Construção de modelo multinível longitudinal para previsão de resultados avaliativos na educação básica

Trabalho Final - Modelagem Estatística

Ana Carolina Erthal
FGV EMAp
Niterói, Brasil
acarolerthal@gmail.com

***Abstract*—Esse artigo aborda resultados avaliativos de alunos e a determinação de variáveis de relevância na determinação desses resultados. Buscamos, utilizando técnicas de Modelagem Estatística - principalmente o ajuste de modelos -, conseguir partir da determinação de um modelo representativo dos dados e variáveis de relevância para prever notas avaliativas. Levaremos em conta, ainda, medidas de erro para quantificar a assertividade de nossos resultados. O artigo tem fins educacionais, então conta com uma parcela significativa de discussões referentes à disciplina de Modelagem Estatística da FGV EMAp.**

***Index Terms*—Educação, Modelagem Estatística, dados longitudinais, modelos multinível, ajuste de modelos.**

I. Introdução

A. O problema

A educação é, antes de tudo, um tema muito complexo. O processo que envolve o aprendizado de cada indivíduo depende não apenas dos agentes diretamente envolvidos (discente e docente), mas de uma série de fatores que têm grande influência no desempenho acadêmico. Não foge muito ao senso comum perceber que um aluno que não conta com apoio familiar, ou mora muito longe do local em que estuda, por exemplo, têm mais dificuldades em dedicar horas de estudo do que um aluno em condições mais favoráveis.

No entanto, o objetivo desse artigo é partir dessa assunção inicial e utilizar dados reais e técnicas de Estatística para observar se, de fato, essa conjectura se comprova no conjunto de dados escolhidos. É claro que, utilizando dados restritos a poucos resgistros em uma localização específica como faremos, não podemos fazer qualquer tipo de afirmação sobre verdades gerais para a Educação, apenas sobre o contexto em que estamos inseridos, mas gostaríamos de observar quais são as variáveis de maior relevância na vida de um estudante, dentre as que dispomos, para determinar quão bem esse aluno se sairá em avaliações didáticas.

Buscamos entender, na prática, se conseguimos determinar essas variáveis de maior importância no resultado avaliativo de alunos, e com isso chegar ao nosso objetivo: tentar prever, para um estudante em circunstâncias designadas por essas variáveis, que desempenho esperamos, levando em conta as medidas de desvio necessárias, e quantificar o quão bem performamos em nossa tentativa de antecipação de resultados utilizando técnicas aprendidas na disciplina de Modelagem Estatística.

É importante observar, desde o princípio, que nossos esforços de predição têm, ainda, forte intenção de levar em conta multiplas avaliações (de mesmos critérios) realizadas pelos mesmos estudantes em momentos distintos do anos letivo. Isto é, queremos observar aspectos longitudinais para estudar se relações entre variáveis de interesse e a realização desses exames através do tempo culminam em boas predições.

B. Relevância

Estudar a escola significa reinventar a Educação a todo momento. Ao redor do mundo, especialistas de renome se debruçam sobre a Teoria da Educação para encontrar formas melhores e mais eficientes de transmitir conhecimento, mas é importante reconhecer a barreira existente entre a construção dessas teorias e a aplicação prática delas nas escolas ao redor do mundo.

Vivemos em contextos muitos diferentes, e muitas vezes essas teorias são desenvolvidas levando em conta circunstâncias distantes das vividas por muitos estudantes. Por exemplo, quando espera-se que o aluno dedique horas de estudo fora da escola, ignora-se com frequência o fato de que este aluno pode não ter condições propícias a estudo em casa, dividindo-a com muitos familiares ou sem ter acesso à internet. Assim, por detrás do estudo puro da educação, é importante que a pesquisa dos fatores socioeconômicos que rodeiam os agentes da educação esteja a todo vapor.

Por isso, é importante pensar a educação além da sala de aula, e o desenvolvimento desse tipo de pesquisa deve ser tão relevante quanto o estudo das ferramentas de ensino. Não levar em conta essa situação resulta em um ensino excludente.

É evidente que esse assunto não surge pela primeira vez nesse artigo. Apesar de haver a linha de estudos que desenvolve teorias educacionais puras, há também aqueles que se dedicam ao estudo dessas diferenças de contexto social, e como eles impactam na educação. No contexto das pesquisas estatísticas em educação, esse tema é muito

recorrente, sendo pauta de diversas publicações. Já em 2003, (Ferrão e Fernandes; 2003) [1] categorizava os fatores que podem influenciar o desempenho acadêmico de um estudante como características culturais, sociais e econômicas da família, habilidades do aluno e fatores escolares, e o assunto permanece tão atual quanto na época.

O tema abordado por esse artigo é, portanto, de extrema relevância para a construção de uma educação inclusiva, já que corrobora com a afirmação de que as variáveis extra-escola são significativas no desenvolvimento educacional de um indivíduo, e constitui um passo no desenvolvimento de uma pesquisa mais geral sobre o assunto, levando em conta dados mais amplos e de diferentes contextos culturais.

