AS DESIGUALDADES EXISTENTES NO BRASIL VISTAS NO RESULTADO DO ENEM 2017

Magno Tairone de Freitas Severino[1 2] – *magno@ime.usp.br* – www.ime.usp.br/∼magno

[1]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP, Brasil.

[2]O presente trabalho foi realizado com apoio da CAPES - Código de Financiamento 001.

1 Introdução

A principal forma de ingresso no ensino superior atualmente no Brasil é através do Exame Nacional do Ensino Médio, o ENEM. Realizado anualmente, tem duração de dois dias, requer uma redação e 180 questões objetivas, divididas em quatro grandes áreas: Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Linguagens, Códigos e suas Tecnologias e Matemática e suas Tecnologias.

Os dados do exame são de acesso público através do portal do Ministério da Educação, veja (INEP, 2018). Ao relacionar a nota do candidato e seus dados socioeconômicos, o principal objetivo deste trabalho é investigar se a desigualdade social que é bastante presente no cotidiano brasileiro se reflete nas notas dos inscritos no ENEM. Um outro proposito é identificar a existência de padrão espacial no desempenho médio dos alunos residentes em Minas Gerais.

2 Motivação

De acordo com (UNITED NATIONS DEVELOPMENT PROGRAMME, 2017), o Brasil figura no décimo lugar no ranking dos países mais desiguais do mundo. A desigualdade pode ser vista em aspectos de gênero, raça, idade, renda, entre outros. Sendo assim, é razoável suspeitar que tais fatores sejam relevantes em relação à performance dos candidatos do ENEM.

Em 2017, 6.731.342 candidatos se inscreveram. Associados à cada um existem 139 variáveis explicativas, vamos considerar apenas as sociais (36 no total). A Síntese de Indicadores Sociais (ver COORDENAÇÃO DE POPULAÇÃO E INDICADORES SOCIAIS DO IBGE, 2018) e a análise exploratória deste conjunto de dados são as principais motivações para a busca de um modelo que ajuste a performance de um aluno do ENEM, veja Figura 1.

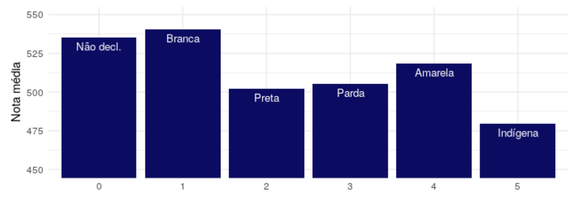

Figura 1: Nota média dos candidatos de acordo com a raça declarada.

3 Regressão linear múltipla

Do total de 6,7 milhões de inscritos, desconsideramos aqueles que foram eliminados do exame, restando 4.426.755. Para estes dados, considere o modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, em que $\mathbf{Y} = (y_1, \dots y_n)^\top$ é o vetor de notas dos candidatos, $\mathbf{X}$ é a matriz $n \times p$ com as variáveis explicativas, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top$ é o vetor de coeficientes do modelo e $\epsilon = (\epsilon_1, \dots \epsilon_n)^\top$ é o vetor com erros aleatórios, sob a suposição de que ϵ_i's são independentes e identicamente distribuídos $N(0, \sigma^2)$.

O modelo foi ajustado utilizando o *software* R (R CORE TEAM, 2016). Considerando todos os níveis das variáveis categóricas, temos 131 coeficientes obtidos para o modelo, apenas alguns não foram significativos ao nível de 5% de significância.

A Tabela 1 mostra os coeficientes estimados pelo método de máxima verossimilhança para o tipo de administração e localização da escola do candidato, a forma de ensino e a raça declarada. Além disso, observamos que o intercepto é estimado em 519,01, valor próximo da média amostral (518,55).

