Investigando as causas de ausências em escolas de ensino médio: Uma análise com o modelo de regressão Binomial Negativo

Samuel M. Medeiros*

07 fevereiro, 2023

Resumo

Este trabalho apresenta uma investigação sobre as causas de ausências em escolas de ensino médio em Portugal. A presença regular dos alunos nas aulas é fundamental para o sucesso acadêmico e o desenvolvimento pessoal. Portanto, é importante identificar as causas que levam a estas ausências. Para fazer isso, é apresentado um modelo para dados de contagem a fim de examinar a relação entre as características domésticas e pessoais dos alunos e suas faltas às aulas. Este modelo permitiu que fosse avaliado a influência dessas características na frequência dos alunos nas aulas e identificassem possíveis soluções para melhorar a presença dos estudantes.

Introdução

Este trabalho se baseia em uma pesquisa original que tenta estabelecer uma relação entre o consumo de álcool e as notas baixas dos estudantes (Cortez e Silva 2008), foram utilizados modelos com tarefas de classificação binária/politômica e regressão.

Este estudo utiliza uma metodologia baseada em modelos lineares generalizados de contagem que se adequa ao escopo do assunto em questão e fornece uma contribuição significativa para a compreensão das causas de ausência nas escolas de ensino médio, apresentando resultados importantes para entender dinâmicas sociais quanto a relação aluno escola na atualidade.

Banco de Dados

Com base em um banco de dados obtido na plataforma Kaggle sobre o consumo de álcool entre estudantes e suas respectivas notas em matemática e português, este estudo foi realizado para compreender a relação entre as características pessoais e sociais dos alunos e o número de faltas em um ano (Variável do tipo numérica inteiro), O banco de dados está disponível no repositório do *GitHub* para este trabalho. Devido ao alto número de variáveis explicativas, se comparado ao número de observações, uma seleção a *priori* foi necessária. As variáveis utilizadas podem ser verificadas na Tabela 1, onde é apresentado também a indicação para o tipo de variável bem como sua indicação no banco de dados originais.

Tabela 1: Variaveis selecionadas para analise

Codigo	Variavel	Classificacao
school	Escola	Categorica

*Universidade Federal do Espírito Santo, samuel.medeiros@edu.ufes.br

Codigo	Variavel	Classificacao
sex	Sexo	Categorica
reason	Razao para escolher a escola	Categorica
Dalc	Consumo de alcool em dias uteis	Categorica
age	Idade	Numerica
internet	Acesso a internet em casa	Categorica

Exploratória dos dados selecionados

É possível observar pela Tabela 1 que os dados são majoritariamente categóricos, podemos observar pela Tabela 2 como se dá a frequência de cada uma das categorias para cada uma das covariáveis em estudo.

Para uma melhor adequação do modelo e uma maior explicabilidade de forma geral, uma recategorização da variável 'Dalc', nível de consumo de álcool, originalmente dividia em: muito baixo(1), baixo(2), moderado(3),alto(4) e muito alto(5), devido à baixa presença de observações para as categorias moderado, com 0.07 dos dados totais, alto, com 0.02 dos dados totais e muito alto, com 0.02 dos dados totais. O reagrupamento foi realizado considerando variáveis com 'Dalc' igual ou superior a 3 (Nível de consumo moderado) como a nova categoria 'high' e observações com nível de consumo de alcool inferior a 3 como 'low'. As novas categorias podem ser observadas na Tabela 2.

Tabela 2: Tabela de frequencia de categorias por variavel

Variavel	Codigo	Categoria	Frequencia	Proporcao
school	GP	Gabriel Pereira	349	0.88
	MS	Mousinho da Silveira	46	0.12
sex	F	Mulher	208	0.53
	M	Homem	187	0.47
reason	course	Preferencia de curso	145	0.37
	home	Perto de casa	109	0.28
	other	Outro	36	0.09
	reputation	Reputacao	105	0.27
Dalc	low	baixo	351	0.89
	high	alto	44	0.11
internet	no	Sim	66	0.17
	yes	Nao	329	0.83

Nota-se que o número de observações para as duas escolas é bem discrepante, ou seja, um número muito inferior de observações da escola 'MS', ou Mousinho da Silveis, em relação a escola Gabriel Pereira (GP).