C. Os dados

Para a construção dessa pesquisa seguindo o tema definido, utilizaremos um conjunto de dados disponível no Kaggle chamado "Student Grade Prediction" que aborda dados de estudantes do Ensino Médio em duas escolas portuguesas. O dataset é originalmente do departamento de Machine Learning da University of California, Irvine, e conta com dados socioeconômicos e educacionais de alunos. Existem, no total, 33 colunas de dados, das quais 3 representam os resultados de G1, G2 e G3 [1], três avaliações realizadas pelos discentes ao longo de um ano letivo.

Nas 30 outras colunas, temos dados principalmente referentes a questões familiares dos alunos (se os pais vivem juntos, se têm educação formal, quantas pessoas vivem na mesma casa, etc.), pessoais (se têm relacionamentos, consomem álcool, se fumam, etc.) e educacionais (quantas horas dedicam ao estudo, se faltam às aulas com frequência, reprovações anteriores). Não há nenhum dado faltante, e a base é bem mantida pela UCI, então não há necessidade de descartarmos linhas ou fazermos qualquer limpeza nesse sentido.

É importante observarmos, no entanto, que o fato de os dados serem provenientes de duas escolas distintas deve ser observado. Devemos observar, no entanto, que a distribuição desses dados é desigual. Temos cerca de 350 linhas de dado para a escola Gabriel Pereira (GP) e apenas 50 linhas de dado para a escola Mousinho da Silveira (MS), então podemos ter um resultado mais fortemente influênciado pela escola Gabriel Pereira.

II. Métodos

A. Modelos

Como o nosso objetivo é ajustar um modelo que se adapte aos dados e seja eficaz em prever as variáveis de nota, devemos iniciar a nossa definição de metodologia tratando de nossas escolhas de modelo. A princípio, observamos, conforme descrito em I-C, que por termos resultados G1, G2 e G3, podemos ajustar um modelo que leve em conta o aspecto longitudinal dos dados para realizar predições.

[1]Na prática, essas 3 colunas representam dados longitudinais, isto é, medidas repetidamente ao longo do tempo nos mesmos agentes

No entanto, é importante destacarmos o seguinte comentário, feito na página em que os dados são disponibilizados pela UCI:

> "Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful"

Isto é, G1, G2 e G3 têm, é claro, uma forte correlação. No entanto, é enunciado que prever G3 sem as outras duas covariáveis é mais complicado. Assim, para fazermos um ajuste de modelo levando em conta a característica longitudinal dos dados, isto é, estabeleceremos um modelo multinível para medições repetidas usando outras variáveis para prever as notas.

Esse tipo de ajuste também é frequentemente chamado de COORTE, e é bastante relevante no campo da Estatística já que permite a criação de modelos em que a variável tempo (que no nosso caso se traduz como a ocorrência de cada uma das três provas) passa a ter relevância na predição da variável em questão (no nosso caso, as notas).

Para elaborar um modelo COORTE, devemos decidir se queremos variar o intercepto, o coeficiente angular (*slope* ou ambos para cada nível definido, mas essa escolha será feita ao testarmos diferentes relações e compararmos resultados obtidos a partir de cada um. Além disso, podemos ter (e é importante para nossa investigação que de fato haja) outras variáveis relevantes para a predição, e as definiremos mais adiante, após explorarmos os dados cuidadosamente para observar relações.

Note que não vamos explorar tentar prever G3 a partir de G1 e G2, como o site descreve ser menos complicado, dado que a correlação é alta demais para o resultado ser muito útil, principalmente no escopo do problema que estamos buscando endereçar. Isto é, prever G3 partindo de G1 e G2 não nos ajuda a fazer asserções sobre informações socioeconômicas, apenas sobre tendências lineares em resultados acadêmicos.

Quando avançarmos para a análise exploratória de dados, será possível visualizar com mais clareza as correlações descritas acima, quanto às assunções para escolha inicial de modelo e também quanto decisões de escolhas de variáveis preditoras.

B. Ajuste

Nesse ponto, é importante destacarmos, que, para além do modelo, devemos tomar nosso partido dentro das abordagens da Estatística. Faremos nosso ajuste seguindo os precedentes bayesianos ou frequentistas?

A escolha feita nesse artigo é utilizar a abordagem frequentista, e utilizaremos métodos no `R` que utilizam estimador de máxima verossimilhança. Essa escolha se justifica pelo tamanho do nosso conjunto de dados que, na Estatística, é suficientemente grande para considerarmos que temos uma quantidadade relevante de dados anteriores para

encontrarmos o EMV (ao contrário da mentalidade de Machine Learning, em que em geral precisamos de uma quantidade muito maior de dados).

A escolha se deve, também, ao fato de que as ferramentas frequentistas de ajuste de modelo são mais utilizadas em livros didáticos sobre modelos multinível. Os ajustes aqui realizados são baseados principalmente nos capítulos sobre modelos multinível de (Roback e Legler; 2020) [2] e (Gelman e Hill;2007) [3], e ambos os livros abordam o tema utilizando funções como `lmer`, do pacote `lme4`. Sendo assim, já que esse artigo tem fins didáticos, utilizarei a abordagem frequentista.