Coeficiente	Estimativa	Erro padrão	p-valor
Intercepto	519.00977	21.40007	$< 2 \times 10^{-16}$
Administração da escola: Federal (base)	-	-	-
Administração da escola: Estadual	-72.48417	0.32792	$< 2 \times 10^{-16}$
Administração da escola: Municipal	-65.09012	0.63228	$< 2 \times 10^{-16}$
Administração da escola: Privada	-47.00081	14.55794	0,001244
Localização da escola: Urbana (base)	-	-	-
Localização da escola: Rural	-9.83936	0.32955	$< 2 \times 10^{-16}$
Ensino: Regular (base)	-	-	-
Ensino: Especial	-19.08266	0.83596	$< 2 \times 10^{-16}$
Ensino: Jovens e Adultos	-30.00682	0.23325	$< 2 \times 10^{-16}$
Raça: não declarado (base)	-	-	-
Raça: Branca	1.51055	0.42283	0,000354
Raça: Preta	-5.84954	0.44450	$< 2 \times 10^{-16}$
Raça: Parda	-4.85479	0.42154	$< 2 \times 10^{-16}$
Raça: Amarela	1.58823	0.54867	0,003796
Raça: Indígena	-14.92869	0.81091	$< 2 \times 10^{-16}$

Tabela 1: Estimativas de alguns coeficientes do modelo, com respectivos erro padrão e nível de significância.

As Figuras 2 e 3 mostram os coeficientes estimados relacionados à escolaridade dos pais do candidato e a renda mensal familiar. Além desses, O coeficiente relacionado ao número de pessoas moram na residência é −2,567. O coeficiente obtido para o fator que discrimina gênero não foi significativo à 5%.

O tipo de escola que o candidato frequentou o Ensino Médio é significativo; categorias e coeficientes estimados: somente em escola pública (base), parte em escola pública e parte em escola privada *sem* bolsa de estudo integral (4,451), parte em escola pública e parte em escola privada *com* bolsa de estudo integral (13.272), somente em escola privada *sem* bolsa de estudo integral (28.083), somente em escola privada *com* bolsa de estudo integral (35.070)

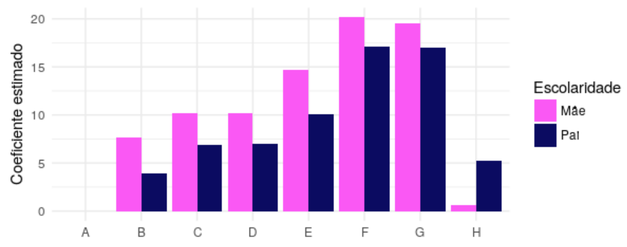

Figura 2: Níveis de escolaridade dos pais: A (Nunca estudou), B (Não completou o 5º ano do EF), C (5º ano, 9º ano do EF incompleto), D (9º ano, EM incompleto), E (EM, faculdade incompleta), F (Faculdade, pós-graduação incompleta), G (Completou a Pós-graduação), H (Não sei).

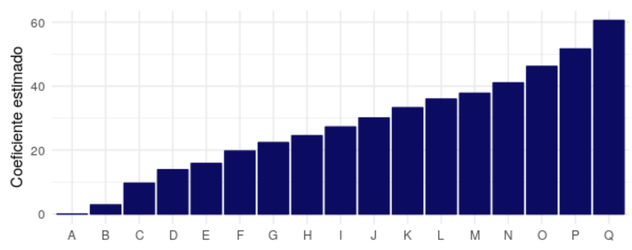

Figura 3: Renda mensal familiar, partindo de nenhuma renda (A), 1 salário (B), e acréscimos de meio salário em cada nível até 20 salários (Q).

Em conclusão, para este modelo, temos que $R^2 = 0{,}3882$, um valor que pode ser considerado baixo. Embora as variáveis presentes no banco de dados são significativas, elas não são suficientes para explicar a performance de cada candidato no ENEM. Vários outros fatores, que não estão descritos no conjunto, influenciam na nota.

4 Modelo espacial

Para este modelo, vamos considerar a performance média dos candidatos residentes no estado de Minas Gerais, agregando pelo município de residência todos os inscritos que compareceram às todas provas de 2017 e calculando a nota final média em cada unidade municipal.

A Figura 4 mostra o resultado obtido com a agregação dos dados. Quanto mais forte o tom de azul, maior é a nota média registrada em 2017 para aquele município. O mapa sugere uma associação espacial do desempenho médio dos candidatos entre as cidades mineiras, que se assemelha com o padrão espacial apresentado pelo IDH municipal do estado.

