo grupo tende a ser maior que a entre indivíduos de grupos diferentes. Para podermos analisar o efeito que o grupo tem sobre eles, devemos usar um modelo de regressão denominado multinível ou hierárquico, pois eles levam em consideração a variabilidade intragrupos e entre grupos.

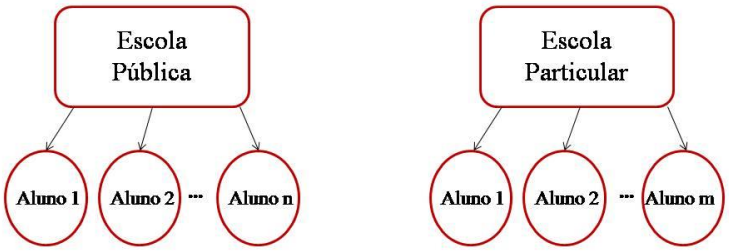
Ainda de acordo com Pinheiro (2005), o termo modelo multinível, foi introduzido em 1972 por Lindley e Smith, mas também estão relacionados a um estudo feito por Benett em 1976. Neste estudo, ele observou que crianças expostas a uma maneira formal do aprendizado da leitura exibiam desempenho maior que as não expostas. Cerca de cinco anos depois, em 1981, Atkins et al demonstraram que ao conduzir o estudo com as crianças agrupadas em classes escolares, a diferença de aprendizado sumia.

Assim, quando os dados que têm estrutura em vários níveis são analisados sem levar em conta esse arranjo, podem existir dois principais tipos de problemas. Um deles é amostral: quando agrupamos os dados de um nível mais baixo, como por exemplo, leitos, em um nível acima, como hospitais, temos uma considerável redução no tamanho da amostra original. O outro tipo de problema pode ocorrer na interpretação dos dados, tais como tirar conclusões de um nível e emiti-las sobre o outro nível, seja do nível individual para o agregado ou vice-versa.

Pela ótica de Bergamo (2002), não levar em consideração a estrutura hierárquica dos dados pode implicar em uma superestimação dos coeficientes do modelo em estudo. A principal

diferença está no fato de que um modelo de regressão tradicional tem uma suposição de independência entre os indivíduos, entretanto, quando há uma organização em formas hierárquicas, indivíduos pertencentes a um mesmo grupo raramente são independentes. Isso acontece porque esses indivíduos apresentam características semelhantes. Dessa forma, a suposição de independência para o modelo tradicional foi violada e o modelo deverá levar em consideração essa estrutura com a correlação entre indivíduos. Para o contexto educacional, um exemplo de estruturação hierárquica seria considerarmos alunos como um primeiro nível dentro de um tipo de Rede de ensino - pública, privada, etc. O desenho esquemático deste exemplo encontra-se na Figura 3.

Figura 3 Esquema hierárquico com aplicação no contexto educacional



Fonte: Do autor

* n e m são os tamanhos das populações dos grupos

Em linhas gerais, o exemplo esquemático da Figura 3 representa a estrutura dos dados analisados neste trabalho, cujo objetivo é avaliar o impacto do tipo de escola na nota média do aluno no ENEM. Utilizaremos as variáveis explicativas ao nível do indivíduo (nível 1) e um agrupamento em tipo de rede de ensino (nível 2).

No próximo capítulo iremos apresentar as características da base de dados e da metodologia proposta, no capítulo 3 traremos outros resultados com o enfoque do objetivo principal.

2 Métodos e Dados

2.1 Origem e Base de Dados

A base de dados contém informações dos alunos relacionadas à prova aplicada de forma não identificada, as respostas do questionário socioeconômico respondido pelos próprios candidatos e algumas informações de controle. Entre as informações sobre a prova aplicada encontramos, por exemplo, variáveis indicando a presença em cada uma das provas e algumas indicando a condição de deficiências de diversos tipos. Para efetuarmos a leitura e limpeza dos dados, utilizamos o software R (R Core Team (2018)).

Neste processo, excluímos as variáveis que não foram alvo deste estudo, selecionamos apenas os alunos presentes em todas as provas e que não foram eliminados em nenhuma delas e selecionamos apenas os alunos que responderam ao questionário socioeconômico. Além disso, também selecionamos apenas os alunos que realizaram a prova no estado de MG, unidade da federação que escolhemos para análise. Para este trabalho, calculamos a nota final de cada aluno como sendo a soma das notas em todas as competências (a nota presente na base de dados do INEP está separada por competência)[2]. Usaremos aqui a nota média definida como:

$$\mu_{nota} = \frac{Soma\ das\ Notas\ das\ Competências\ +\ Nota\ da\ Redação}{5},$$

Em função da grande quantidade de observações faltantes e possível viés nas respostas para a variável renda, optamos por utilizar como indicador do *status* socioeconômico do aluno o sistema de classificação da Associação Brasileira das Empresas de Pesquisa (ABEP), através do Critério Brasil 2014. Nele há um sistema de pontos a cada pergunta respondida no

[2] Linguagens, Códigos e suas tecnologias, Matemática e suas tecnologias, Ciências Humanas e suas tecnologias e Ciências da Natureza e suas tecnologias

questionário socioeconômico e a soma deles classifica o indivíduo em cada uma das quatro classes. Os valores de pontuações e de corte podem ser conferidos no Anexo I.

Após serem realizadas todas as modificações, obtivemos uma base de dados para análise de tamanho reduzido, o que permitiu realizar este trabalho diretamente em um computador pessoal. A base de dados utilizada poderá ser encontrada em www.ufjf.br/cursoestatistica, em csv, disponível para reprodução deste trabalho.

