Uma investigação :

Regressão Logística com estatísticas do ENEM

Davi Araujo Martins NºUSP 10337787

Resumo

Este estudo emprega a análise de regressão logística no contexto do Exame Nacional do Ensino Médio (ENEM), utilizando dados socioeconômicos para modelar o desempenho dos estudantes. Integrando a perspectiva sociológica de Pierre Bourdieu, examina-se como o capital cultural, social e econômico impacta a performance na prova e logo a trajetória educacional, contribuindo para a reprodução de desigualdades

Introdução

Regressão Logística

A regressão logística é uma técnica estatística utilizada para modelar a relação entre uma variável dependente binária (por exemplo, sucesso ou fracasso) e um conjunto de variáveis independentes. No contexto de análises socioeconômicas, a regressão logística torna-se valiosa ao examinar fenômenos complexos, como o desempenho em exames educacionais em relação a variáveis socioeconômicas, permitindo a quantificação das influências de fatores como renda, escolaridade dos pais e acesso a recursos culturais.

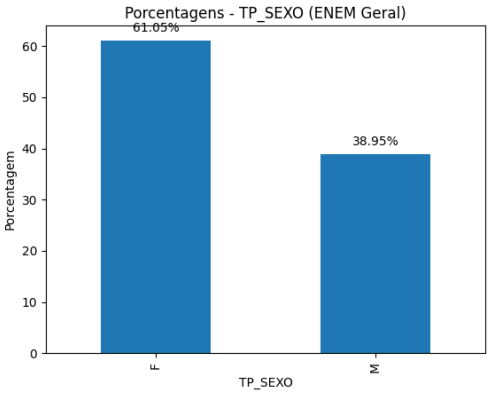
A introdução de variáveis dummy, ou variáveis indicadoras, é uma prática comum na regressão logística para lidar com categorias não métricas. Em análises socioeconômicas, essas dummies podem representar, por exemplo, diferentes estratos de escolaridade, regiões geográficas ou grupos socioeconômicos. A inclusão dessas variáveis dummy enriquece a análise, permitindo a captura de nuances e desigualdades específicas presentes nos dados. Adotando a abordagem "k-1", escolhe-se uma dummy como referência, simplificando a interpretação dos resultados ao comparar o efeito de cada categoria com a categoria de referência, contribuindo para uma compreensão mais refinada dos fatores que influenciam os resultados socioeconômicos em estudo.

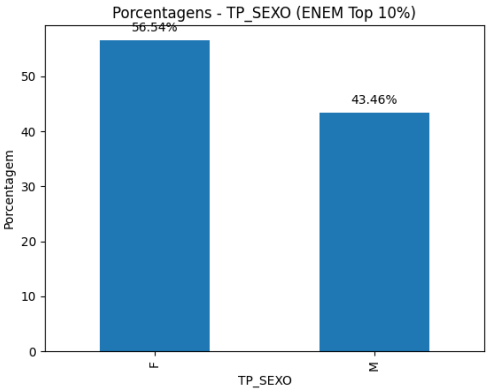
Sobre os inscritos e redivisão do dataset

Os microdados do ENEM 2020 representam por volta de 3 milhões de inscritos e tem 39 colunas com informações técnicas relacionadas a prova e dados socioeconômicos. Nesta análise foram considerados os seguintes fatores socioeconômicos: sexo, raça, escolaridade do pai, escolaridade da mãe, profissão do pai, profissão da mãe, renda familiar, tipo de escola do inscrito e acesso à internet na residência. Muitos dos dados socioeconômicos coletados pelo ENEM já sugerem herança e hereditariedade, então a escolha foi baseada em esses fatores serem mais gerais e menos específicos se comparados a por exemplo número de geladeiras ou quartos na residência. E para o fator de performance foi considerado inscritos que tiveram as 10% maiores notas gerais na prova. Originalmente este trabalho foi idealizado com os dados Fuvest e alunos da USP, talvez esses 10% representem um grupo similar.

Nas próximas seções terão gráficos desses fatores sobre o grupo geral e o grupo dos 10% e comentários sobre a redivisão desses fatores e possíveis outros comentários.

Sexo

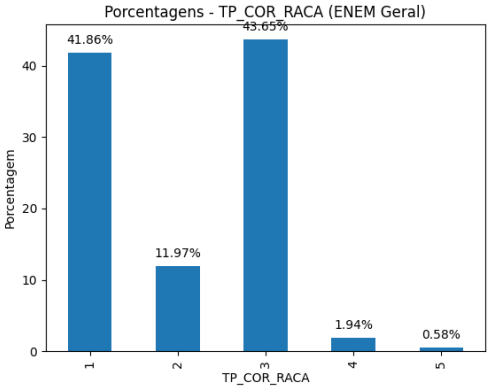


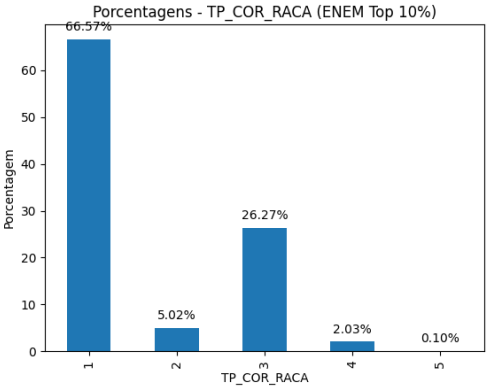


Nessa questão para regressão logística foi considerado o sexo feminino como referência, então fica como implícito e variável que aparece é o sexo masculino.