C. Avaliação

Tendo nosso modelo e abordagem definidos, precisamos esclarecer, também, que método de avaliação de resultados será utilizado, já que buscamos, após estabelecer o modelo que deverá prever resultados, poder avaliar a capacidade de se adaptar aos dados desse modelo.

Desde o início estabelecemos que nosso objetivo principal é, de fato, conseguir prever as notas de estudantes partindo de outras características, sendo estas relativas a questões escolares, pessoais ou familiares, mostrando que de fato há uma relação estruturada entre essas variáveis e o desempenho escolar.

Nesse sentido, gostaríamos de, após realizar as predições, conseguir quantificar o quanto acertamos - isto é, se nosso modelo de fato obteve sucesso em prever a variável de interesse: resultados de avaliações. Por isso, utilizaremos medidas de capacidade preditiva de nosso modelo.

Na prática, como temos uma regressão, é bastante imediato pensarmos em conferir valores de MSE (Mean Squared Error) ou RMSE (Root Mean Squared Error). Também é importante conferirmos valores de AIC (critério de informação de Akaike), buscando o menor possível.

Ainda que haja, nesse artigo, algum interesse em explicar relações entre variáveis (secundario a realizar predições), a abordagem clássica de realizar testes de bondade do ajuste através de R^2 esbarra em uma questão de nosso modelo. A medida em geral relevante na estatística, o R^2, encontra certa divergência entre a comunidade científica por não ser considerado igualmente quantificável em modelos multinível. Calcular o coeficiente de determinação significa quantificar de variância dos dados que é explicada pelo modelo, e fazer isso de uma forma geral se temos diferentes níveis é uma questão que, nos últimos anos, tem sido alvo de publicações e criação de pacotes com diferentes abordagens, mas sem que haja acordo claro entre a comunidade científica. [4] [5]

Por isso, no campo da bondade do ajuste, utilizaremos apenas a medida de log-verossimilhança, que quantifica o ajuste de nosso modelo aos dados em questão. É importante lembrar que, apesar de podermos inferir conclusões sobre nosso modelo a partir dos valores absolutos dessas medidas (caso estes sejam muito diferentes do razoável), o principal objetivo de realizá-las é para compararmos a performance obtida entre modelos, e as utilizaremos para escolher o melhor ajuste de modelo.

III. Resultados

A. Análise exploratória dos dados

Para podermos, de fato, avançar para realizar o ajuste de nosso modelo e realizarmos constatações e predições, precisamos decidir quais variáveis serão utilizadas como preditoras, isto é, que variáveis influenciam mais fortemente nos resultados avaliativos dos estudantes. Para isso, foi utilizada no *python* a biblioteca *pandas-profiling*, que fornece uma visão geral dos dados coluna a coluna e também das interações entre elas, além de descrever alguns alertas, como de correlações altas demais, ou possíveis valores nulos.

Inicialmente, para conhecermos melhor as variáveis, vamos analisar o seguinte gráfico de correlações, representado por um heatmap:

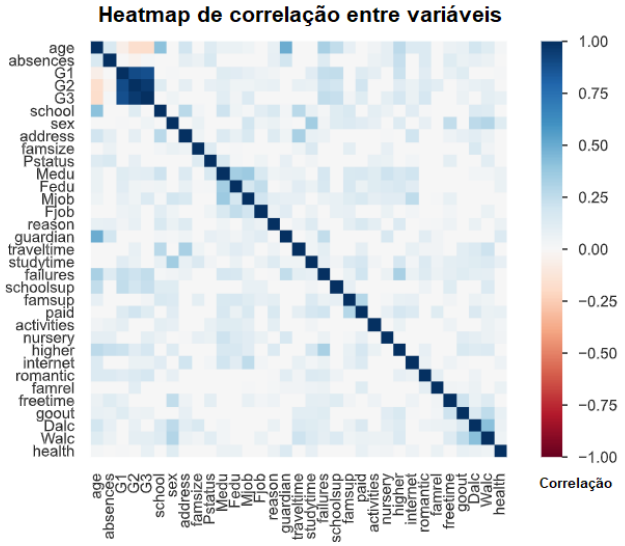

Fig. 1: Gráfico de correlação entre as variáveis do *dataset*

Como já discutimos anteriormente, G1, G2 e G3 têm uma correlação quase absoluta (Fig. 4), mostrando que alunos que performam bem em uma prova normalmente também tem bons resultados nas outras (e o contrário também). No entanto, nosso objetivo é descobrir que outras variáveis também se relecionam bem a essas variáveis. Observe que as outras correlações são consideravelmente mais baixas, mas é possível visualizar certa dependência entre as notas em avaliações e as variáveis *failures* (f) e *schoolsup* (s), seguidas por *romantic* (r) e *higher* (h).

Nesse ponto, é interessante explicarmos o que cada uma dessas colunas representa. A coluna *failures* representa o número de disciplinas reprovadas anteriormente (numa escala de 1 a 4, sendo o último representativo de 4 reprovações ou mais). *schoolsup* é uma variável binária que indica se o aluno conta com suplemento educacional ou não, isto é, se frequenta

aulas extra-escola. A variável *romantic* também é binária, e informa se o estudante está envolvido em um relacionamento, e por fim *higher* é uma variável binária que indica se o aluno planeja ingressar no Ensino Superior.