2.2 Modelagem dos dados

Os modelos lineares são amplamente utilizados em problemas de diversas áreas, e podem ser escritos da forma:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} + e_i,$$

Com i = 1, ... , n e j = 1, ... , k

Em que:

Y_i É a variável explicada no i-ésimo indivíduo;

β_0 Representa o intercepto da reta com o eixo Oy;

$\beta_1, \beta_2, \dots, \beta_k$ são os coeficientes da regressão;

X_{ij} É o valor do i-ésimo indivíduo na j-ésima variável explicativa;

e_i É o erro associado ao i-ésimo indivíduo.

Entretanto, para que esse modelo seja válido, os dados devem seguir os seguintes pressupostos (Bussab & Morettin, 2004):

Os erros são uma variável aleatória e seguem uma distribuição normal com variância constante e média zero. Em notação, $e_i \sim N(0,\ \sigma^2)$;

1. As variáveis aleatórias $e_1, e_2, \dots, e_n$ são independentes;

2. As variáveis explicativas X_1, X_2, ... , $X_j$são não correlacionadas, ou seja, não há multicolinearidade entre as variáveis explicativas;

Porém, em estruturas de dados em que indivíduos são agrupados segundo alguma característica, há, possivelmente, a quebra do pressuposto de independência entre eles, uma vez que unidades de um mesmo grupo tendem a apresentar características semelhantes entre si.

Por conta desta característica, utilizaremos modelos hierárquicos para estimar o efeito da escola pública na nota final do aluno no ENEM, onde os alunos serão considerados no nível 1 e os tipos de escolas serão considerados no nível 2. Para atenuar um possível efeito de nível de estado, selecionamos apenas os alunos que realizaram a prova no estado de Minas Gerais. Desta forma, teremos um modelo hierárquico de dois níveis.

2.3 Modelos Hierárquicos de Dois Níveis

Nesta classe de modelos, os dados possuem uma organização tal que a variável resposta é observada ao nível do indivíduo e as variáveis explicativas podem ser observadas tanto no nível individual quanto nos níveis dos grupos.

Para Scott, Shrout e Weinberg (2013), no processo de identificação de um modelo hierárquico, devemos iniciar pelo formato mais simples, semelhante à um modelo de regressão linear simples, com apenas uma variável explicativa e sem nenhum grupo. Após isto, vamos inserindo as informações a fim de explicar a variação dos dados. Um exemplo de modelo hierárquico de dois níveis pode ser escrito na forma:

$$Y_{ij} = \beta_2 X_{ij} + \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}$$

Em que temos dois efeitos aleatórios e um coeficiente de efeito fixo. Os efeitos de grupo podem ser escritos como:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$$

Em que u_{0j} e u_{1j} são os erros do segundo nível da hierarquia. Assume-se que os erros de nível 1 são normalmente distribuídos com variância comum σ^2, em todos os grupos. Os erros de nível 2, u_{0j}e u_{1j}são independentes dos erros e_{ij}e têm distribuição normal multivariada com médias iguais a zero.

2.4 Estimação dos Parâmetros

Para estimarmos os parâmetros de um modelo linear simples, precisamos que todos os pressupostos do modelo sejam respeitados, isto é: os erros distribuem-se ao redor da média $\alpha + \beta x$, com média zero E $(\varepsilon_i|x) = 0$; Todos os erros têm a mesma variabilidade em torno dos níveis de X, Var $(\varepsilon_i|x) = \sigma_\varepsilon^2$ e os erros são não-correlacionados. Além disso, supomos que a variável X é fixa (ou sem erro ou determinística) e que os estimadores de α e β serão aqueles que minimizem as somas dos quadrados dos erros. (Bussab & Moretin, 2004). Como:

$$\varepsilon_i = y_i - (\alpha - \beta x_i), \qquad i = 1, \dots, n$$

Então temos que:

$$\text{SQ}(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2$$

Derivando a equação em relação à α e β e igualando a zero, chegamos ao resultado que:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \text{ e}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Quando consideramos modelos hierárquicos, a estimação pode ser feita para três tipos de parâmetros: os efeitos fixos, efeitos aleatórios de primeiro nível e os componentes de variância e covariância. Segundo Hox (1995), os estimadores mais usados em modelos hierárquicos são os de máxima verossimilhança que ao estimarem valores populacionais maximizam a função de verossimilhança e obtém estimativas para os parâmetros do modelo. Essa estimação requer, em muitos casos, um método computacional iterativo, como o algoritmo EM. De acordo com

Hox (1998), habitualmente os *softwares* utilizados geram valores iniciais a partir de estimativas de mínimos quadrados. Já na segunda iteração são obtidos valores estimados de mínimos quadrados generalizados, que são utilizados para calcular os estimadores do segundo nível da hierarquia. Essas estimativas são utilizadas para estimar as variâncias e covariâncias no primeiro e segundo níveis da hierarquia.