A divisão nacional dos sexos no mesmo ano é 51% feminino e 49% masculino segundo o censo 2022 - Panorama. Mesmo considerando essa pequena diferença percentual, nos dois grupos a prevalência feminina é significativa. Não cabe dentro do escopo deste trabalho questões de escolha de carreira, algo que Bourdieu aborda bastante, mas isso sugere que cursos tradicionalmente masculinos e mais difíceis de entrar não sejam significativamente assim por fatores de performance de prova.

Raça



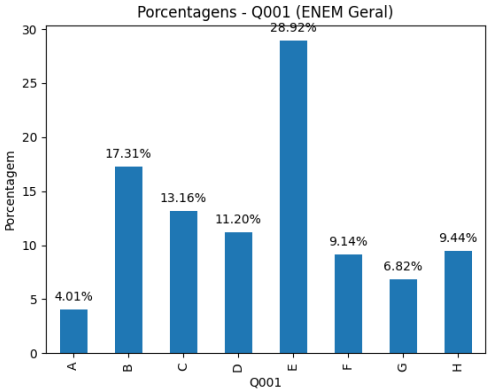


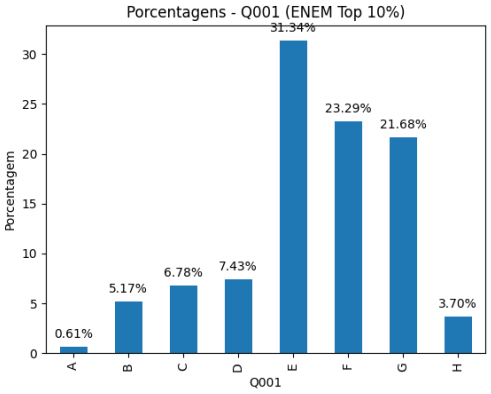
- 0 Não declarado
- 1 Branca
- 2 Preta
- 3 Parda
- 4 Amarela
- 5 Indígena
- 6 Não dispõe da informação

A legenda inclui “Não declarado” e “Não dispõe da informação”, mas na primeira tem apenas menos de 2% dos inscritos e na segunda 0%. E foi escolhido remover essas duas categorias do conjunto de dados e esse foi o único filtro aplicado. A categoria referência foi “Branca”.

Nesse caso a expectativa do senso comum se comprova, a porcentagem de brancos aumenta e de pretos e pardos diminui drasticamente do grupo geral para os top 10%. Curiosamente a porcentagem de amarelos se mantém e isso parece ter resultados que vão ser abordados posteriormente.

Escolaridade do pai - Q001 - Até que série seu pai, ou o homem responsável por você, estudou?

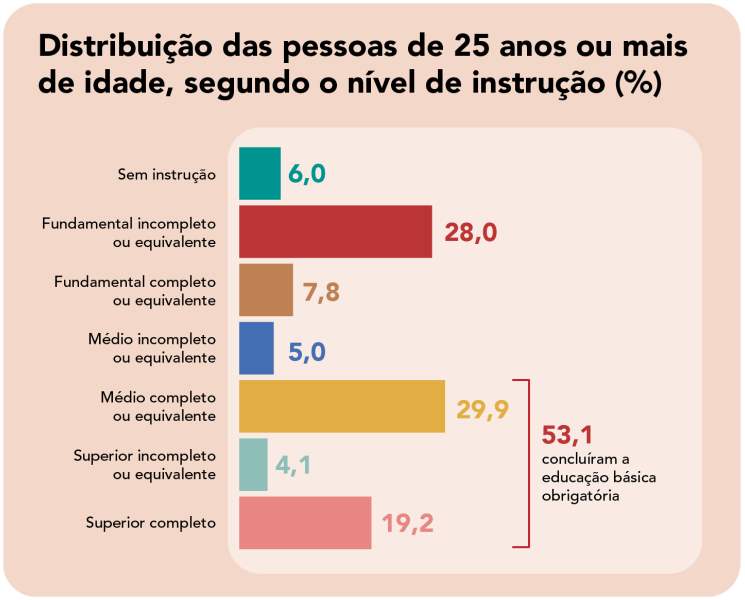




- A Nunca estudou.
- B Não completou a 4ª série/5º ano do Ensino Fundamental.
- C Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
- D Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
- E Completou o Ensino Médio, mas não completou a Faculdade.
- F Completou a Faculdade, mas não completou a Pós-graduação.
- G Completou a Pós-graduação.
- H Não sei.

Nessa questão para regressão logística foi separada entre os que tinham menos que ensino médio completo, os que tinham ensino médio completo e os que tinham mais que o médio completo sendo que a categoria referência é “Não sei”. A intenção foi dividi-los em 3 grupos de proporções similares, mas a examinação mais estratificada dos grupos com menos estudo poderia ser mais revelador.