Vamos explorá-las mais a fundo, mas começaremos checando como elas se relacionam em valores:

TABLE I: Correlações entre as variáveis objetivo suas e possíveis variáveis preditoras

	G1	G2	G3	f	s	r	h
G1	1.00	0.89	0.87	0.23	0.24	0.11	0.21
G2	0.89	1.00	0.95	0.20	0.18	0.13	0.11
G3	0.87	0.95	1.00	0.24	0.23	0.17	0.16
f	0.23	0.20	0.24	1.00	0.00	0.12	0.32
s	0.24	0.18	0.23	0.00	1.00	0.05	0.00
r	0.11	0.13	0.17	0.12	0.05	1.00	0.07
h	0.21	0.11	0.16	0.32	0.00	0.07	1.00

Observando a tabela, conseguimos ter uma visão mais clara das correlações, e observamos que mesmo as mais relevantes não são muito altas. Ainda assim, tentaremos ajustar o modelo utilizando-as, já que é necessário trabalhar com as ferramentas que dispomos.

É importante lembrarmos que, apesar de o passo inicial ser sempre checar quais variáveis são mais relacionadas à variável resposta, devemos ter o cuidado de remover multicolinearidades, isto é, não utilizar dois preditores que tenham a mesma influência sobre nossa variável de interesse, já que conhecendo uma, não há ganho em conhecer a outra.

Assim, checando a tabela podemos detectar possíveis multicolinearidades. É bastante imediato perceber que *failures* (f) tem uma correlação relevante com *higher* (h). Podemos excluir, então, uma dessas covariáveis, e sendo *failures* mais fortemente relacionado a G1, G2 e G3, não utilizaremos *higher* no nosso ajuste de modelo.

O mesmo pode ser dito da variável *romantic*, que tem uma correlação com *failures* na mesma medida que consideramos suficiente para ser considerada relacionada a G1, G2 e G3. Assim, também não utilizaremos essa variável.

Através da análise exploratória, conseguimos limitar nossa variável resposta à influência de duas variáveis preditoras: *failures* e *schoolsup*. Veremos mais claramente a influência entre essas colunas e os resultados nos exames.

Utilizamos, a princípio, a coluna GM, que representa a média entre as colunas G1, G2 e G3 (isto é, a nota média por aluno entre os três exames). Note que, de fato, há uma correlação clara entre as variáveis, e o aumento do número de *failures* é associado a uma tendência de média menor nas avaliações. Esse padrão se repete para cada um dos três gráficos entre *failures* e cada avaliação, mas os omitiremos aqui por simplicidade.

No entanto, temos interesse em saber, também, se a influência de *failures* ocorre de forma diferenciada em cada uma das avaliações, o que justificaria a abordagem de modelos multinível longitudinais. Veja, se um valor maior em *failures* resulta em uma distribuição de resultados pior em G1 do que em G3, por exemplo, temos motivo para desconfiar que

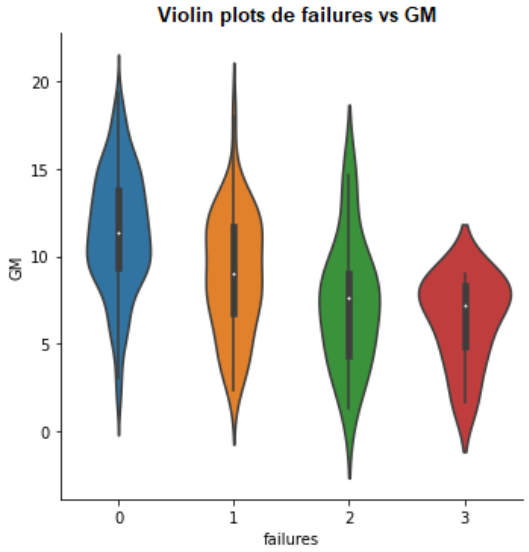

Fig. 2: Gráfico de violino entre *failures* e GM

essa variável impacta diferentemente cada grupo, e abordar COORTE parece bastante razoável.

Perceba na imagem abaixo que, de fato, observa-se que em cada nível temporal a distribuição de indivíduos entre *failures* é diferente. Por exemplo, alunos que reprovaram em disciplinas anteriores estão associados a um desempenho comparativamente pior do que alunos que nunca reprovaram. Essa diferença, no entanto, parece maior em G1 do que em G3!

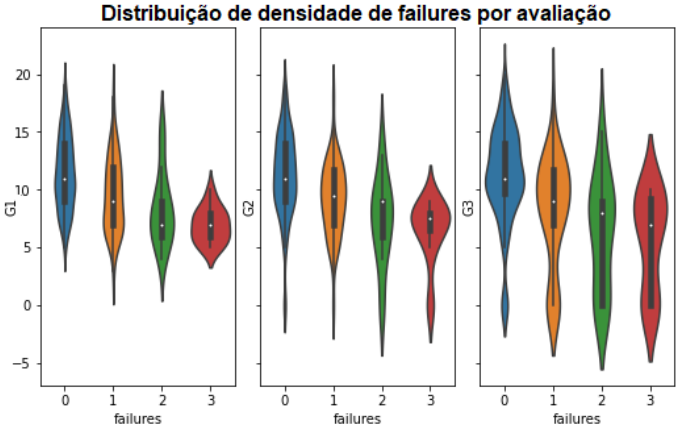

Fig. 3: Gráfico de violino com densidade *failures* por cada avaliação

Vamos fazer, também, um gráfico que represente a relação entre as notas e a variável *schoolsup*.