Existem duas formas de estimação de máxima verossimilhança: A máxima verossimilhança completa (ML) e a máxima verossimilhança restrita (MLR), sendo que a MLR é uma versão do processo de ML, para modelos mistos. Nesse método, cada observação é dividida em duas partes independentes, uma que se refere aos efeitos fixos e outra aos efeitos aleatórios. A soma da função densidade de probabilidade de cada parte é a função de densidade de probabilidade de cada observação, e a maximização da função de densidade de probabilidade dos efeitos aleatórios, em relação aos componentes de variâncias, elimina o viés resultante da perda de graus de liberdade na estimação dos efeitos fixos do modelo. Além disso, os estimadores dos componentes de variância de máxima verossimilhança restrita não são formas explícitas, ou seja, o estimador de cada componente só pode ser encontrado por meio de métodos iterativos, pois estão em função dos estimadores dos outros componentes. (Camarinha Filho, 2003).

Nos capítulos seguintes, iremos mostrar como essa metodologia pode ser aplicada aos dados do ENEM 2014 e os resultados obtidos.

3 Resultados

Para este trabalho, selecionamos as seguintes variáveis explicativas do nível 1: idade (em anos completos), sexo (Masculino e Feminino), escolaridade da mãe (Em classes de escolaridade) e classe de *status* socioeconômico (A, B, C ou D). A escolaridade da mãe foi escolhida em detrimento à do pai pois apresentava um menor número de respostas na categoria "Não Sei". Como variável do nível 2, escola, escolhemos o tipo de escola em que o aluno estudou a maior parte do ensino médio, classificada em "privada" ou "pública". Vale ressaltar que a categoria "Pública" da variável de escola, agrega diferentes dependências administrativas, Federal, Estadual ou Municipal, diversidade que será ignorada no nosso trabalho. Além disso, vamos considerar o efeito de um possível terceiro nível, cidade, como homogêneo. Esse nível não será considerado uma vez que as informações pertinentes à escola de cada candidato apresentaram um alto índice de não resposta e as provas do ENEM não são realizadas em todas as cidades do Estado.

Tabela 1 Medidas de Frequências e Comparação de Notas

Característica	Níveis	Quantidade de Alunos	Nota Média	Diferença para Referência
Tipo de Escola	Privada	70.944	598,24	Referência
	Pública	446.169	505,63	-92,61
Sexo	Feminino	298.801	512,18	Referência
	Masculino	218.312	526,76	14,58
Classe Socioeconômica	A	14.736	631,08	Referência
	B	127.904	553,61	-77,47
	C	293.003	507,09	-123,99
	D	81.470	483,01	-148,07
Nível de Escolaridade da Mãe	Não estudou	22.429	471,40	Referência
	Da 1ª à 4ª série do Ensino Fundamental (antigo primário)	159.925	494,86	23,46
	Da 5ª à 8ª série do Ensino Fundamental (antigo ginásio)	95.284	505,22	33,82
	Ensino Médio (antigo 2º grau) incompleto	30.005	521,92	50,52
	Ensino Médio (antigo 2º grau)	110.129	532,57	61,17
	Ensino Superior incomplete	14.154	557,63	86,23
	Ensino Superior	46.244	571,67	100,27
	Pós-graduação	26.234	588,09	116,69
	Não sei	12.709	481,39	9,99
Total	-	517.113	518,34	-