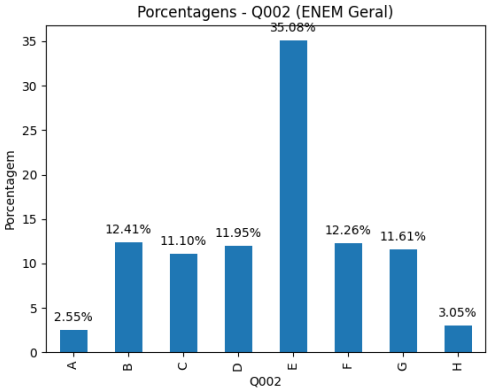
As comparações com a escolaridade da mãe são mais fáceis com a regressão logística, mas aqui vale ressaltar a diferença entre evasão escolar, considerando a soma dos grupos A e B, escolaridade no geral e as respostas "Não sei" no grupo geral. 21,32% dos pais têm menos do que o 5° ano do ensino fundamental completo enquanto 14,96% das mães partilham dessa escolaridade. A mediana das mães e dos pais é ensino médio, mas as mães têm mais ensino no geral porque é percentualmente menor nos ensinos abaixo que o ensino médio e percentualmente maior no ensino médio e nos ensinos acima. Para obter uma média de fato, teria que utilizar uma métrica como anos de estudo. Sobre as respostas "Não sei", a distribuição para os pais é de 9,44% e para as mães de 3,05%. E essa diferença existe também para as informações de profissão, 11,94% para os pais e 8,59% para as mães. Não tem como confirmar ou negar, mas abandono parental é uma hipótese que se levanta. Mas considerando que a diferença diminui de uma pergunta para outra, já sugere que seriam fatores múltiplos que explicam essas diferenças e própria diferença da evasão escolar pode ser um fator, a incerteza de quando o pai ou mãe parou de estudar.

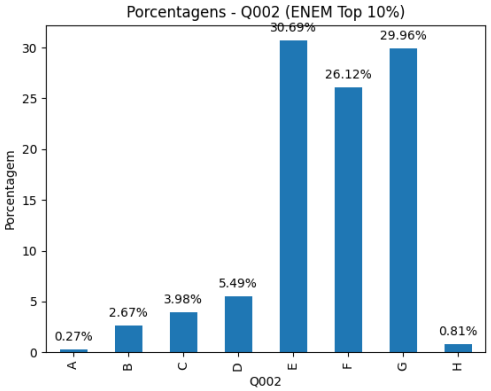


Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Pesquisas por Amostra de Domicílios, Pesquisa Nacional por Amostra de Domicílios Contínua 2022.

Comparando com a população geral brasileira, apesar de utilizar divisões diferentes, pais e mães dos inscritos do ENEM não diferem muito da população geral. Nos top 10%, a distribuição favorece inscritos com pais e mães com ensino superior completo ou mais, quase metade.

Escolaridade da mãe - Q002 - Até que série sua mãe, ou a mulher responsável por você, estudou?

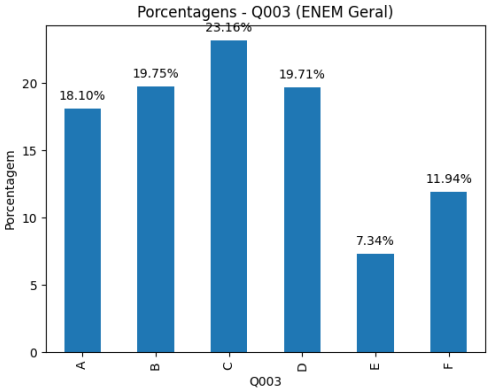


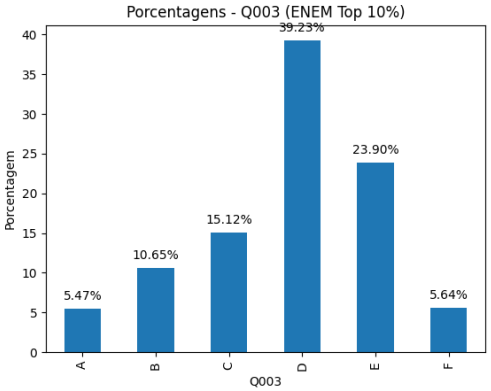


- A Nunca estudou.
- B Não completou a 4ª série/5º ano do Ensino Fundamental.
- C Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
- D Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
- E Completou o Ensino Médio, mas não completou a Faculdade.
- F Completou a Faculdade, mas não completou a Pós-graduação.
- G Completou a Pós-graduação.
- H Não sei.

Nessa questão para regressão logística foi separada entre as que tinham menos que ensino médio completo, os que tinham ensino médio completo e os que tinham mais que o médio completo sendo que a categoria referência é “Não sei”. A intenção foi dividi-los em 3 grupos de proporções similares, mas a examinação mais estratificada dos grupos com menos estudo poderia ser mais revelador.

Profissão do pai - Q003



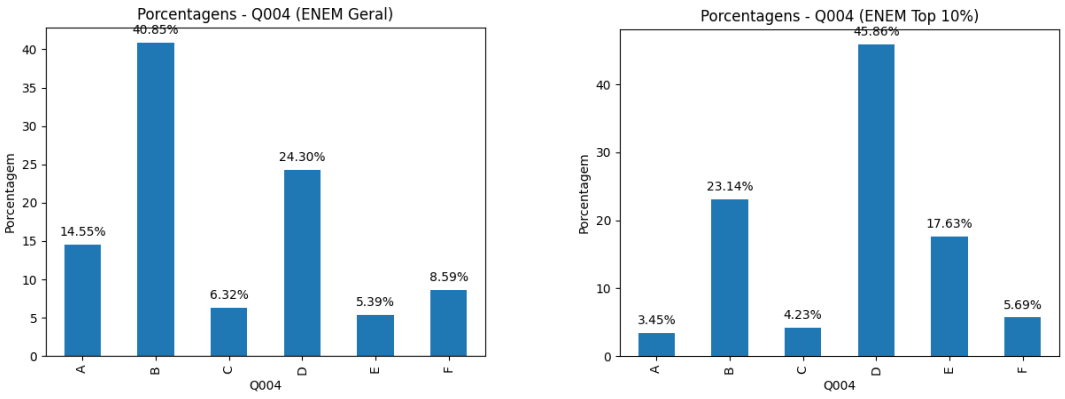


- A Grupo 1: Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista.
- B Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista 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.
- C 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.
- D 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.
- E 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.
- F Não sei.