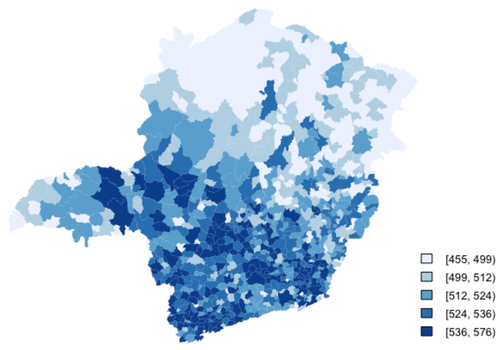

Figura 4: Nota média na prova do ENEM em cada um dos municípios de Minas Gerais

Uma medida utilizada para mensurar a associação espacial entre unidades de áreas é o C de Geary (RIPLEY, 1991),

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{\left(\sum_{i \neq j} w_{ij}\right)\sum_i (Y_i - \bar{Y})^2},$$

em que n é o número de áreas, Y_i a nota média observada na região i ($i = 1, \dots, 853$), $\bar{Y} = \sum_i Y_i/n$ e w_{ij} indica se i é vizinho de j. Para os dados considerados, $C = 0{,}845$, valor que indica associação espacial positiva.

Um modelo popular que incorpora associação espacial entre as vizinhanças é o modelo autorregressivo condicional (CAR), proposto por (BESAG, 1974). Assim, a distribuição do efeito aleatório θ_i, condicionado nos seus vizinhos, fica na forma

$$\theta_i | \theta_{j \neq i} \sim N\Big[\sum_j w_{ij}\theta_j, \tau^2/w_{i+}\Big],$$

em que $w_{i+} = \sum_j w_{ij}$, o número de vizinhos da região i.

Com isso, podemos ajustar um modelo Bayesiano hierárquico para estimar os efeitos aleatórios espaciais. Seja $\theta = (\theta_1, \dots, \theta_{853})$ o vetor de efeitos aleatórios espaciais associados à cada região. Considere então

$$\begin{aligned} Y_i \mid \theta_i &\sim N(\mu + \theta_i, \sigma^2), \\ \boldsymbol{\theta} \sim CAR(\tau^2, \rho), &\qquad \mu \sim N(500, 100), \\ \sigma^2 \sim GI(2{,}001, 1{,}001), &\qquad \tau^2 \sim GI(2{,}001, 1{,}001), \end{aligned}$$

em que μ é um escalar que representa o valor médio global da performance dos municípios, θ_i é o efeito aleatório espacial da região i, σ^2 é o efeito de variância não espacial, τ^2 é o efeito de variância espacial e ρ é o parâmetro que garante que a matriz de precisão $\boldsymbol{\Sigma}^{-1}$ seja inversível (aqui, consideramos $\rho = 0.9$). Além disso, os hiperparâmetros foram escolhidos para determinar distribuições à priori vagas. Observe que os valores α e β escolhidos para as distribuições de σ^2 e τ^2 determinam variáveis aleatórias gama-inversa com média 1 e variância 10^3.

Utilizamos o programa JAGS (PLUMMER, 2003), juntamente com o módulo GeoJAGS (FREITAS SEVERINO, 2018), para executar o algoritmo MCMC (Markov chain Monte Carlo) e amostrador de Gibbs (S. GEMAN; D. GEMAN, 1984; GELFAND; SMITH, 1990). O algoritmo foi configurado para realizar 60,000 iterações, com *burn-in* de 10,000 iterações e 5,000 observações formando a amostra a posteriori. Para evitar autocorrelação, consideramos um *lag* de 10 passos.

Parâmetros	Média	95% HDP	
μ	517,136	516,300	517,800
σ^2	126,154	101,200	152,000
τ^2	0,004	0,003	0,005

Tabela 2: Estimativas a posteriori dos parâmetros do modelo espacial: média e intervalo de credibilidade (HPD).