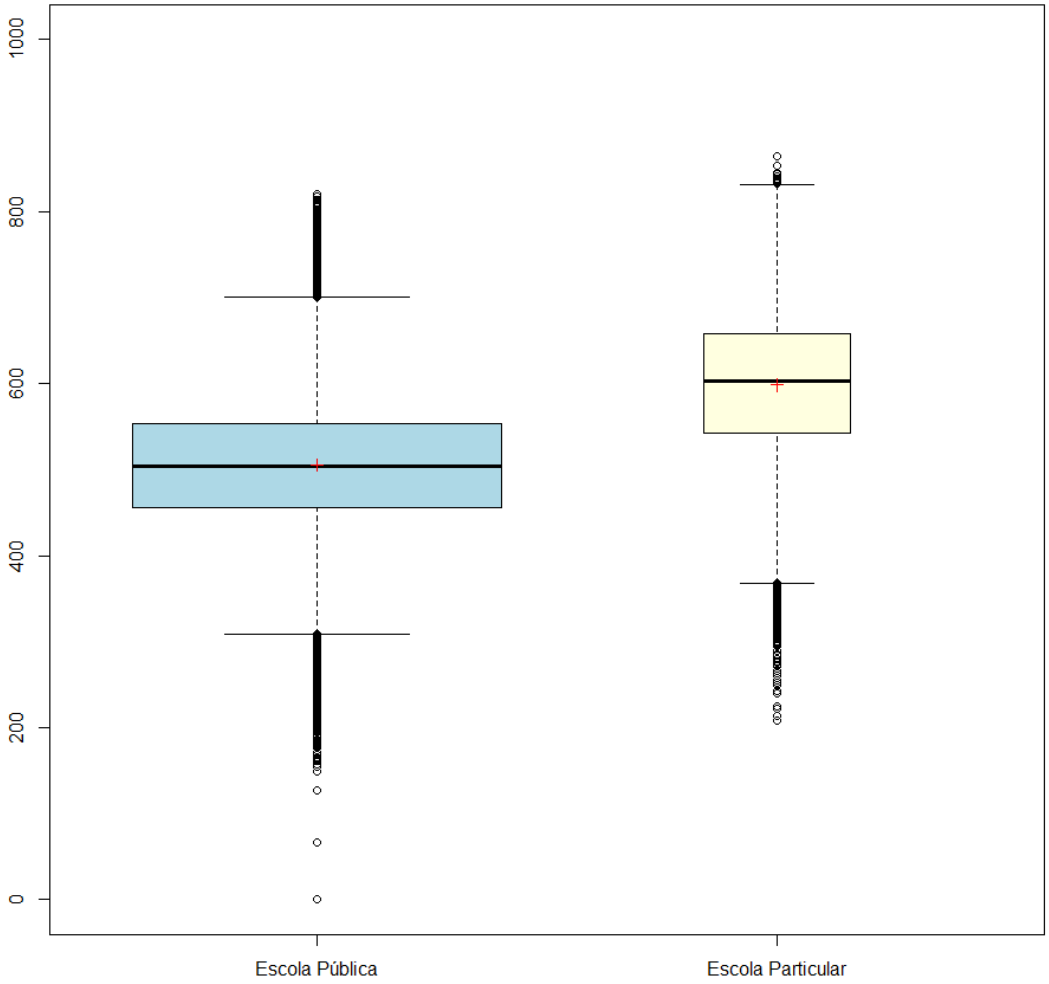
Fonte: Do Autor

Em relação às diferenças observadas entre escolas públicas e privadas, conseguimos perceber a partir da Tabela 1 que os percentuais de alunos que frequentaram escolas privadas, representam apenas 13,72% do total, e em relação à média das notas, percebemos que elas mostraram diferença de 92,61 pontos no total. Para entendermos a relevância dessa diferença, analisamos os pontos de corte do Sistema de Seleção Unificada (SISU) de todos os cursos da Universidade Federal de Juiz de Fora (UFJF) – Campus Juiz de Fora – em 2015 na chamada regular e na lista de espera.

Suponhamos, por exemplo, que um aluno de escola pública tenha a nota do ENEM igual à média das notas dos alunos de escola pública, 505,63 pontos, e que outro aluno tenha nota igual a 598,24, nota média dos alunos de escolas particulares. O aluno de escola pública seria aprovado em apenas um curso na lista de espera para o segundo semestre na instituição, o curso de Letras – Libras que tem ponto de corte 465,68. Na mesma análise, verificamos que um aluno com média de 598,24 pontos seria aprovado também em um curso, porém já na chamada regular e obteria aprovação em mais um no primeiro semestre, os cursos do Bacharelado Interdisciplinar em Ciências Humanas (566,14) e Matemática (566,96). Além disso, no segundo semestre ele seria aprovado em 18 cursos diferentes.

Sob outra ótica, podemos olhar como se dividem as notas em Minas Gerais. O resultado está na Figura 4.

Figura 4 Distribuição das Notas Separadas por Tipo de Escola



Fonte: Do Autor

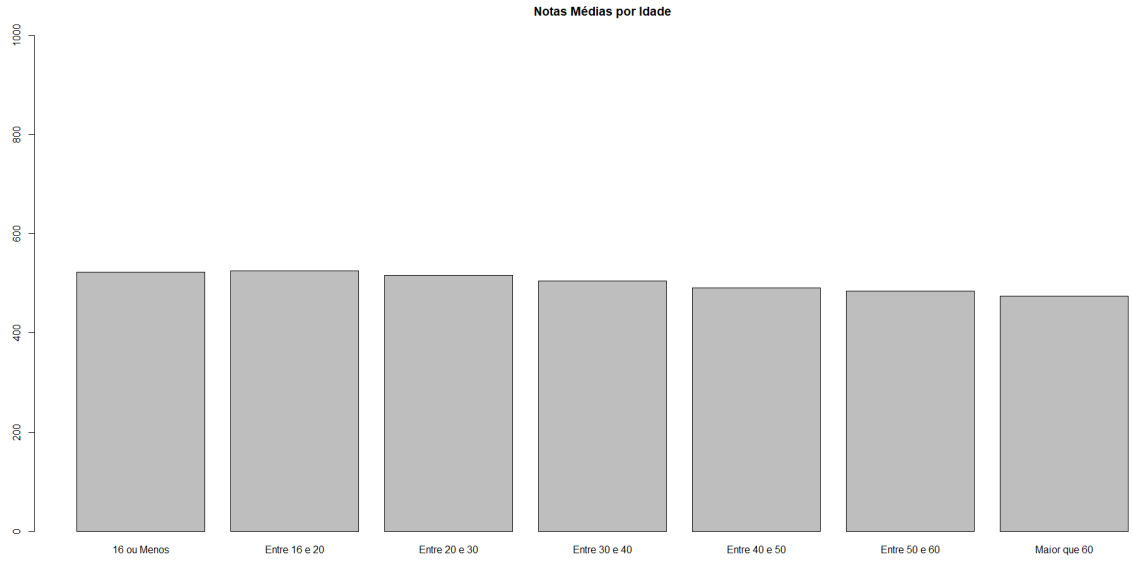
O gráfico da Figura 4 mostra os quartis das notas no estado e a faixa em que estariam inseridos os alunos de escola pública e escola privada. É possível notar que existem 3 faixas de candidatos entre eles e analisando os dados, chegamos ao resultado de que a diferença de 92 pontos no ano analisado, resultaria em uma perda de 196.755 posições na classificação do ENEM, isso representa aproximadamente 38% dos candidatos do exame. Segundo notícia vinculada em 19/01/2015 pelo portal G1 utilizando dados publicados pelo MEC, no SISU de 2015 as universidades federais de MG disponibilizaram 24.900 oportunidades, ou seja, se cada aluno se inscrevesse para uma delas o aluno médio da escola Pública não teria nenhuma vaga à sua disposição.