Nessa questão para regressão logística foi separada nos mesmos grupos descritos e a categoria referência é “Não sei”.

Na população geral de inscritos existe distribuição bem próxima dos pais no grupos 1 ao 4. E nos 10%, não têm distribuições próximas e os grupos 4 e 5 são mais prevalentes.

Profissão da mãe - Q004



- A Grupo 1: Lavradora, agricultora sem empregados, bóia fria, criadora de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultora, pescadora, lenhadora, seringueira, extrativista.
- B Grupo 2: Diarista, empregada doméstica, cuidadora de idosos, babá, cozinheira (em casas particulares), motorista particular, jardineira, faxineira de empresas e prédios, vigilante, porteira,

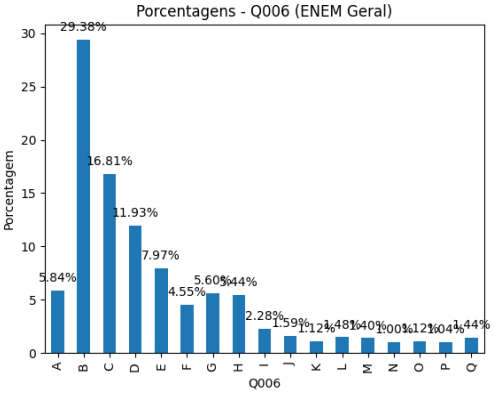
carteira, office-boy, vendedora, caixa, atendente de loja, auxiliar administrativa, recepcionista, servente de pedreiro, repositora de mercadoria.

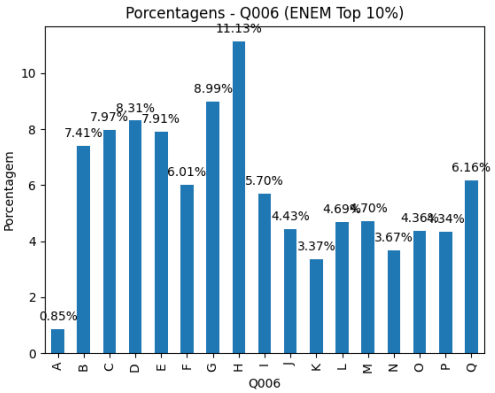
- C Grupo 3: Padeira, cozinheira industrial ou em restaurantes, sapateira, costureira, joalheira, torneira mecânica, operadora de máquinas, soldadora, operária de fábrica, trabalhadora da mineração, pedreira, pintora, eletricista, encanadora, motorista, caminhoneira, taxista.
- D Grupo 4: Professora (de ensino fundamental ou médio, idioma, música, artes etc.), técnica (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretora de imóveis, supervisora, gerente, mestre de obras, pastora, microempresária (proprietária de empresa com menos de 10 empregados), pequena comerciante, pequena proprietária de terras, trabalhadora autônoma ou por conta própria.
- E Grupo 5: Médica, engenheira, dentista, psicóloga, economista, advogada, juíza, promotora, defensora, delegada, tenente, capitã, coronel, professora universitária, diretora em empresas públicas ou privadas, política, proprietária de empresas com mais de 10 empregados.
- F Não sei.

Nessa questão para regressão logística foi separada nos mesmos grupos descritos e a categoria referência é “Não sei”.

Na população geral de inscritos tem uma prevalência de mães pertencentes ao grupo 2, que contém profissões como diarista, empregada doméstica e faxineira de empresas e prédios. E nos 10%, mães que pertencem ao grupo 4, que contém profissões como professora e técnica.

Renda familiar - Q006 - Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)



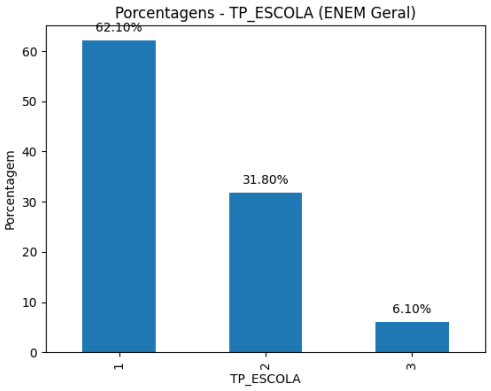


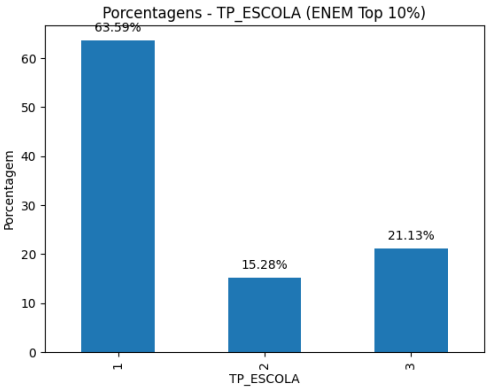
- A Nenhuma Renda
- B Até R$ 1.212,00
- C De R$ 1.212,01 até R$ 1.818,00.
- D De R$ 1.818,01 até R$ 2.424,00.
- E De R$ 2.424,01 até R$ 3.030,00.
- F De R$ 3.030,01 até R$ 3.636,00.
- G De R$ 3.636,01 até R$ 4.848,00.