Para termos uma visão completa ao explorar os dados, queremos explorar a relação dessa variável com cada avaliação. Portanto, na Figura 4 exploramos também essa associação, abordando em meio às correlações entre avaliações a distribuição entre a variável binária *schoolsup*.

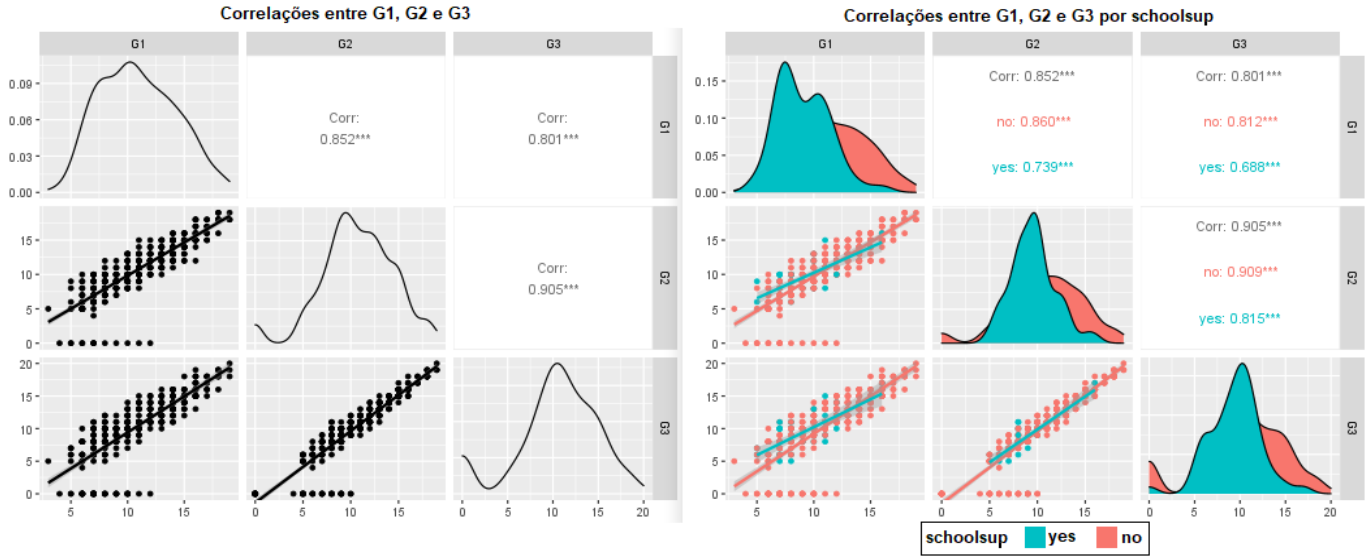

Fig. 4: Correlações entre variáveis de interesse e considerando um preditor binário (*schoolsup*)

Note que, novamente, observando os gráficos entre *schoolsup* e G1, G2 e G3 individualmente, concluímos resultados são indicadores da necessidade de utilizar um modelo que considere o aspecto longitudinal dos dados, já que há uma mudança na distribuição dos dados entre estudantes que realizaram ou não o suplemento escolar de acordo com a passagem do tempo (avaliações). Temos indícios, então, a partir da análise exploratória, de que devemos utilizar *schoolsup* como covariável à organização dos efeitos aleatórios do nosso modelo.

Além disso, na Fig. 5 utilizamos, novamente, a média GM, e a correlação entre as variáveis é observável (ainda que não seja possível dizer apenas visualizando se a média de resultados aumenta ou diminui conforme *schoolsup*, é imediato visualizar que há um comportamento diferente quando temos ”sim” ou ”não”).

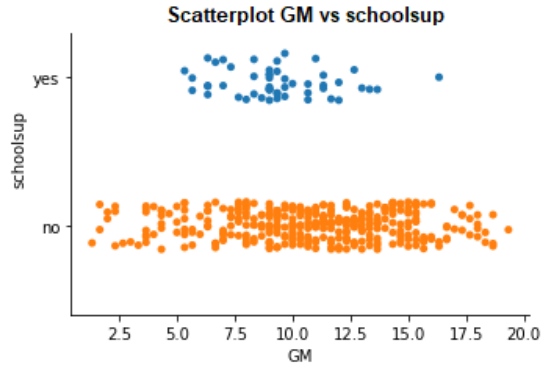

Fig. 5: Scatterplot entre GM e *schoolsup*

Assim, determinamos nossas covariáveis preditoras, e podemos partir para a aplicação prática do ajuste de modelos.