Ainda com os dados da tabela 1, conseguimos ver que o comportamento das notas para a variável indicadora de *status* socioeconômico analisada está de acordo com o observado na literatura. Segundo Tratvitzki (2013), a renda familiar influencia diretamente no desempenho dos alunos, chegando a apresentar diferenças de mais de 100 pontos, explicando 7% da variação das notas dos candidatos. Na tabela 1, temos, por exemplo, que a classe C apresenta menor nota média que as classes A e B que, juntas, contém apenas 27,3% dos alunos. A diferença da classe C para a classe A chega a ser maior que 120 pontos no exame. Realizando a mesma análise em relação às vagas na UFJF, obtivemos a diferença de aprovação que foi de apenas 1 curso para ingresso o segundo semestre para um aluno com a média de 507,09 pontos pertencente à classe de renda D contra 15 cursos aprovados na chamada regular para o primeiro semestre, 29 na lista de espera para o mesmo semestre e 26 para o segundo semestre de um aluno com a média 631,08 pertencente à classe de renda A.

Essa característica corrobora um estudo feito por Menezes Filho (2007) em que as variáveis relativas às características familiares dos alunos, tais como a escolaridade da mãe do candidato, tiveram a maior influência no seu desempenho. Através das análises exploratórias conseguimos perceber que as notas médias têm relação positiva com esse critério: uma vez que a escolaridade da mãe cresce, as notas chegam a aumentar até 18,04 pontos na diferença entre ensino médio incompleto e ensino fundamental completo e, se comparadas à categoria "Não Estudou" que temos como referência, a diferença chega a 116,99 pontos para a mãe que tem pós-graduação.

Analisando as características pessoais dos estudantes conseguimos avaliar a diferença de notas entre os dois sexos. Embora sejam a maioria no exame, as mulheres apresentam nota média inferior com diferença de apenas 14,58 pontos. Essa diferença não representa nenhuma aprovação a mais no SISU 2015. Já em relação às idades dos candidatos, é possível verificar que há uma relação entre idade e nota média. Quanto mais distante da faixa etária correta para a realização do exame, entre 17 e 18 anos, pior será o desempenho do aluno. Esse resultado está associado à explicação de Tratvitzki (2013), que avalia que, a defasagem escolar é uma importante influência negativa para o desempenho do aluno e pode ser observador na Figura 5.

Figura 5 Gráfico de Barras Separadas por Faixas Etárias



Fonte: Do Autor

Embora a diferença entre as médias das classes de idades seja sutil, conseguimos perceber que quanto mais ela se afasta da faixa entre 16 e 20 anos, pior é a nota média.

3.1 Análise do Modelo Proposto

De acordo com o objetivo do trabalho, pretendemos selecionar um modelo final que nos permita avaliar o impacto que o tipo de escola predominante no ensino médio tem na nota do aluno. Para Finch, Bolin e Kelley (2014) devemos começar do modelo mais simples e adicionar as variáveis de efeito fixo e aleatório um a uma realizando o teste ANOVA a fim de verificarmos se o modelo com a nova variável tem maior capacidade preditiva que o anterior. Como as variâncias são heterocedásticas, aplicamos uma correção ditando a heterocedasticidade dentro dos grupos. Com isso, começamos com o modelo abaixo:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

Em que: $\beta_{0j} = \gamma_{00} + \gamma_{01}(Tipo\ de\ Escola_j) + u_{0j}$

Após as adições dos efeitos fixos, construímos um modelo de regressão linear para explicar a variação das notas, utilizando como fatores de controle o sexo, a escolaridade da mãe, a idade e a renda dos candidatos.

$$Y_i = \beta_0 + \beta_1(Sexo_i) + \beta_2(Idade_i) + \beta_3(Escolaridade\ Materna_i) + \beta_4(Classe\ Econômica_i) + \varepsilon_i$$

Para verificarmos se o efeito aleatório do tipo de escola era necessário no modelo proposto, fizemos um teste ANOVA entre um modelo linear com e sem o efeito aleatório.
No teste F, utilizamos as hipóteses:

H_0: Coeficiente aleatório = 0, ou seja, γ_{01} = 0
H_1: Coeficiente da parte aleatória ≠ 0, ou seja, $\gamma_{01} \neq 0$

O resultado do valor-p do teste foi < 2.2e-16, isto é, temos evidências para rejeitar a hipótese nula e considerar, desta forma, a importância da variável aleatória do tipo de escola.
Após a seleção do modelo, ficamos com a forma final:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Sexo_{ij}) + \beta_{2j}(Idade_{ij}) + \beta_{3j}(Escolaridade\ Materna_{ij}) + \beta_{4j}(Classe\ Econômica_{ij}) + \varepsilon_{ij}$$

Em que: $\beta_{0j} = \gamma_{00} + \gamma_{01}(Tipo\ de\ Escola_j) + u_{0j}$

Os coeficientes estimados juntamente com o intervalo de confiança de 95% podem ser encontrados na tabela abaixo:

Tabela 2 Coeficientes estimados e intervalos de confiança de 95% do modelo proposto