- H De R$ 4.848,01 até R$ 6.060,00.
- I De R$ 6.060,01 até R$ 7.272,00.
- J De R$ 7.272,01 até R$ 8.484,00.
- K De R$ 8.484,01 até R$ 9.696,00.
- L De R$ 9.696,01 até R$ 10.908,00.
- M De R$ 10.908,01 até R$ 12.120,00.
- N De R$ 12.120,01 até R$ 14.544,00.
- O De R$ 14.544,01 até R$ 18.180,00.
- P De R$ 18.180,01 até R$ 24.240,00.
- Q Acima de R$ 24.240,00.

Nessa questão para regressão logística foi separada nos grupos de renda familiar até R$1212,00 entre R$ 1212,00 e R$ 3030,00 e acima de R$ 3030,00. A intenção era dividir em 3 grupos com mais ou menos a mesma quantidade. E a categoria referência são os que têm renda familiar acima de R$ 3030,00. A intenção foi dividi-los em 3 grupos de proporções similares, mas a examinação mais estratificada dos grupos com mais renda familiar poderia ser mais revelador e mostrar questões de rendimentos decrescentes

Tipo de escola - Tipo de escola do Ensino Médio



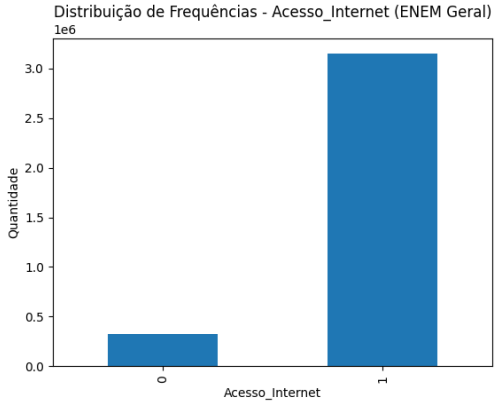


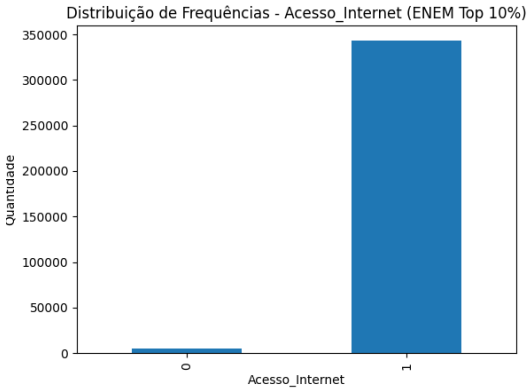
- 1 Não Respondeu
- 2 Pública
- 3 Privada

Nessa questão para regressão logística foi separada nos grupos descritos mesmos sendo que a categoria referência é “Não Respondeu”.

A taxa de “Não respondeu” é tão grande nessa questão que quase qualquer análise seria inconclusiva e o melhor seria tirar da análise, mas a questão de comparar a escola pública com a escola privada é tão grande que foi mantida.

Acesso à internet - Na sua residência tem acesso à Internet?





Geral		Top 10%	
Sim/1	90.610957%	Sim/1	98.662137%
Não/0	9.389043%	Não/0	1.337863%

A internet é uma ferramenta poderosa de informação e por isso foi a única variável mais específica escolhida. Inicialmente a suposição era que o acesso à Internet em casa seria menos prevalente, mas ainda é consideravelmente relevante não ter esse acesso em casa para performance na prova.

Redivisão do dataset

Variáveis Dependente	
Topl_Geral	Medidor de performance

Variáveis Independentes	**Categoria referência**
Sexo_Masculino	Sexo Feminino
Preta	Branca
Parda	Branca
Amarela	Branca
Indigena	Branca
Pai_Menos_que_Medio	“Não sei”
Pai_Ensino_Medio_Completo	“Não sei”

Pai_Ensino_Superior_Mais	“Não sei”
Mae_Menos_que_Medio	“Não sei”
Mae_Ensino_Medio_Completo	“Não sei”
Mae_Ensino_Superior_Mais	“Não sei”
Pai_Grupo_1	“Não sei”
Pai_Grupo_2	“Não sei”
Pai_Grupo_3	“Não sei”
Pai_Grupo_4	“Não sei”
Pai_Grupo_5	“Não sei”
Mãe_Grupo_1	“Não sei”
Mãe_Grupo_2	“Não sei”
Mãe_Grupo_3	“Não sei”
Mãe_Grupo_4	“Não sei”
Mãe_Grupo_5	“Não sei”
Ate_1212	Renda familiar acima de 3030
Entre_1212_e_3030	Renda familiar acima de 3030
Escola_Privada	“Sem Resposta”
Escola_Publica	“Sem Resposta”
Acesso_Internet	Não ter acesso