Podemos entender melhor o comportamento do número de faltas ao observar a Figura 1, onde é identificável um grande volume de observações para alunos com nenhuma falta no ano letivo, e algumas poucas observações, *outliers* possivelmente, para alunos com mais de 40 faltas totais no ano letivo. É possível identificar pela Figura 2, ao analisar a frequência de faltas por categoria para cada uma das covariáveis, o perfil dos dados citados como possíveis *outliers*. É possível identificar uma maior variabilidade para categoria baixo consumo de álcool durante a semana bem como para a aquelas observações onde a razão para escolha do colégio foi a localidade. Como esperado em consequência do tipo de amostra em estudo, alunos de ensino médio, a distribuição de idade se concentra principalmente entre 16 e 18 anos, com alguns ainda entre 18 e 20.

Podemos, pela Figura 3, previamente indicios ou não de alguma relação entre a interação das covariáveis idade e uso de internet na resposta número de faltas. Note que, o comportamento para idades como 15 e 20, por mais que falte observações o suficiente é possível identificar um comportamento diferente das

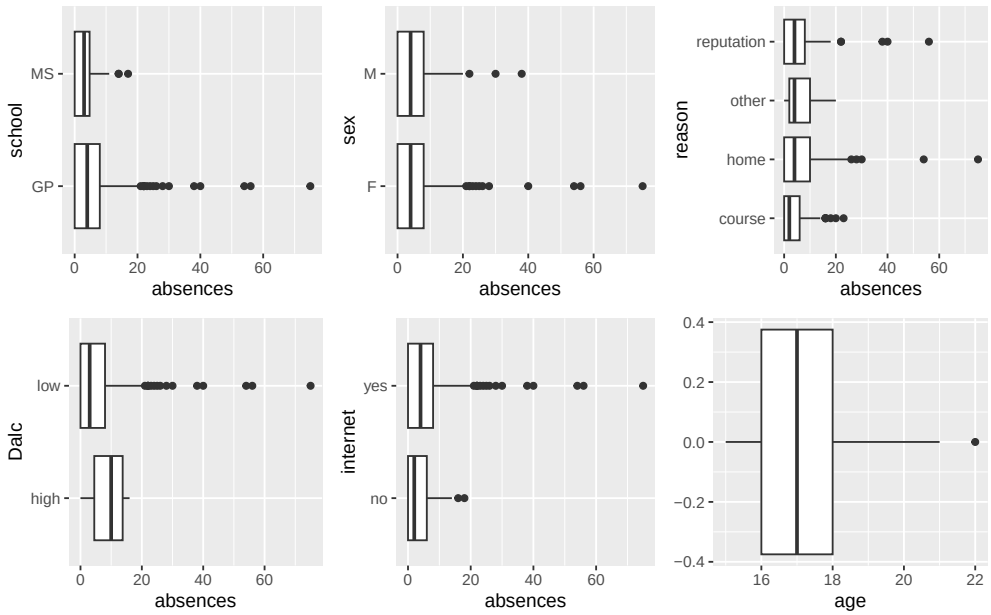

Figura 1: Número total de faltas por categoria de cada covariável

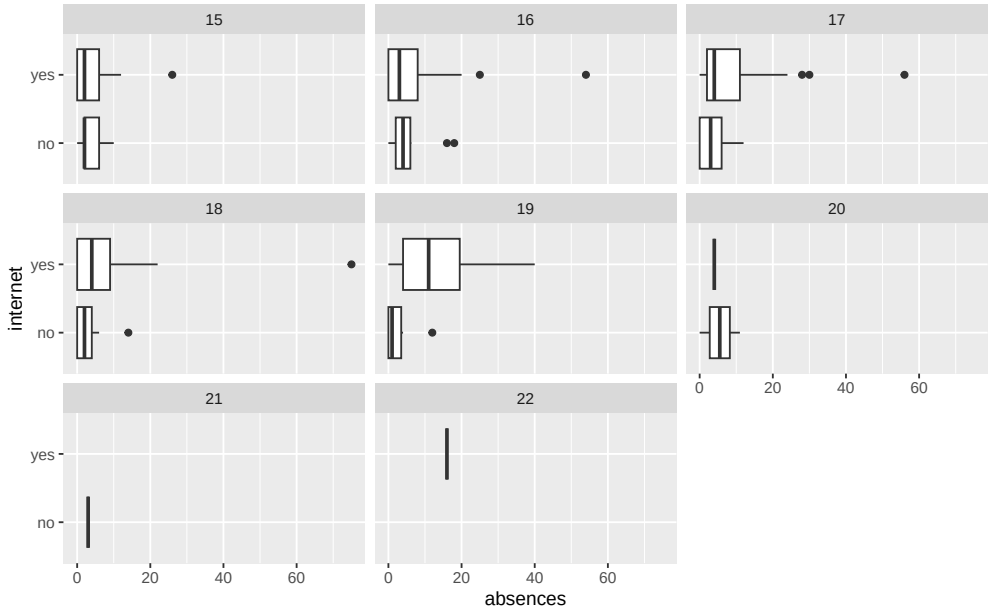

Figura 2: Número total de faltas por acesso a internet para cada idade

outras categorias, que possuem uma maior variabilidade quando o aluno tem acesso a internet no domicílio, possuindo também números de faltas mais elevados quando possuem o acesso.