Variáveis	Níveis	Limite Inferior	Estimado	Limite Superior
Efeitos Fixos				
	Intercepto	498,72	560,06	621,41
Sexo	Sexo Masculino	8,38	8,78	9,18
Idade	Idade	-0,18	-0,16	-0,13
Escolaridade da Mãe	Da 1ª à 4ª série do Ensino Fundamental (antigo primário)	17,84	18,87	19,90
	Da 5ª à 8ª série do Ensino Fundamental (antigo ginásio)	21,91	23,02	24,13
	Ensino Médio (antigo 2º grau) incompleto	33,42	34,73	36,04
	Ensino Médio (antigo 2º grau)	37,05	38,18	39,30
	Ensino Superior incomplete	51,10	52,69	54,29
	Ensino Superior	48,41	49,72	51,02
	Pós-graduação	56,25	57,70	59,14
	Não sei	0,28	1,87	3,46
Classe Econômica	Classe B	-35,63	-34,32	-33,01
	Classe C	-53,55	-52,16	-50,78
	Classe D	-66,10	-64,62	-63,13
Efeito Aleatório				
Tipo de Escola	Escola Pública	-15,60	-44,24	-125,43

Fonte: Do autor

Com o resultado obtido no modelo hierárquico utilizado, podemos concluir que, controlando por outras variáveis que possam influenciar na nota do aluno, o tipo da escola em que ele cursou predominantemente o ensino médio é responsável por diminuir 44,24 pontos em sua nota média se a instituição de ensino for pública.

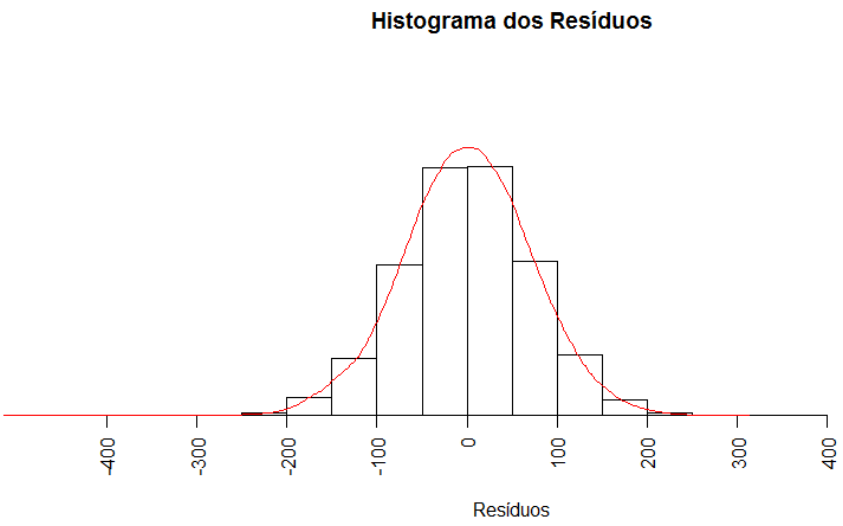
Para entendermos a dimensão da responsabilidade do tipo de escola na nota do aluno, faremos uma simulação com as taxas de alunos aprovados no SISU 2015 da UFJF – Campus Juiz de Fora. Para essa análise escolhemos as notas de corte de três cursos: Medicina (798,62); Direito (761,58) e Estatística (679,02). Com as notas regulares, apenas 19 alunos oriundos de

escolas públicas obteriam aprovação no curso de Medicina; após a adição de 44,24 pontos esse número subiria para 319. Analisando o curso de Direito, essa aprovação passaria de 227 para 1.632. Já para o curso de Estatística, o número que era de 5.901 chegou a 19.884, um aumento de mais de 300%.

3.2 Análise dos resíduos

De acordo com Cordeiro e Lima Neto (2006) as técnicas de diagnósticos devem ser utilizadas para verificar problemas com os ajustes dos modelos de regressão. Como pressuposto do modelo, devemos ter resíduos com distribuição normal, $N \sim (0, \sigma^2)$. Para assegurar essa condição e verificarmos a qualidade do ajuste do modelo, utilizamos de técnicas gráficas.

Figura 6 Histograma dos Resíduos do Modelo

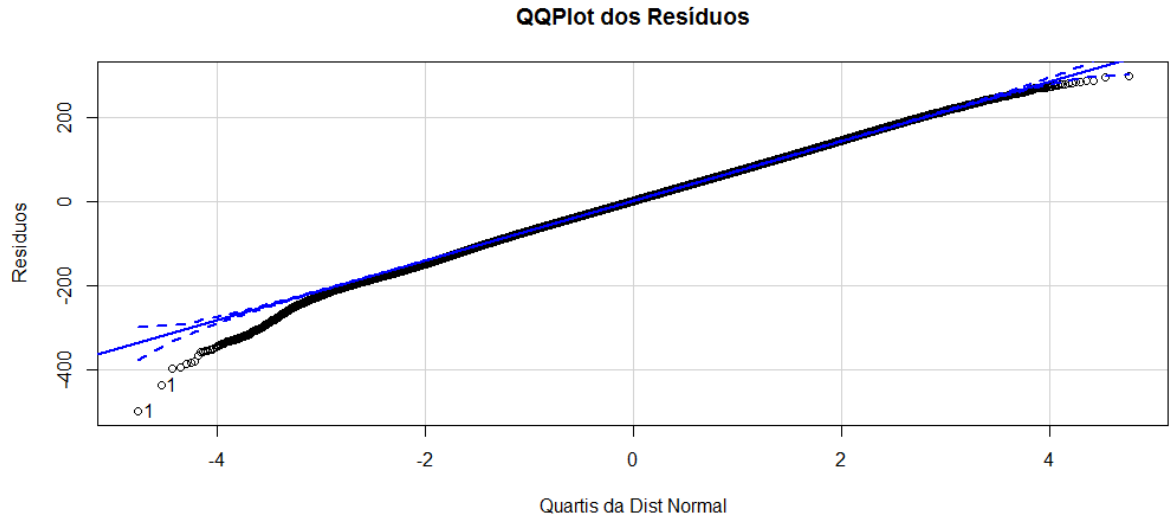


Fonte: Do autor

Com o histograma apresentado na figura 6, conseguimos ver a formação de uma curva de distribuição muito semelhante à normal, centrada no zero. Ao calcularmos o intervalo de confiança de 95% verificamos que, de fato, o valor estimado é muito próximo de zero com o $IC_{95\%}$ = [-0.197; 0.197]. Ao realizarmos o teste *t* para comparar a média dos resíduos com o valor nulo, obtivemos um valor-p de próximo de 1.

