# DAG e modelo multinível em análise de dados do ENEM 2019

José Arthur

Junho de 2021

## 1 Introdução

O objetivo deste trabalho é usar algum conhecimento sobre DAG's e modelagem estatística multinível que eu tenho para analisar dados do ENEM, o Exame Nacional do Ensino Médio, do ano de 2019.

## 2 Os dados

Os dados para este trabalho foram coletados do site do INEP, Instituto Nacional de Educação e Pesquisas. Os dados vieram numa grandíssima planilha de excel e continham muita informação de todo tipo. No entanto, quase que imediatamente meus olhos recairam sobre os dados socioeconômicos das inscrições dos alunos. Estes dados contém informações sobre, por exemplo, renda familiar, posses da família, infraestrutura a disposição do aluno, etc e me pareceu interessante tentar encontrar alguma relação entre esses dados e o desempenho dos alunos no exame.

Além disso, também extratifiquei os dados por gênero porque achei que talvez eu pudesse encontrar alguma correlação entre gênero e desempenho em alguma(s) área(s) cobrada(s) pela prova. Eu acabei analisando só o desempenho nas áreas de matemática e suas tecnologias e ciências humanas e suas tecnologias. Pensei que, devido a questão de mulheres na matemática, por exemplo, pudesse perceber alguma diferença e essas áreas também pareciam bem distintas.

Cada uma das planilhas, a com dados socioeconômicos e a com dados por gênero e desempenho, também tinha uma divisão por região e cada região tinha seus respectivos estados. Infelizmente, não havia dados de alunos individualmente, o que tornaria as coisas mais interessantes, mas tudo bem, faz sentido que não haja.

Uma vez que eu não ia usar todos os dados, fiz uma rápida limpeza. Para cada estado, deixei apenas os dados contendo as quantidades de alunos que na incrição declararam ter acesso a internet, de alunos e que tiraram menos e mais que 600 em matemática/humanas, de alunas que tiraram menos e mais que 600

em matemática/humanas e, é claro, de alunos no total. Talvez valha a pena dar uma olhadinha na planilha.

# 3 DAG

Bom, eu montei o seguinte DAG p'ra analisar os dados:

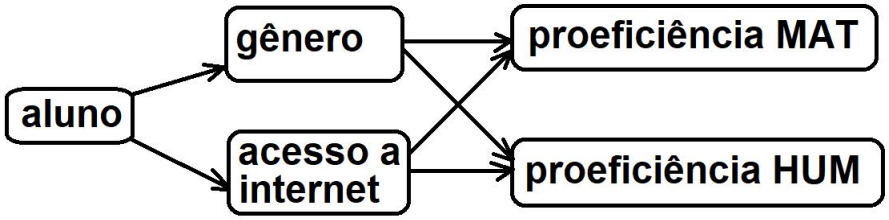


Figure 1: DAG completo

Nesse DAG, queremos analisar a relação do aluno com a proeficiência em cada área. Uma vez que gênero e acesso a internet não devem interferir muito um no outro e eu nem teria dados que relacionam um ao outro, eu dividi o DAG em dois e vou analisar cada um separadamente.

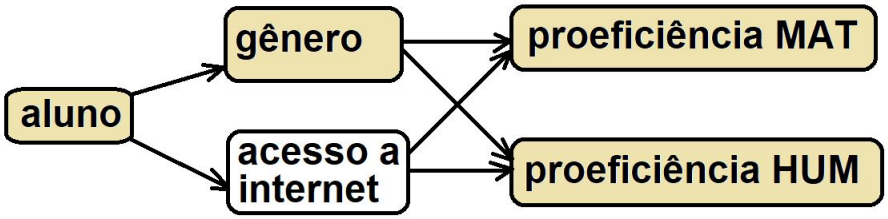


Figure 2: DAG para análise do acesso a internet

É claro que, no fundo no fundo, cada proeficiência teria um DAG próprio. Mas coloquei as duas juntas porquê a visualização não fica comprometida e no fim vamos querer comparar as duas mesmo.

Nos DAG's, 'gênero' e 'acesso a internet' são variáveis intermediárias. Isso torna tudo mais fácil, mas talvez um pouco menos interesssante, dado que as maiores discussões sobre o tema são sobre outros tipos de variáveis (de colisão e confusão, em especial).

Obs: existem dados "não observados" (na verdade, só não coletados por mim) importantes para prever o desempenho de um aluno. Raça e condição social, por exemplo, devem fazer uma grande diferença. Porém, eu optei por

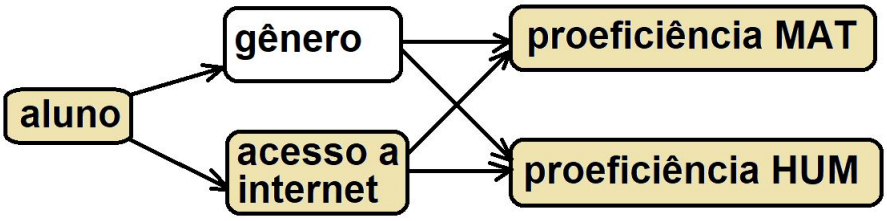


Figure 3: DAG para análise de gênero

restringir um pouco mais a análise e nisso esses dados acabaram ficando de fora (até por uma certa facilidade).