Resultados e Análise

Regressão Logística

```
Coefficients:
                          Estimate Std. Error  z value Pr(>|z|)
(Intercept)             -3.517e+00  2.943e-02 -119.493  < 2e-16 ***
Sexo_Masculino           4.240e-02  4.478e-03    9.469  < 2e-16 ***
Preta                   -6.348e-01  9.661e-03  -65.708  < 2e-16 ***
Parda                   -3.737e-01  5.128e-03  -72.885  < 2e-16 ***
Amarela                 -3.501e-02  1.592e-02   -2.199   0.0279 *
Indigena                -1.211e+00  6.109e-02  -19.826  < 2e-16 ***
Pai_Menos_que_Medio      7.519e-02  1.241e-02    6.059 1.37e-09 ***
Pai_Ensino_Medio_Completo  2.944e-01  1.213e-02   24.270  < 2e-16 ***
Pai_Ensino_Superior_Mais   5.660e-01  1.261e-02   44.893  < 2e-16 ***
Mae_Menos_que_Medio      1.786e-01  2.329e-02    7.668 1.75e-14 ***
Mae_Ensino_Medio_Completo  5.808e-01  2.282e-02   25.444  < 2e-16 ***
Mae_Ensino_Superior_Mais   8.590e-01  2.303e-02   37.296  < 2e-16 ***
Pai_Grupo_1              7.939e-03  1.374e-02    0.578   0.5635
Pai_Grupo_2              1.174e-01  1.124e-02   10.447  < 2e-16 ***
Pai_Grupo_3              1.529e-01  1.071e-02   14.277  < 2e-16 ***
Pai_Grupo_4              5.272e-01  1.021e-02   51.658  < 2e-16 ***
Pai_Grupo_5              5.603e-01  1.153e-02   48.607  < 2e-16 ***
Mae_Grupo_1              7.721e-07  1.570e-02    0.000   1.0000
Mae_Grupo_2              1.170e-01  1.006e-02   11.632  < 2e-16 ***
Mae_Grupo_3              1.268e-01  1.354e-02    9.368  < 2e-16 ***
Mae_Grupo_4              3.151e-01  9.894e-03   31.852  < 2e-16 ***
Mae_Grupo_5              3.544e-01  1.160e-02   30.550  < 2e-16 ***
Ate_1212                -1.180e+00  8.673e-03 -136.095  < 2e-16 ***
Entre_1212_e_3030       -6.400e-01  5.690e-03 -112.471  < 2e-16 ***
Escola_Publica          -4.916e-01  5.944e-03  -82.707  < 2e-16 ***
Escola_Privada           5.646e-01  6.291e-03   89.758  < 2e-16 ***
Acesso_Internet          7.310e-01  1.747e-02   41.833  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Os coeficientes de cada uma das variáveis faz parte de uma fórmula que calcula a probabilidade de um indivíduo fazer parte do grupo dos top 10%. Essa fórmula tem esse formato:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}.$$

Nenhum dos coeficientes é particularmente grande e todas as variáveis são binárias, no máximo podem ir até 1, então a comparação relativa é mais relevante. E isso é mais fácil junto com o conceito de odds ratio (razão de chances), que aparece na próxima tabela. Ainda assim, os coeficientes negativos indicam uma contribuição negativa para as chances de um inscrito fazer parte dos top 10% em relação à categoria referência. Então, por exemplo, todas as outras raças estão em desvantagem em relação à branca. E o que é de importância primordial são a significância das variáveis e seus

coeficientes na última coluna mais à direita. Quanto menor melhor, e maior que 0.05 seria melhor que fosse retirado do modelo por falta de relevância estatística. A maioria dos coeficientes são significativos, apenas alguns que não. “Amarela” tem uma significância menor, possivelmente pela porcentagem similar nos dois grupos. E pais e mães que pertencem ao grupo 1 de profissões, profissões mais ligadas a agropecuária e o meio rural, idealmente seriam retirados do modelo. Isso sugere que as profissões dos inscritos que não sabem as profissões dos pais seja similar a essa em termos de impacto na alta performance na prova.

Odds Ratio - Razão de Chances

	Predictor	Coefficient	Odds_Ratio
(Intercept)	(Intercept)	-3.516849e+00	0.02969284
Sexo_Masculino	Sexo_Masculino	4.239687e-02	1.04330846
Preta	Preta	-6.347719e-01	0.53005640
Parda	Parda	-3.737184e-01	0.68817065
Amarela	Amarela	-3.501367e-02	0.96559222
Indigena	Indigena	-1.211099e+00	0.29786978
Pai_Menos_que_Medio	Pai_Menos_que_Medio	7.519142e-02	1.07809049
Pai_Ensino_Medio_Completo	Pai_Ensino_Medio_Completo	2.943664e-01	1.34227560
Pai_Ensino_Superior_Mais	Pai_Ensino_Superior_Mais	5.659991e-01	1.76120651
Mae_Menos_que_Medio	Mae_Menos_que_Medio	1.785665e-01	1.19550234
Mae_Ensino_Medio_Completo	Mae_Ensino_Medio_Completo	5.807576e-01	1.78739198
Mae_Ensino_Superior_Mais	Mae_Ensino_Superior_Mais	8.589988e-01	2.36079597
Pai_Grupo_1	Pai_Grupo_1	7.939281e-03	1.00797088
Pai_Grupo_2	Pai_Grupo_2	1.173981e-01	1.12456704
Pai_Grupo_3	Pai_Grupo_3	1.529076e-01	1.16521735
Pai_Grupo_4	Pai_Grupo_4	5.272425e-01	1.69425402
Pai_Grupo_5	Pai_Grupo_5	5.602556e-01	1.75112006
Mae_Grupo_1	Mae_Grupo_1	7.721142e-07	1.00000077
Mae_Grupo_2	Mae_Grupo_2	1.170369e-01	1.12416089
Mae_Grupo_3	Mae_Grupo_3	1.268447e-01	1.13524076
Mae_Grupo_4	Mae_Grupo_4	3.151459e-01	1.37045930
Mae_Grupo_5	Mae_Grupo_5	3.543976e-01	1.42532185
Ate_1212	Ate_1212	-1.180349e+00	0.30717150
Entre_1212_e_3030	Entre_1212_e_3030	-6.399638e-01	0.52731150
Escola_Publica	Escola_Publica	-4.915904e-01	0.61165284
