UNIVERSIDADE FEDERAL DO PARANÁ
CURSO ESTATÍSTICA

RAFAEL BUTTINI SALVIATO GRR - 20160206
IGOR HOLTRUP HORROCKS GRR - 20160230

MODELAGEM DA NOTA DE REDAÇÃO DO ENEM A PARTIR DE VARIÁVEIS DE CARÁTER SOCIAL PARA CANDIDATOS DO ESTADO DO PARANÁ

CURITIBA
Junho de 2018

Resumo

Com o objetivo de explicar a nota da prova de redação, o presente trabalho ajustou um modelo de regressão linear múltiplo com variáveis categóricas de cunho socioeconômico. Da base de dados do microdados do Inep, foram extraídas as informações socioeconômicas dos candidatos do ENEM da edição de 2016, bem como seus desempenhos na prova. Dado que fatores computacionais constrageram o escopo da análise, foi decidido utilizar o estado do Paraná para ser estudado. Também foi feito uma seleção de variáveis pelos autores, e posteriormente através do algorítmo *stepwise*. O devido diagnístico do modelo foi realizado, e não foi observado nada que comprometesse os pressupostos do modelo de regressão linear. Como resultado, o modelo não teve um coeficiente de explicação grande, entretanto as variáveis significativas corroboram com a tese de que deve-se intensificar as políticas afirmativas de justiça social e combate ao sucateamento do ensino básico no estado do Paraná.

Introdução

O Exame Nacional do Ensino Médio (ENEM) é uma das maiores avaliações educacionais do mundo. Ele é composto por cinco provas, onde quarto são objetivas e uma é discursiva (a redação). Além de servir como diagnóstico para o ensino médio do país, o ENEM também serve como exame de admissão para várias instituições de ensino superior (IES) no Brasil e no exterior (e.g: Instituto Politécnico de Coimbra, em Portugal). No que diz respeito a repercursão a nível nacional (tanto na mídia quanto nas instituições de ensino), a redação do ENEM é disparadamente a avaliação mais polêmica do exame. Além de mostrar um retrato da capacidade de argumentação dos brasileiros (as) que estão encerrando o ensino básico, os resultados dessa avaliação são determinantes para definir quais alunos (as) vão ingressar em boa parte das IES do país. Isso posto, o presente trabalho investigou as possíveis causas que melhor explicam a nota do (a) estudante na redação do ENEM. A base de dados utilizada é oriunda da edição de 2016, disponibilizada pelo Ministério da Educação (MEC) na pasta dos microdados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). A base de dados possui 166 variáveis com aproximadamente 8 milhões de observações, e o método de análise utilizado para o presente trabalho foi a análise de regressão linear. Dada certas restrições em termos de capacidade de máquina e análise, o escopo do estudo foi restringido para as pessoas que realizaram a prova no estado do Paraná. Ademais, pessoas que não preencheram alguns dados (i.e: apresentaram "non avaliable/NA" em algumas variáveis) também foram excluídas da análise. Com relação a seleção das variáveis para o modelo de regressão, algumas delas foram eliminadas em um primeiro momento por meio de avaliações subjetivas que buscaram se pautar por motivos epistemológicos inerentes a presente atividade (i.e: foram eliminadas "padrão de resposta das alternativas", "código do município", etc). Num segundo momento, outras variáveis foram excluídas por algorítimos de seleção de covariáveis. O presente trabalho foi dividido da forma que segue: a "Introdução" foi responsável por descrever a importância do trabalho, trazer luz ao problema de pesquisa e delimitar o seu escopo. A "Metodologia" ficou incubida de relatar de forma assertiva os métodos e procedimentos utilizados no presente trabalho, e a "Resultados e Discussões" foi responsável por apresentar os principais resultados e discutir as descobertas mais importantes. Por último, a "Conclusão" encerra o trabalho, destacanto suas limitações e ensaiando possíveis investigações futuras.

Material e métodos

A análise foi realizada integralmente com o software estatístico livre R. A base de dados analisada foi extraída dos microdados do Inep com informação a respeito da prova do ENEM de 2016. Esses microdados estão disponibilizados para download como arquivo compactado no formato (.zip) no próprio site do Inep. O motivo pelo qual os dados estão de maneira compactada é devido ao tamanho da base de dados (aproximadamente 6 GB). Além do exame ser nacional e a base comportar o desempenho dos alunos realizantes do teste, o microdados do Inep contém também variáveis de caráter social dos alunos (e.g: "Qual é a renda mensal de sua família?") o que torma a base extremamente grande. Como os recursos computacionais disponíveis não permitiriam analisar alunos de todo o território nacional, optou-se a limitar o estudo para apenas candidatos que realizaram a prova no estado do Paraná. Para fins de esclarecimento, a base do ENEM original foi de 8.627.367 candidatos, selecionando apenas candidatos do estado do Paraná a base ficou com 420.641 observações, e deste universo selecionado, observações com dados omissos nas variáveis de interesse foram removidos - restando apenas 32.772 candidatos.