# 4 Gênero

Como gênero é uma variável intermediária, eu apenas separei os dois gêneros e tracei uma linha de tendência. Nos dois gráficos, a esquerda vê-se o desempenho do gênero feminino e a direita, o do masculino.

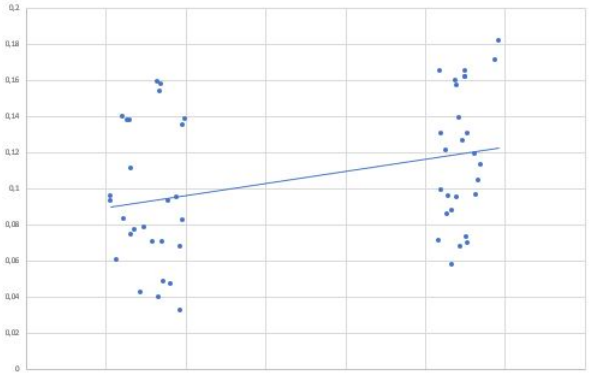

Figure 4: Desempenho por gênero de cada estado em matemática

Usei os dados em porcentagem porque não sabia direito como ponderar por população. Talvez pudesse traçar retas que ligavam os desempenhos dos dois gêneros em cada estado e ponderar o coeficiente angular das retas de acordo com a população.

Isso permite notar que em matemática de fato o gênero masculino se sai melhor que o feminino. Em humanas, os dois gêneros se saem parecidos. É claro que isso não significa que homens sejam naturalmente melhores que mulheres em matemática, não sejamos tão rasos. Sabemos, por exemplo, que desde a infância mulheres são desencentivadas a seguir numa área como essa.

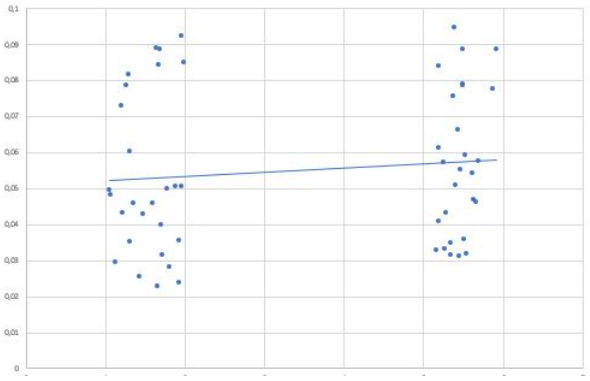

Figure 5: Desempenho por gênero de cada estado em humanas

# 5 Acesso a internet

Novamente, como acesso a internet é uma variável intermediária, bastou fazer uma simples linha de tendência para obsevar a correlação entre isso e o desempenho do aluno no ENEM. Dessa vez usei a porcentagem de alunos com acesso a internet de cada estado.

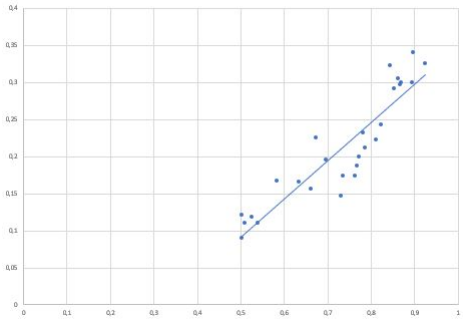

Figure 6: Desempenho de cada estado em matemática pela sua porcentagem de alunos com acesso a internet

Podemos observar que a internet tem um efeito bem positivo na proeficiência dos candidatos. Além disso, como pode se ver abaixo, o efeito é maior do desempenho em matemática (em azul) do que em humanas (em laranja).

# 6 Modelo multinível

Bom, meio rapidamente, poderíamos querer construir um modelo multinível para esses dados. Teríamos primeiro uma média nacional que dita como prediríamos a nota de um participante sem saber sua região, por exemplo. Depois, para

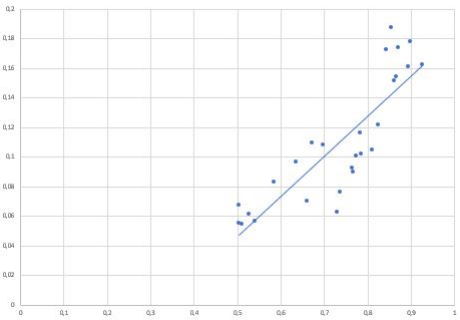

Figure 7: Desempenho de cada estado em humanas pela sua porcentagem de alunos com acesso a internet