B. Ajustes e aplicação

Ajustar de fato um modelo requer um processo extensivo de compreensão dos dados com os quais estamos trabalhando. Mas quando esse processo parece estar encerrado, e de fato vamos a nossa ferramenta computacional para criar nosso modelo, o processo de explorar as relações entre variáveis é retomado em certa perspectiva. [6]

Precisamos experimentar um pouco, trabalhar diferentes relações entre covariáveis, e, no caso de modelos multinível, descobrir se o ideal é variar o intercepto, o coeficiente angular ou ambos por nível. Ainda mais, precisamos descobrir quais covariáveis devem influenciar nessa escolha de efeito aleatório (definição de intercepto/coeficiente por grupos).

Essa experimentação, comum ao trabalho do estatístico, será omitida aqui, e partiremos para explicitar o modelo que melhor explica os dados seguindo a abordagem e método determinadas:

$$y_i = \alpha + \beta_i f_i + \phi_i s_i + \gamma_{t[i]} f_i + \omega_{t[i]} s_i + \epsilon_i$$

Onde:

- y_i: predição para a linha i;
- f_i: *failures* na linha i;
- s_i: *schoolsup* na linha i;
- β_i e ϕ_i: coeficiente angular na linha i;
- $\gamma_{i[j]}$ e $\omega_{i[j]}$: coeficientes angulares na linha i para o grupo (tempo) t;
- ϵ_i: erro em i.

Para ajustar esse modelo, fazemos uma modificação na estrutura dos dados, realizando um `pivot_longer` para termos uma coluna ”G”, onde 0 representa a aplicação G1, 1 representa G2 e 2 representa G3, e uma coluna ”grade”, contendo a nota daquele aluno naquela aplicação de exame. Isso é importante para que possamos utilizar a coluna ”G”

como coluna dos grupos e ”grade” como variável de interesse. Tendo os dados prontos, utilizaremos 80% deles para ajustar o modelo, guardando 20% dos dados para, posteriormente, podermos realizar predições e avaliar sua performance.

Sendo assim o modelo, quando ajustado utilizando o R e a função $lmer$, usando 80% de nosso conjunto de dados, tem os seguintes coeficientes e incertezas:

TABLE II: Coeficientes de efeito aleatório no modelo ajustado

	Intercept	*failures*	*schoolsup*
0	11.49956	-1.593356	-1.714785
1	11.49956	-1.836984	-1.539789
2	11.49956	-2.174398	-1.297426

Aqui exibimos os coeficientes definidos pelo modelo para os chamados efeitos aleatórios - isto é, que mudam de acordo com os níveis, que aqui são cada uma das aplicações de avaliação. Assim, obtivemos valores constantes de intercepto (já que, de fato, definimos no modelos que ele não deveria variar entre grupos) e valores que variam para os coeficientes angulares.

É interessante observar que, enquanto a variável *failures* diminui conforme as avaliações (isto é, a passagem do tempo no ano letivo), *schoolsup* aumenta. Isto é, ter realizado aulas suplementares passa a ter um efeito mais positivo. Um leitor atento poderia observar que *schoolsup* ter um valor negativo para o coeficiente angular não parece correto, mas observe a imagem 5.

No entanto, além de observar os valores absolutos, é importante analisarmos também suas incertezas:

TABLE III: Incertezas dos coeficientes de efeito aleatório

	Variance	Std.Dev
failures	0.12045	0.3471
schoolsup	0.06214	0.2493
Residual	13.18360	3.6309

Note que, comparativamente com os valores absolutos, temos desvios baixos, o que fala positivamente de nosso modelo. Além disso, ao somarmos o desvio padrão a nossas estimativas (ou duas vezes o desvio padrão, tratando de um intervalo de confiança de 95%) não incluiremos o zero, o que novamente fala a favor do modelo.

Queremos tratar, também, dos chamados coeficientes de efeito fixo (note que, obviamente, esse efeito varia conforme o dado entregue ao modelo para ajuste, e o nome ”efeito fixo” se refere a sua não-variação conforme a mudança de níveis do modelo multinível). Vejamos a tabela de coeficientes:

TABLE IV: Coeficientes de efeito fixo no modelo ajustado

	Estimate	Std. Error
(Intercept)	11.4996	0.1226
failures	-1.8682	0.2456
schoolsup	-1.5173	0.3459

Assim, conseguimos ajustar nosso modelo e obter coeficientes. É importante, agora, avaliarmos a performance de nosso modelo. Vamos iniciar checando o gráfico de resíduos e garantindo a normalidade destes.

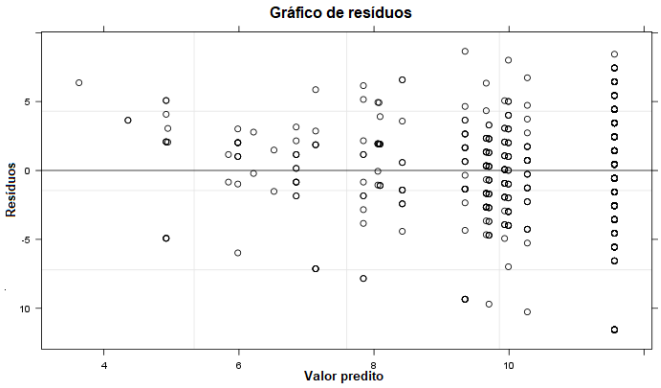

Fig. 6: Gráfico residual após ajuste do modelo

Parece que, de fato, os resíduos se distribuem igualmente acima e abaixo da linha de ajuste, confirmando a normalidade dos resíduos ratificando nosso modelo.