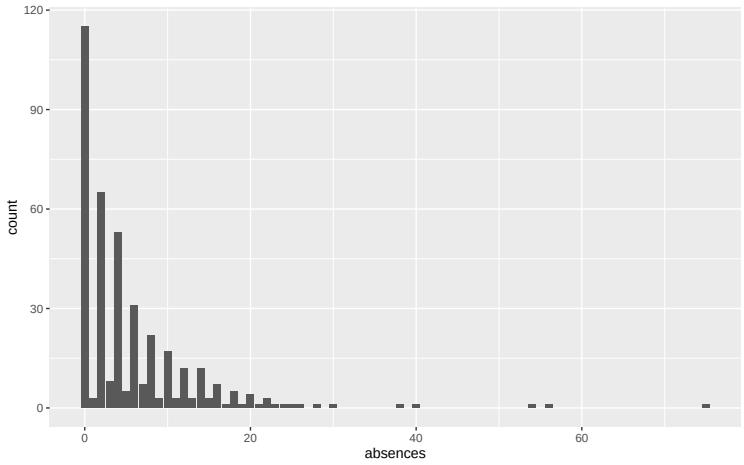

Figura 3: Frequência de número de faltas

Aplicação do modelo

Dentro da literatura, buscamos o modelo mais parcimonioso, ou seja, aquele que consegue uma boa explicabilidade de maneira mais simples possível. Considerando a variável de interesse no estudo como sendo número de ocorrência de determinado evento em função de certas características e/ou situações, a categoria de modelos mais adequada no caso são os log-lineares. No caso em estudo, número de faltas, as quais serão modeladas utilizando o modelo mais básico, modelo de Poisson, assumindo sua ligação canônica *log*. Denotando Y_i como número de ausências para o aluno dada as i-ésimas características . Supondo Y_i como sendo uma variável de distribuição Poisson(μ_i) em que μ representa a taxa média de faltas para o indivíduo com as características citadas. Dada as definições é estabelecido então o modelo:

$$log(\mu_i) = \alpha + \beta_1 DalcH_i + \beta_2 Age_i + \beta_3 SexF_i + \beta_4 SchoolM_i + \\ \beta_5 InternetN_i + \beta_6 ReasonH_i + \beta_7 ReasonO_i + \beta_8 ReasonR_i$$

Modelo ajustado

Utilizando-se do modelo descrito, é perceptível pela Figura 4, o gráfico envelope, que o modelo Poisson não é o adequado para modelagem do banco de dados. Em virtude da péssima qualidade de ajusta vista, uma proposta que consiga lidar melhor com a sobredisperssão é necessário para o estudo.

Com essas considerações então uma proposta de intervenção ao problema é a utilização do modelo Binomial Negativo para a variável. A abordagem é trabalhada na subsessão a seguir.

Modelo com resposta Binomial Negativa.

Como proposta de modelo que se adequa melhor a sobredisperssão, ou a variância superior a resposta média, considere Y_i variável aleatória de distribuição BN(μ_i, ϕ), estimando também o parâmetro de dispersão. Para

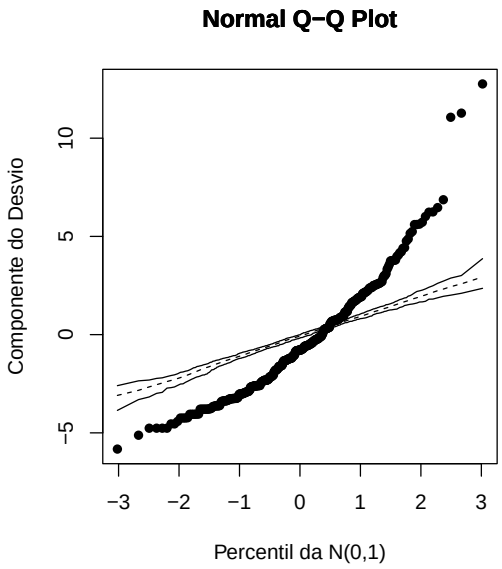

Figura 4: Envelope para o modelo Poisson

o caso de ligação canônica, aqui utilizado, sendo ele a ligação g(.) como logaritmica, definimos o modelo como sendo:

$$log(\mu_i) = \beta_0 + \beta_1 DalcL_i + \beta_2 SexM_i + \beta_3 Age_i + \beta_4 SchoolMS_i + \\ \beta_5 ReasonH_i + \beta_6 ReasonO_i + \beta_7 ReasonR_i + \beta_8 InternetY_i$$

Onde a resposta em estudo é a resposta média para número de faltas.