Após essa análise gráfica também fizemos a análise do QQPlot dos resíduos, construído com o auxílio do pacote *car*.

Figura 7 QQPlot dos Resíduos



Fonte: Do autor

Analisando o gráfico dos quartis da distribuição dos resíduos confrontados com os quartis da distribuição normal teórica, podemos verificar que existem alguns pontos fora da linha reta ideal, entretanto conseguimos ver que esses pontos existem em pouca quantidade, não tendo capacidade de influenciar o resultado do modelo. O restante dos dados tem uma formação do que se aproxima de uma linha reta, indicativo de que os erros são normalmente distribuídos.

Dessa forma, concluímos que o modelo está bem ajustado aos dados, respeitando os pressupostos do modelo e não apresentando nenhum problema com os resíduos.

4 Discussão

Após a análise final do modelo ajustado, podemos perceber o tamanho do impacto do tipo de escola na nota média do aluno no momento de realização do ENEM. Esse impacto chega a ser tão grande que faz a diferença entre o aluno ser ou não aprovado para uma vaga.

Com base nos resultados do Censo Escolar 2016, disponibilizado pelo MEC 28,3 mil escolas no Brasil disponibilizam o Ensino Médio. Desse total, 70,8% são públicas (68,1% Estaduais, 1,8% Federais e 0,9% Municipais) e essas escolas atendem mais de 7,08 milhões de alunos. Entretanto, essa grande massa de estudantes não compete em igualdade de condições com os alunos de escolas particulares. Segundo Vasconcelos e Lima (2004), o ensino público nas escolas do Brasil é comprometido por diversos fatores de diversas naturezas e complexidades, entre eles: deficiências de infraestrutura e de material didático; segurança precária; motivação insuficiente tanto partindo dos discentes quanto dos docentes – originadas por práticas pedagógicas inadequadas, remuneração insuficiente e desatualização dos docentes frente à novas tecnologias e metodologias de ensino. Esses fatores contribuem para a explicação do tamanho da discrepância observada neste trabalho, chegando a 44 pontos.

Não conseguimos, neste trabalho, solucionar problemas tais como a multicolinearidade das variáveis, problemas de interação e heterogeneidade das variâncias. Portanto, todos os efeitos aqui obtidos têm cunho preditor e não causal, conforme orientação de Rindskopf (2013).

Como o questionário do ENEM é respondido pelos próprios participantes, e as perguntas relativas ao questionário não são obrigatórias, não pudemos levar em conta os fatores da escola e do município que o aluno estudou. Essas informações seriam interessantes uma vez que, como ressaltado por Vasconcelos e Lima (2004), eles são de extrema importância para o resultado do grupo. Essa análise seria interessante para verificarmos o que causa tamanha diferença e podermos apontar possíveis soluções, uma vez que a Lei 12.711/2012, de 29 de Agosto de 2012, institui que as reservas de vagas para alunos oriundos de escolas públicas, hoje 50%, deverá ser revista dez anos após sua aplicação.

Apêndice I – Comandos Utilizados no *software* R

```
# Para instalar o pacote utilizado
install(nlme)
library(nlme)

# Modelos Finais Construídos
modelo1= lme (nota_media ~ 1, random = ~1|TP_ESCOLA)
modelo2= lme (nota_media ~ IDADE +  EscMae + classe + TP_SEXO, random = ~1|TP_ESCOLA, weights = varIdent(form = ~1|TP_ESCOLA))
modelo3= lm(nota_media ~ IDADE +  EscMae + classe + TP_SEXO)

# Para analisarmos a importância dos componentes do modelo

anova(modelo1, modelo2)

anova(modelo3, modelo2)

# Para construirmos os intervalos de confiança para os coeficientes do modelo

intervals(modelo2) #Essa função também é pertencente ao pacote nlme

# Gráficos Construídos
#Construção dos Decis
quantile(nota_media, prob=c(0,0.10,0.20,0.30,0.40,0.50,0.60,0.70,0.80,0.90,0.100))
hist(nota_media, main = "Distribuição das Notas dos Alunos", axes = F, xlab = "Notas", ylab = "", breaks = dec, freq=TRUE)
points(602.42, 10000, col="blue", pch=4)
points(505.70, 10000, col="red", pch=4)
axis(1,at = dec, pos=0, lty=2, pch = 5, las=3)
```

```
#Análise dos Resíduos

install(car)

library(car)

re=resid(modelo2)

qqPlot(re, ylab="Residuos", xlab="Quartis da Dist Normal", main = "QQPlot dos Resíduos")

 hist(re, main= "Histograma dos Resíduos", xlab="Resíduos", ylab="", axes=F,
ylim=c(0,0.007), prob=T)

lines(density(re), col="red")

axis(1,at = c(-400,-300,-200,-100,0,100,200,300,400), pos=0, las=3)

t.test(re)  #Para encontrarmos o intervalo de confiança e testarmos se a média é 0
```