Assim sendo, a fim de trazer mais novidades com o modelo final do trabalho, as variáveis referentes ao desempenho do canditado - por exemplo: sua nota nas provas de matemática e ciências naturais - foram desconsideradas da análise. Estas variáveis trazem ganho significativo para o R^2 do modelo ajustado por elas, entretanto, resultados como o citado acima não são nenhuma novidade (alunos que possuem um bom desempenho geral, tendem a ter um desempenho bom em redação). Por conseguinte, seguiu-se um estudo visando ajustar um modelo com variáveis de caráter socioeconômicas. Porque além de trazerem benefícios para ciêcnias sociais, também trarão resultados inovadores que vão contribuir com o ineditismo do presente trabalho.

A base de dados conta com aproximadamente 50 variáveis de cunho socioeconômico, onde boa parte delas são variáveis categóricas. Dado que: (i) o número de variáveis categóricas foi muito grande e (ii) não temos a parcimônia de um especialista na área para selecionar aquelas que fazem mais sentido para o modelo, se fez necessário reduzir este contigente de 50 variáveis a mão, e porteriormente dar o devido tratamento nelas. Por isso, antes iniciar qualquer análise preliminar, foram criadas variáveis *dummy* para cada uma destas variáveis, buscando sempre deixar o resultado o mais interpretável possível (i.e.: interpretação do β_0).

Isso posto, a calibração do modelo se deu da seguinte forma: primeiramente foi ajustado um modelo com todas as variáveis, onde se observou um $R^2 \approx 0,3$ para o ajuste. Sequencialmente, foi aplicado o algoritmo *stepwise* para seleção de variáveis no modelo (utilizando o AIC como critério de não exclusão). Como ainda restaram algumas variáveis que, pelo teste F não foram significativas, removemos as mesmas e ajustamos um novo modelo para ser diagnosticado e interpretado. Entretanto, ao removermos tais variáveis e fazer este novo ajuste, verificamos que novas variáveis não significativas apareceram no modelo - e isto motivou um segundo *stepwise* que deu origem ao modelo final.

Resultados e discussões

A tabela abaixo mostra a relação dos coeficientes estimados para o modelo de regressão ajustado final. O intercepto representa a média de nota na redação para aqueles que: não moram na capital, são do sexo feminino, não estão inscritos para se formar no ensino médio, escolheram a língua inglesa como opção de língua estrangeira, possuem renda mensal da familia acima de R$ 1770,00, não começaram atividade remunerada antes dos 18 anos, são brasileiros nato, se formaram em escola particular, têm nacionalidade brasileira e suas escolas se situam no meio urbano. E ainda, o intercepto é a média de nota sem considerar o efeito da idade e da quantidade de pessoas que moram na mesma residência que a candidata. Para as variáveis *dummy*, se a resposta for "sim", recebe o valor de 1 e se for "não", recebe o valor de 0:

	Variável	Descrição	Valor β
1	(Intercept)	Vide descrição do intercepto no primeiro parágrafo	626,5
2	ID_CAPITAL	Mora na capital? (sem considerar RMC)	15,9
3	NU_IDADE	Qual a idade do candidato?	-1,9
4	TP_SEXO	Sexo masculino?	-31,9
5	IN_CERTIFICADO	Está incrito para se formar no Ensino Médio?	-13,2
6	TP_LINGUA	Opção de língua estrangeira	-20,1
7	Q005	Número de moradores na residência	-2,7
8	Q006	Renda mensal da família: Acima de R$ 1770,00?	14,1
9	Q027	Começou atividade remunerada abaixo de 18 anos?	-9,2
10	Q049_DIURNO	Estudou no período diurno?	23,4
11	Q049_NOTURNO	Estudou no período noturno?	-10,7
12	Raca_Parda_ou_Preta	Se autodeclara negro ou pardo?	-7,2
13	Nasc_Bras_Natura	É naturalizado brasileiro ou nasceu no exterior?	-18,2
14	Nasc_Estrangeiro.a.	Nacionalidade estrangeira?	-26,2
15	TP_Publica	Esudou em escola pública?	-49,4
16	Educ_p.Jovens_e_Adultos	Estudou em escola para jovens e adultos?	-15,5
17	Loc_Esc_Rural	A escola estava situada no meio rural?	-18,5

Sobre as variáveis que apresentaram uma relação positiva com o valor esperado da nota na redação, as maiores - e únicas - foram (respectivamente): estudou no período diurno (23,4), mora na capital (15,9) e renda mensal familiar acima de R$ 1700,00 (14,1). E sobre as variáveis que apresentaram uma relação negativa com a variável resposta, as menores foram (respectivamente): estudou em escola pública (-49,4), é do sexo masculino (-31,9) e se é de nacionalidade estrangeira (-26,2). O coeficiente de explicação do modelo (R^2) foi de 13,6 %. Na sequencia, será feito o devido diagnóstico do modelo:

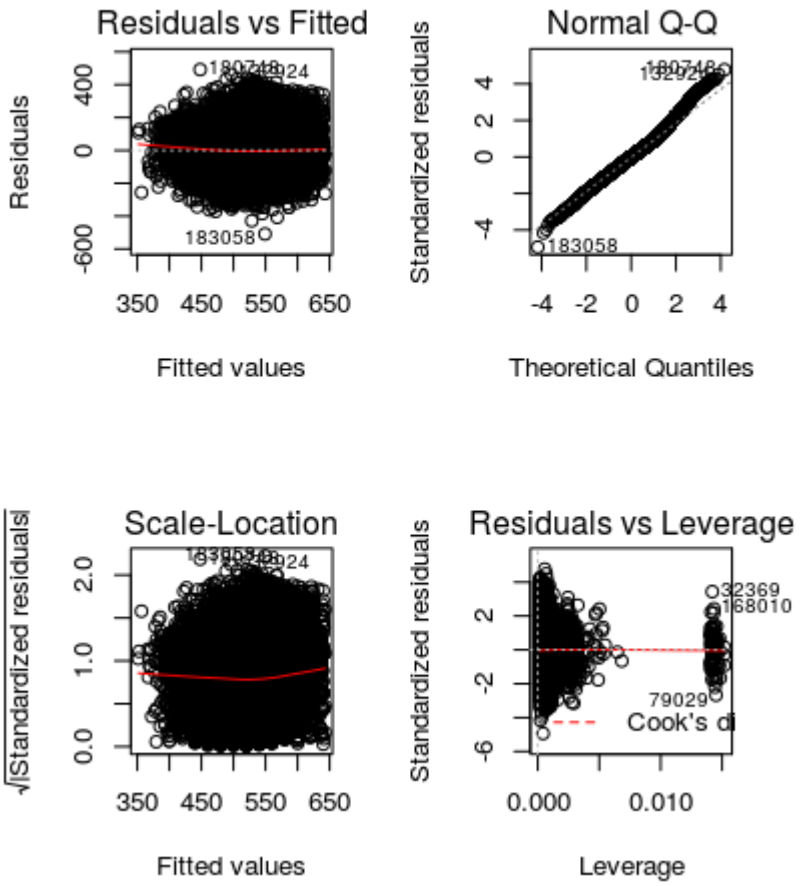

Figura 1: Plots da análise de diagnóstico

De fato, visualmente, não foi observado heteroscedasticidade nos resíduos. Com base no *qqplot*, verifica-se que os resíduos possuem distribuição normal, entretanto, foi observado caudas pesadas - de modo moderado. Ademais, não foi observado nenhum ponto de alavancagem. E, também nada que seja muito expressivo no que diz respeito à presença de *outliers* (i.e.: os únicos candidatos que se aproximaram a essa categoria, foram os pontos 183058 e 187048, que constam no gráfico *Residuals vs Fitted*). Na sequência, será exposto o resultado para verificar multicolinearidade através do *variance inflation factor* (vif's):

	Nome da variável	VIF
1	ID_CAPITAL	1.03
2	NU_IDADE	1.80
3	TP_SEXO	1.05
4	IN_CERTIFICADO	1.38
5	TP_LINGUA	1.04
6	Q005	1.03
7	Q006	1.10
8	Q027	1.08
9	Q049_DIURNO	1.44
10	Q049_NOTURNO	1.37
11	Raca_Parda_ou_Preta	1.02
12	Nasc_Bras_Naturalizado.a..Nasc_no_Exterior	1.01
13	Nasc_Estrangeiro.a.	1.00
14	TP_Publica	1.06
15	Educ_p.Jovens_e_Adultos	1.85
16	Loc_Esc_Rural	1.01

Conforme observado, nenhum valor está acima de 10. Logo, não foi observada multicolinearidade.

Conclusão

O presente trabalho teve como objetivo verificar o impacto que variáveis socioeconômicas têm no desempenho dos candidatos do ENEM na prova de redação. A restrição que os recursos computacionais trouxeram para o trabalho, impediram de se utilizar todo o território nacional, bem como avaliar todas as variáveis do bando de dados disponibilizado pelo Inep. Sobre o modelo, pesar do baixo coeficiente de explicação que o modelo apresentado demonstrou, foi satisfatório verificar resultados que corroboram com a tese de justiça social.

O ensino básico do sistema público em todas as suas esferas demonstrou um baixíssimo desempenho frente ao desempenho das escolas privadas. Isso mostra de fato um descaso com o ensino básico por parte das autoridades competentes. Ademais, foi observado também que, as variáveis relacionadas com a renda, diretamente ou indiretamente, (e.g.: renda familiar, idade que começou a exercer atividade remunerada, estudar no período diurno, etc) tiveram grande importância para explicar o bom desempenho do aluno na prova de redação (i.e.: quanto maior a renda, melhor o desempenho esperado).

Como sugestão para trabalhos futuros, seria interessante extender o escopo da análise para outras unidades federativas, a fim de verificar se os mesmos padrões são observados entre regiões. E também, com o auxílio de um profissional da área de educação, tornar-se-á possível uma melhor seleção e resumo de variáveis, de modo a corroborar significativamente com a qualidade epistemológica da análise. E por último, pode-se sugerir outros modelos de regressão que não necessáriamente seja o modelo de regressão linear. Talves, um modelo para variáveis com distribuição truncada (pois a nota vai de zero a mil), ou para variáveis discretas.