Podemos avançar, então, para a parte final de nossos resultados, em que utilizamos os dados de teste separados no início para prever resultados de avaliações. Utilizando o método `predict` do `R`, obtivemos um vetor de predições, e então foi possível obter as medidas de capacidade preditiva do modelo.

Começamos avaliando nosso Root Mean Squared Error. Obtivemos um RMSE de 3.28, que parece ser razoável quando estamos tratando de nosso intervalo de valores entre 0 e 20 (notas possíveis para cada prova). Isto é, não estamos tão exatos quanto gostaríamos, mas é possível enxergar que estamos bem distantes de chutes às cegas.

Errar por 3 pontos, no entanto, não é ideal, e idealmente realizaríamos uma predição um pouco mais precisa. Esse valor, porém, é resultante do modelo multinível longitudinal de menor AIC.

Avaliando modelos que combinavam as variáveis que observamos serem de maior relevância de diversas formas diferentes (alterando quais variáveis influenciariam nos efeitos fixos e efeitos aleatórios, inclusive dentre esses se/como alteraríamos intercepto e *slope*), escolhemos, de fato, o que tinha o menor Critério de informação de Akaike (AIC), isto é, mensuramos a qualidade dos modelos e escolhemos o melhor deles. Na prática, obtivemos um AIC de 6437.62, o que ainda parece alto também.

Quanto à bondade do ajuste, temos o valor da log-verossimilhança de nosso modelo escolhido: -2527.35. Sabemos que a interpretação dessa medida é que quanto maior for o valor, temos um modelo que se ajusta melhor a nossos dados. De fato escolhemos, entre os modelos obtidos, o de maior log-verossimilhança, mas novamente este não parece ser, em termos absolutos, um valor promissor.

De forma geral, a avaliação de resultados parece satisfatória no ponto em que confirma a relação entre as covariáveis preditoras e a variável resposta. Terminamos, porém, com certa animosidade pelos valores avaliativos qualitativamente inferiores do que esperávamos ao longo do desenvolvimento

do ajuste. Discutiremos mais a fundo possíveis justificativas e caminhos futuros na próxima seção.

IV. Conclusão

A. Discussão: o que aprendemos?

A nível geral, podemos destacar algumas conclusões obtidas a partir do desenvolvimento deste artigo. Através do desenvolvimento da análise exploratória nos pareceu fortemente que as variáveis que determinamos de fato tinham uma influência relevante no resultado de avaliações. Particularmente, as variáveis de *failures* e *schoolsup*, que apresentavam as maiores correlações com G1, G2 e G3, aparentavam ter influências diferentes para cada realização do exame, isto é, vislumbramos um aspecto longitudinal dos dados.

Ao construirmos o modelo, no entanto, tivemos um resultado que parece bastante mediano. Realizar predições com erro médio de aproximadamente 3 unidades em uma escala de 0 a 20 não é ideal, mas também demonstra que o modelo tem algo a oferecer: estamos estabelecendo relações que de fato existem. As possíveis origens dessa disparidade entre o resultado esperado e o obtido são várias, e listarei aqui as que considero mais relevantes.

Existe a possibilidade em que, apesar de termos 3 aplicações ao longo do tempo sobre os mesmos agentes, não houvesse uma ligação tão forte entre as covariáveis que escolhemos (e talvez com nenhuma variável) e mudanças no desempenho ao longo do tempo. Na prática, uma prova pode ter sido simplesmente um pouco mais complicada do que a outra, ou mais complicada para alunos de uma escola do que para alunos de outra. Nesse caso, abordar um modelo linear simples teria um resultado mais satisfatório.

É importante observar aqui que houveram tentativas no sentido de realizar modelos mais simples, usando o pacote `lm` e as mesmas covariáveis, mas tentando prever a nota média GM, e o resultado pareceu ligeiramente melhor. [2] No entanto, iniciamos esse projeto com objetivo de observar características de medições repetidas, e tendo ou não sucesso absoluto, temos interesse de discutir e comentar o que observamos nesse aspecto. Sendo assim, realizar essa tentativa adicional tinha objetivo de confirmar a relação entre as covariáveis e a variável resposta.

Nesse tópico, encontramos um possível motivo adicional de insucesso. As correlações que observamos a princípio na tabela I entre as variáveis escolhidas e G1, G2 e G3 não eram muito altas. Ora, dado o intuito do trabalho tínhamos de observar as mais relevantes, mas aqui observamos que uma correlação de 0.25 ou 0.2 pode não ser suficiente para estabelecermos um modelo que prevê resultados com tanta asserção se baseado nessas variáveis.

Um terceiro e último possível motivo que abordaremos é a questão pontuada anteriormente de tratarmos de duas escolas diferentes, mas endereçaremos melhor essa questão em Limitações (IV-B).