Escola_Privada	Escola_Privada	5.646328e-01	1.75880192
Acesso_Internet	Acesso_Internet	7.309540e-01	2.07706121

Uma razão de chances de 1 indica que a condição ou evento sob estudo é igualmente provável de ocorrer nos dois grupos. Maior do que 1, mais provável. E menor do que um, menos provável. Por exemplo, um inscrito preto tem 0,53 menos chances de alta performance prova que um branco, ou, um inscrito branco tem quase 2 vezes mais chances que um inscrito preto de alta performance na prova.

Comparações entre grupos de uma “mesma variável original” pode ser feita dividindo duas razões, por exemplo inscritos com pais de ensino superior completo ou mais tem 1,64 mais chances de

alta performance na prova que com pais que têm menos do que o ensino médio completo. Para as mães esse valor é de 1,98. Ou comparando os valores diretamente já que eles têm a mesma referência.

Entre grupos diversos a comparação entre os coeficientes e razões de chances são válidas, mas não são diretas. Por exemplo, a diferença entre inscritos brancos e inscritos negros pretos é similar a diferença entre ter acesso a internet em casa e não ter.

Entre os maiores coeficientes e razões de chances estão mães com o ensino superior, acesso à internet em casa, mães com o ensino médio completo, escola privada e pais do grupo 5. E entre os menores coeficientes e razões de chances estão indígenas, renda familiar até R$ 1212,00, renda familiar entre R$ 1212,00 e R$ 3030,00, pretos e escola pública.

O aumento de escolaridade das mães se prova mais relevante que o aumento de escolaridade dos pais na performance dos filhos, os coeficientes e as razões são maiores para as mães e as diferenças entre níveis de escolaridade também. O impacto de um pai com ensino superior é similar ao de uma mãe com ensino médio completo. E em relação às profissões é o "contrário", os grupos de profissões não seguem uma escala hierárquica como a escolaridade, mas ainda têm uma divisão social que parece considerar fatores como trabalho mais ou menos manual e nível de instrução requerido. Os grupos 4 e 5 são mais próximos entre si e mais distantes dos grupos 2 e 3 e do grupo referência "Não sei". E os valores são mais altos para os pais. Mas nos incritos o sexo não tem tanto impacto, o sexo masculino leva apenas uma pequena vantagem.

O acesso à internet em casa aparenta ser bem democrático entre os inscritos, então é mais que a falta de acesso crucialmente negativa.

A análise da comparação entre escola privada e escola pública é comprometida pela quantidade de falta de respostas. Mas considerando que uma escola é pública ou é privada, a escola privada é uma grande vantagem comparada a um grupo misto de escolas. Mas os números reais são incertos.

Sobre as diferenças raciais, a dominação branca é evidente. Sendo que os amarelos têm alta performance próxima e os indígenas têm a maior desvantagem nesse aspecto. E as diferenças entre pretos e pardos podem ser explicadas por colorismo ou, ao contrário, fundamentam o colorismo enquanto fenômeno social.

E sobre as diferenças de renda familiar, como mencionado anteriormente, teria sido melhor estratificar mais os grupos. Percebe-se um ganho bem significativo entre os dois primeiros extratos. Mas não tem como perceber o que seria um ponto de "neutralidade social" ou que a renda deixa de ser tão significativa para alta performance. Ou até mesmo, corroborar ou não se as anedotas do fenômeno do ENEM não ser uma prova tão relevante para os mais ricos.

	Odds_Ratio_CI_Lower	Odds_Ratio_CI_Upper
(Intercept)	0.02802849	0.03145602
Sexo_Masculino	1.03419265	1.05250461
Preta	0.52011458	0.54018826
Parda	0.68128933	0.69512148
Amarela	0.93592691	0.99619781
Indigena	0.26425846	0.33575615
Pai_Menos_que_Medio	1.05218600	1.10463275
Pai_Ensino_Medio_Completo	1.31074299	1.37456679
Pai_Ensino_Superior_Mais	1.71821947	1.80526902
Mae_Menos_que_Medio	1.14216173	1.25133404
Mae_Ensino_Medio_Completo	1.70919435	1.86916725
Mae_Ensino_Superior_Mais	2.25659625	2.46980718
Pai_Grupo_1	0.98118382	1.03548925
Pai_Grupo_2	1.10006932	1.14961030
Pai_Grupo_3	1.14101212	1.18993607
Pai_Grupo_4	1.66069860	1.72848744
Pai_Grupo_5	1.71200430	1.79112954
Mae_Grupo_1	0.96970526	1.03124278
Mae_Grupo_2	1.10220997	1.14654898
Mae_Grupo_3	1.10551030	1.16577076
Mae_Grupo_4	1.34413954	1.39729442
Mae_Grupo_5	1.39327993	1.45810066
Ate_1212	0.30199411	0.31243765
Entre_1212_e_3030	0.52146346	0.53322513
Escola_Publica	0.60456869	0.61882001
Escola_Privada	1.73725018	1.78062102
Acesso_Internet	2.00713288	2.14942583

Essa tabela mostra os intervalos de confiança das razões de chance (odds ratio).

Outras medidas de ajuste

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1758947  on 2703635  degrees of freedom
Residual deviance: 1416451  on 2703609  degrees of freedom
AIC: 1416505

Number of Fisher Scoring iterations: 7
```