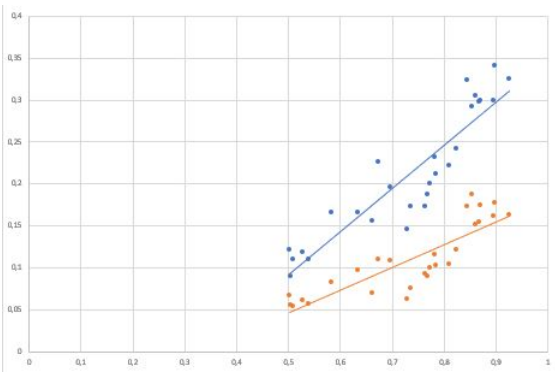

Figure 8: comparação entre os dois últimos gráficos

cada região, podemos fazer um modelo separado com uma média ponderada pela população da média regional e da média nacional.

Eu diminui um pouco a o peso da média nacional porque achei que fazia sentido dar mais peso pros dados "a posteriori" (pensando na média nacional cmo uma priori p'ras médias de cada região).

Na figura abaixo, cada gráfico mostra as linhas de tendência nacional (em azul), da região (em laranja) e a média ponderada entre as duas (em cinza). Cada linha liga o desempenho médio do gênero feminino (a esquerda) como o desempenho médio do gênero maculino (a direita).

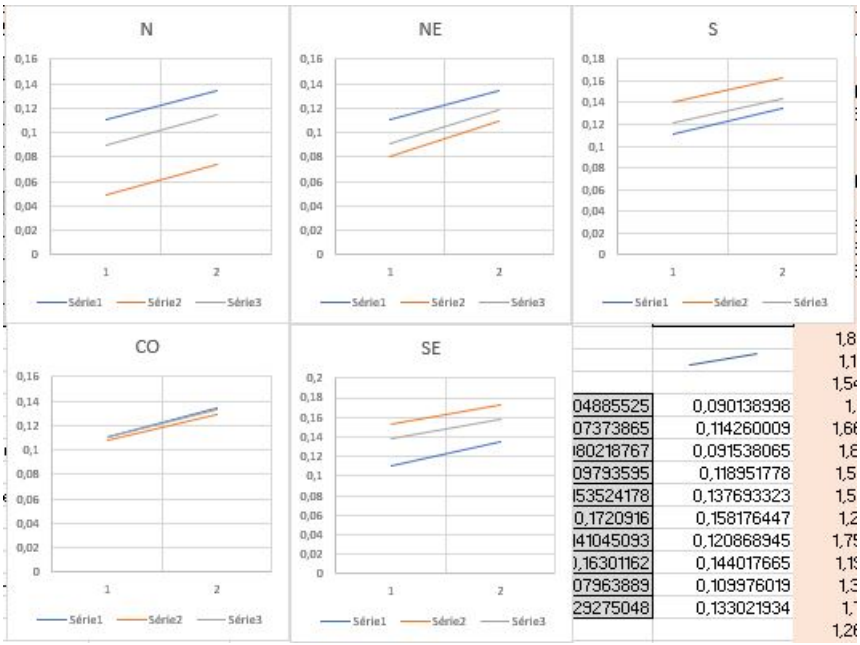


Figure 9: modelos multiníveis para cada região do país

Temos N para norte, NE para nordeste, S para sul, CO para centro-oeste e, finalmente, SE para sudeste. Favor, ignorar o pedaço de tabela que aparece kkkk.

O peso usado para os dados nacionais foi 10. O peso por região foi 5 para o norte, 17 para o nordeste, 5 para o sul, 17 para o sudeste e 4 para o centro-oeste, lembrando, é claro, esses pesos foram dados de acordo com o tamanho da população de cada estado.

Podemos ver que para as regiões com maior peso, noordeste e sudeste, a linha cinza se aproxima mais da linha laranja, construída a partir dos dados da própria região. O centro-oeste por acaso tem uma média muito próxima da média nacional.

Poderíamos ainda descer mais um nível e fazer um modelo por estado, mas não farei isso agora. Também seria bom ainda calcular uma distribuição para os erros de cada região, mas também não farei isso agora.

# 7 Comentários

Eu gostaria de ter usado as bibliotecas do R nesse trabalho, acho que poderia ter me aprofundado na discussão sobre os DAG's e poderia ter feito um modelo multinível melhoriznho, com erro e tudo mais. Isso porque elas facilitam o trabalho, claro. Ainda assim, fico satisfeito de alguma forma com o que consegui fazer. :)

# 8 Referências

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Sinopse Estatísticas do Exame Nacional de Ensino Médio 2018. Brasília: Inep, 2019. Disponível em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/sinopses-estatisticas/enem. Acesso em: 22 de junho de 2021.