A Tabela 2 mostra as médias a posteriori para os parâmetros μ, σ^2 e τ^2, bem como os respectivos intervalos de credibilidade à 95%. O valor médio global μ estimado é próximo ao valor observado, 518,549. Os valores médios obtidos para os 853 efeitos aleatórios são exibidos na Figura 5, variando no intervalo $[-47{,}08, 34{,}83]$, com média 0 e mediana 2,93. Assim, o município cujo efeito aleatório espacial é negativo, tem a sua nota média no ENEM decrescida pelo valor estimado pelo efeito aleatório, ou seja, a localização do município o faz ter a nota média inferior à média global.

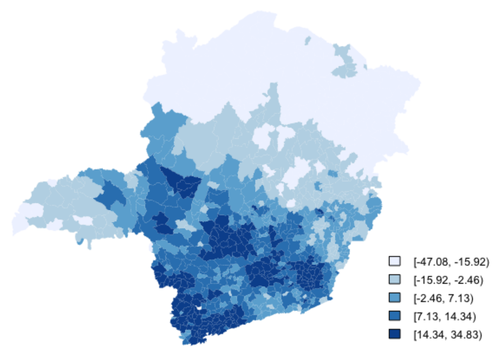

Figura 5: Média a posteriori dos efeitos aleatórios espaciais na performance dos alunos por município no estado de Minas Gerais.

Corroborando com o valor C de Geary calculado na Equação 4, o mapa obtido mostra um arranjo espacial com padrão claro no desempenho médio dos alunos no ENEM nos municípios de Minas Gerais.

5 Discussão

- Apresentação da desigualdade socioeconômica da sociedade brasileira em termos quantitativos.
- Os resultados obtidos com o modelo de regressão linear múltipla reforçam a existência da já sabida desigualdade no Brasil. As estimativas dos coeficientes mostraram que a renda familiar, a raça, o grau de instrução dos pais são significativos na nota final dos candidatos.
- O modelo espacial proposto mostra que o padrão geográfico da desigualdade presente no estado de Minas Gerais também pode ser visto através do desempenho médio dos alunos na prova do ENEM. Estamos, portanto, apenas demonstrando através dos modelos estatísticos o que foi relatado pela (COORDENAÇÃO DE POPULAÇÃO E INDICADORES SOCIAIS DO IBGE, 2018).
- Trabalho futuro: regressão quantílica para ajuste das maiores notas do exame, procurando investigar se obteremos um modelo com coeficientes diferentes.

Referências

BESAG, Julian. Spatial Interaction and the Statistical Analysis of Lattice Systems. English. *Journal of the Royal Statistical Society. Series B*, Wiley for the Royal Statistical Society, v. 36, n. 2, p. 192–236, 1974. ISSN 00359246. Disponível em: ¡http://www.jstor.org/stable/2984812¿.

COORDENAÇÃO DE POPULAÇÃO E INDICADORES SOCIAIS DO IBGE. *Síntese de indicadores sociais : uma análise das condições de vida da população brasileira 2018*. Rio de Janeiro: IBGE, 2018. ISBN 1516-3296.

FREITAS SEVERINO, Magno Tairone de. *Extending JAGS for spatial data*. 2018. Diss. (Mestrado) – Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte.

GELFAND, Alan E.; SMITH, Adrian F. M. Sampling-Based Approaches to Calculating Marginal Densities. v. 85, n. 410, p. 398–409, jun. 1990. ISSN 0162-1459 (print), 1537-274X (electronic).

GEMAN, Stuart; GEMAN, Donald. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 6, n. 6, p. 721–741, nov. 1984. ISSN 0162-8828. DOI: 10.1109/TPAMI.1984.4767596. Disponível em: ¡http://dx.doi.org/10.1109/TPAMI.1984.4767596¿.

INEP. *Microdados - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*. [S.l.: s.n.], 2018. http://portal.inep.gov.br/microdados. Acessado em: 07/12/2018.

PLUMMER, Martyn. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. [S.l.: s.n.], 2003. mcmc-jags.sourceforge.net/.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: ¡https://www.R-project.org¿.

RIPLEY, Brian D. *Statistical inference for spatial processes*. [S.l.]: Cambridge university press, 1991.

UNITED NATIONS DEVELOPMENT PROGRAMME. *Human Development Report 2016: Human Development for Everyone*. [S.l.]: United Nations, 2017. ISBN 9789211264135.