Para ser um bom modelo de regressão logística, os desvios residuais teriam que ser significativamente menores que os desvios de um modelo nulo e é apenas menor.

```
> # Print the confusion matrix and metrics
> print("Confusion Matrix:")
[1] "Confusion Matrix:"
> print(conf_matrix)
      Predicted
Actual      0      1
     0 689674   5773
     1  70883   6139
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.901"
> print(paste("Error Rate:", round(error_rate, 3)))
[1] "Error Rate: 0.099"
```

A matriz de confusão indica uma boa acurácia e um erro pequeno. Mas dada a natureza de ser os top 10%, um modelo nulo teria uma acurácia similar.

```
> print(nagelkerke_r2)
$N
[1] 2703636

$R2
[1] 0.2487857
```

A interpretação do R^2 de Nagelkerke não é direta como a interpretação do R^2 em modelos lineares, em que pode ser interpretada como a proporção de variância explicada . Mas varia entre 0 e 1 e mais próximo de 1 indica um melhor ajuste do modelo preditivo aos dados. O R^2 de Nagelkerke desse modelo foi 0,24, o que indica um ajuste ruim do modelo.

No geral as medidas de ajuste apontam que o modelo não tem um alto poder de predição e que mais fatores explicam o fenômeno como um todo. Mas as relações encontradas entre a variável dependente, ser parte dos inscritos com top 10% da nota geral, e as variáveis independentes, os fatores socioeconômicos, são significativas e estatisticamente relevantes. E logo as relações entre os fatores socioeconômicos entre si considerando esse contexto dos top 10% também são.

Conclusão

Como mencionado anteriormente, este trabalho era inicialmente planejado para utilizar dados da Fuvest, porém a forma que os dados são disponibilizados não permite muita manipulação e recortes. Então esse foi o primeiro problema encontrado, o que diz um pouco sobre transparência e as execuções da lei de acesso à informação e importância para fazer ciência e democracia.

Passado a coleta de dados, tem a escolha dos dados que serão efetivamente utilizados. Nessa parte tem um pouco de expectativas informadas, mas no fundo, pelo menos para um primeiro trabalho, é bem investigativo. E os resultados corroboram ou não com as escolhas dos indicadores sociais, mas tendem sugerir diversas iterações de mudança de escopo e/ou indicadores para obter informações mais relevantes.

E por fim, a parte da análise é o ápice do trabalho e a parte mais difícil. Os números informam correlações e não causalidades e mesmo essas correlações são numéricas e não contém informações “culturais”. Por exemplo, foi encontrada uma relação mais forte entre a escolaridade das mães e alta

performance na prova do que com a dos pais e isso tem razões culturais. Provavelmente pelas mesmas razões culturais que o Bolsa Família prioriza em certos contextos a mãe em relação ao pai para ser o veículo do benefício. Mas para fazer afirmações audaciosas sobre o que são essas razões como o Bourdieu faz sobre outros assuntos, requer mais propriedade do que eu sinto possuir no momento. Ainda assim "miséria de condição" é perceptível pela pelos fatores renda e raciais e mas também "miséria de posição" pode ser inferida pela assimetria entre pais e mães na escolaridade e na profissão. Mesmo com profissões e escolaridade similares, a relação para com os filhos é diferente. Considerando valores tradicionais do homem ser o provedor primário e da mulher ser a cuidadora primária dos filhos, é possível ver essa assimetria como reflexo disso, só que ainda seria um salto.

E sobre a utilização de técnicas mais avançadas estatísticas para tentar melhor observar fenômenos sociais, a regressão logística não foi um modelo bom para análise do todo. Mas para comparação das partes é relevante. Mesmo dando "errado", tem informações valiosas que podem ser tiradas. E no fim, envolve um trabalho investigativo de iterar uma mesmo modelo/técnica e iterar técnicas diferentes.

Bibliografia

BOURDIEU, Pierre. A escolha dos eleitos. In: Os herdeiros: os estudantes e a cultura. Florianópolis: Editora da UFSC, 2014. p. 15-45.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Microdados do ENEM. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 14 dez. 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Panorama do Censo 2022. Disponível em: https://censo2022.ibge.gov.br/panorama/. Acesso em: 14 dez. 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Educa IBGE: Conheça o Brasil - População: EDUCAÇÃO. Disponível em: https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18317-educacao.html#:~:text=N%C3%ADvel%20de%20Instru%C3%A7%C3%A3o&text=No%20Brasil%2C%2053%2C2%25,%2C2%25%20no%20mesmo%20ano. Acesso em: 14 dez. 2023.