A conclusão tomada aqui é que, de fato, observamos relações entre características socioeconômicas dos estudantes e resultados acadêmicos - ainda que não com a assertividade que esperávamos, levando em conta multi-medições -, reforçando as premissas que tínhamos no início. É claro, realizamos esse experimento em um conjunto limitado demais para tomarmos qualquer conclusão como geral, mas há certa satisfação em confirmarmos que o caráter normativo avaliativo deixa de levar em conta características intrínsecas à vida de muitos estudantes, e sentir que qualquer passo na comprovação dessa disparidade também é um passo no sentido da mudança.

Pelo caráter acadêmico e educacional deste trabalho, acredito que abordar aprendizados a nível pessoal também seja razoável. Ao longo do desenvolvimento deste trabalho utilizei mais a fundo ferramentas estatísticas do `R`, que até então tinha sido uma dificuldade maior e uma experiência um pouco frustrante. Além disso, ter escolhido abordar um modelo multinível para medições repetidas gerou a necessidade de pesquisar mais a fundo sobre o assunto, tanto livros acadêmicos quanto artigos em que as técnicas foram aplicadas.

B. Limitações

Desenvolver esse projeto trouxe à tona algumas limitações, sendo estas dos dados (e instransponíveis) quanto metodológicas pessoais. Começamos a abordar as limitações dos dados ao relatar possíveis falhas na construção de modelos, mas vamos aprofundar essa discussão.

Explicitamos na introdução uma questão relevante de nosso conjunto de dados: temos a divisão entre duas escolas, mas nossos registros são divididos de forma bastante desigual entre elas, havendo comparativamente muito mais dados da escola Gabriel Pereira do que da escola Mousinho da Silveira. Essa desproporcionalidade causa uma influência muito maior da escola GP, e possivelmente erros de predição com mais frequência na escola MS.

Na prática, sabemos que essa limitação dos dados pode ser resolvida com uma abordagem estatística já utilizada nesse artigo: os modelos multinível. Nesse caso, poderíamos estabelecer um modelo multinível hierárquico (medindo resultados nos níveis aluno e escola), e fazer predições levando em conta o grupo a que pertencem. Para manter nosso objetivo principal, no entanto, seria necessário construir um modelo bastante complexo, levando em conta os níveis longitudinais e hierárquicos.

Essa possibilidade esbarrou em uma barreira metodológica na medida em que pareceu além do que eu conseguiria desenvolver no prazo desse trabalho. No entanto, a abordagem parece interessante, e possivelmente retornaria resultados melhores, e se ajusta satisfatóriamente como uma possível direção futura.

Por fim, novamente entre as limitações metodológicas, tive certa dificuldade de compreender as discussões da comunidade científica estatística quanto ao uso do R^2 em situações como a

[2] Obtivemos um RMSE de 3.38 (um pouco maior) e um AIC de 2091.89 (bem menor, considerando que o modelo de fato era bem mais simples e o AIC leva isso em conta)

de meu projeto. Cheguei a me deparar com o pacote `r2mlm`, mas pelas publicações que encontrei ele parecia bastante novo e longe de ser um consenso. De qualquer forma, tive bastante dificuldade na interpretação dos resultados ao aplicá-lo em meu modelo.

C. Trabalhos futuros

Por fim, devemos abordar possíveis direcionamentos futuros em nosso trabalho. A principal exploração que esse trabalho deixa como intenção futura é abordar a questão descrita nas Limitações sobre utilizar um modelo multinível hierárquico para as escolas. Mantendo o direcionamento do projeto inicial, gostaríamos de manter a observação de medições sequenciais, isto é, nosso modelo multinível longitudinal, e assim ajustar um modelo multinível nos dois aspectos.

Além disso, tentar realizar a mesma abordagem em um *dataset* diferente também seria bastante interessante, principalmente no contexto de escolas brasileiras, utilizando provas aplicadas pelo SAEB (Sistema Nacional de Avaliação da Educação Básica), podendo abranger aspectos longitudinais mais amplos, já que as provas são aplicadas anualmente, e também hierárquicos mais fortes, havendo a possibilidade de levar em conta os níveis aluno, estados e regiões, por exemplo.

De forma geral, o assunto é bastante amplo e promissor, além de extremamente relevante num contexto mundial, e explorar mais a fundo essas relações se traduzem em contribuições com o desenvolvimento da Educação.

REFERENCES

[1] M. E. F. e Cristiano Fernandes, "O EFEITO-ESCOLA E A MUDANÇA – DÁ PARA MUDAR? EVIDÊNCIAS DA INVESTIGAÇÃO BRASILEIRA," *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, 2003.

[2] J. L. Paul Roback, *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*. Chapman and Hall/CRC, 2020.

[3] J. H. Andrew Gelman, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[4] H. S. Shinichi Nakagawa, "A general and simple method for obtaining r2 from generalized linear mixed-effects models," *British Ecological Society*, 2012.

[5] S. K. S. Jason D Rights, "New recommendations on the use of r-squared differences in multilevel model comparisons," *National Library of Medicine - National Center for Biotechnology Information*, 2020.

[6] "Advanced Statistics using R longitudinal data analysis," https://advstats.psychstat.org/book/longitudinal/index.php, acessado em 23/05/2023.