Anexo I – Critério Brasil de 2014

Tabela 1: Pontuação do Critério Brasil:

A: Posse de itens

	Itens	Quantidade de itens			
		0	1	2	3 ou +
Pontuação	Televisão em Cores	0	1	2	4
	Rádio	0	1	2	4
	Banheiro	0	4	5	7
	Automóvel	0	4	7	9
	Empregada Mensalista	0	3	4	4
	Máquina de Lavar	0	2	2	2
	Videocassete e/ou DVD	0	2	2	2
	Geladeira	0	4	4	4
	Freezer (aparelho independente ou parte da geladeira duplex)	0	2	2	2

Fonte: ABEP <http://www.abep.org/criterio-brasil>

B: Grau de Instrução da mãe	Pontos
Analfabeto / Primário Incompleto	0
Primário Incompleto / Fundamental Incompleto	1
Fundamental Completo / Médio Incompleto	2
Médio Completo /Superior Incompleto	4
Superior Completo	8

Fonte: ABEP <http://www.abep.org/criterio-brasil>

Tabela 2: Critérios de Corte de Classes Critério Brasil 2014:

Classe	Pontos
A	35 – 46
B	23 – 34
C	14 – 22
D – E	0 – 22

Fonte: ABEP < http://www.abep.org/criterio-brasil>

Referências Bibliográficas

[1] ATIKIN M., ANDERSON D. & HINDE J. *Statistical modeling in school effectiveness studies.* Journal of the Royal Statistical Society, série: A, 144. P. 148 – 161, 1981.

[2] BENETT, N. *Teaching styles and pupil progress.* Cambridge: Havard University Press, 1976.

[3] BERGAMO, G. C. Aplicação dos modelos multiníveis na análise de dados de medidas repetidas no tempo. Dissertação de Mestrado defendida na Escola Superior de Agropecuária Luiz de Queiroz – ESALQ. Universidade de São Paulo, 2002.

[4] BUSSAB, W. O. & MORETTIN, P. A., Estatística Básica, 5ª edição Editora Saraiva, 2004.

[5] CAMARINHA FILHO, J. A. Nota Metodológica sobre Modelos Lineares Mistos, Universidade Federal do Paraná (UFPR). Setembro de 2003, Disponível em: http://www.est.ufpr.br/rt/jom03a.pdf.

[6] CORDEIRO, G.M. & LIMA NETO, E. A. Modelos Paramétricos, Universidade Federal Rural de Pernambuco e Universidade Federal da Paraíba, dezembro de 2006.

[7] FINCH, W. R., BOLIN, J. E., KELLEY, K. *Multilevel Modeling Using R*, Statistics in the Social and Behavioral Sciences Series. Editora CRC Press, 2014.

[8] FOX, J. & WEISBERG, S. (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

[9] G1, http://g1.globo.com/educacao/noticia/2015/01/inscricoes-para-250-mil-vagas-do-sisu-2015-comecam-nesta-segunda.html. Acesso em 14-06-2018.

[10] HOX, J. J. *Applied Multilevel Analysis*, TT-Publikaties, 1995. Disponível em: http://joophox.net/publist/amaboek.pdf.

[11] HOX, J. J. *Multilevel Modeling: When and Why*. University of Amsterdam & Utrecht University Amsterdam/Utrecht, the Netherlands, 1998. Disponível em: http://www.joophox.net/publist/whenwhy.pdf.

[12] LINDLEY, D. V. & SMITH, A. F. M. Bayes estimates for the linear model, Journal of the Royal Statical Society, série B, 34, p. 1-41, 1972.

[13] MENEZES FILHO, N. Os Determinantes do Desempenho Escolar no Brasil, Instituto Futuro Brasil, Ibmec-SP e FEA-USP, 2007.

[14] PINHEIRO, J. C.; BATES, D. M.; DEBROY, S; SARKAR, D. R Core Team (2018). nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-131.1, <URL: https://CRAN.R-project.org/package=nlme>.

[15] PINHEIRO, Sandra Maria Conceição. Modelo Linear Hierárquico: Um modelo alternativo para análise de desempenho escolar. Dissertação de Mestrado defendida na Universidade Federal de Pernambuco em Janeiro/2005.

[16] RINDSKOPF, D. *Multilevel Models in the Social and Behavioral Sciences* em *The SAGE Handbook of Multilevel Modeling,* 2013

[17] SCOTT, M. A., SHROUT, P. E., WEINBERG, S. L., *Multilevel Model Notation – Establishing the Commonalities* em *The SAGE Handbook of Multilevel Modeling*, 2013.

[18] SOEIRO, Leda &AVELLINE, Suelly. Avaliação Educacional. Porto Alegre: Sulina, 1982.

[19] TRAVITZKI, Rodrigo. ENEM: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar. 2013. Tese (Doutorado em Educação) - Faculdade de Educação, Universidade de São Paulo, São Paulo, 2013. doi:10.11606/T.48.2013.tde-28062013-162014. Acesso em: 2018-06-14.

[20] Universidade Federal de Juiz de Fora (UFJF) – Pontos de Corte do SiSu 2015. Disponíveis em *http://www.ufjf.br/cdara/sisu-2/sisu/sisu-2015/.*

[20] VASCONCELOS, S. D e LIMA, K. E. C. Inclusão Social e Acesso às Universidades Públicas: o Programa “Professores do Terceiro Milênio”, 2004.