Seleção de Variáveis

Um importante adendo para o estudo é a forma como se deu a seleção de variáveis para o trabalho. Uma pré seletiva foi realizada a fim de reduzir consideravelmente o número de covariáveis. O método *stepAIC* com esse subgrupo foi tido como auxiliar na seletiva, com uma segunda seleção manual realizada após sua aplicação utilizando testes qui-quadrado. Resultando no modelo acima citado. Vale comentar que a variável "Medu", nível de educação da mãe, foi tida como significante para o modelo, porém devido a problema de multicoliariedade com a covariável "Dalc" teve que ser retirada do grupo final.

Ao testar a interação entre as covariáveis selecionadas arbitrariamente, nota-se que apenas a interação *age:internet* é significante a um nível *alpha* de 0.1, com um p-valor de 0.0215. Ao realizar o teste de razão de verossimilhança é notado que a interação é significante a um p-valor de 0.05 somente a um nível *alpha* de 0.01. A interação não foi mantida no modelo, considerando também queda baixa do AIC de 2168.1 para 2166.3 e a alteração do deviance de 449.63 a 386 graus de liberdade para 449.88 a 385 graus de liberdade, priorizando a parcimonia do modelo, apenas os efeitos únicos das covariáveis foram mantidos.

Ao avaliarmos a significância do modelo com a presença da covariável Sex, utilizando o teste de razão de verossimilhança a um *alpha* de 0.05, supondo a hipótese nula de $\beta_2 = 0$, não rejeitamos a hipótese nula a um p-valor de 0.12, como apresentado abaixo.

```
## Likelihood ratio tests of Negative Binomial Models
```

```
##
## Response: absences
##                                          Model     theta Resid. df
## 1       Dalc + age + school + reason + internet 0.6731100       387
## 2 Dalc + sex + age + school + reason + internet 0.6792423       386
##      2 x log-lik.   Test    df LR stat.   Pr(Chi)
## 1       -2150.501
## 2       -2148.126 1 vs 2     1 2.374769 0.1233098
```

Restando o modelo final :

$$log(\mu_i) = \beta_0 + \beta_1 DalcL_i + \beta_2 Age_i + \beta_3 SchoolMS_i + \\ \beta_4 ReasonH_i + \beta_5 ReasonO_i + \beta_6 ReasonR_i + \beta_7 InternetY_i$$

com os parâmetros estimados:

Tabela 3: Tabela de estimativas para o modelo final

	Estimativa	Erro Padrão	valor z	Pr(>|z|)
(Intercept)	-1.9343	1.0159	-1.904	0.05692
Dalclow	-0.6355	0.3116	-2.040	0.04140
age	0.2224	0.0554	4.015	5.95e-05
schoolMS	-0.7075	0.2289	-3.091	0.00200
reasonhome	0.5094	0.1655	3.077	0.00209
reasonother	0.2486	0.2467	1.008	0.31352
reasonreputation	0.4555	0.1688	2.698	0.00698
internetyes	0.3295	0.1809	1.821	0.06862

Com um deviance final de 449.27 a 387 graus de liberdade e um AIC de 2168.5.

Diagnóstico

Considerando os modelos dentro do escópo da matéria, nota-se pela Figura 5 o ganho em explicabilidade do valor esperado de ausências para o modelo com resposta Binomial Negativa se comparado com o anteriormente aplicado modelo com resposta Poisson. Note que ainda sim temos alguns poucos valores fora das bandas de confiança do envelope do modelo. Esse fato provavelmente é fruto do grande número de zeros para variável resposta, mostrando a necessidade da aplicação de um modelo Binomial Negativo inflacionado. O mesmo não foi testado pois foge do conteúdo estudado no curso. Mas ainda sim vemos uma boa explicabilidade da representação do modelo utilizado.

Levando em consideração esse perceptível ganho, o modelo de resposta binomial negativa foi tido como mais adequado. Observando a Figura 6, pode-se considerar, devido a linearidade do preditor linear $\eta = X\underline{\beta}$, a ligação realizada, logaritmica, como adequada para descrever a relação da variável resposta aos dados. É possível perceber também os possíveis candidatos a pontos de alavanca, aberrantes e de influência do modelos pelos gráficos (a),(b) e (c) da Figura 6. Esses dados podem ser observados na Tabela 4.

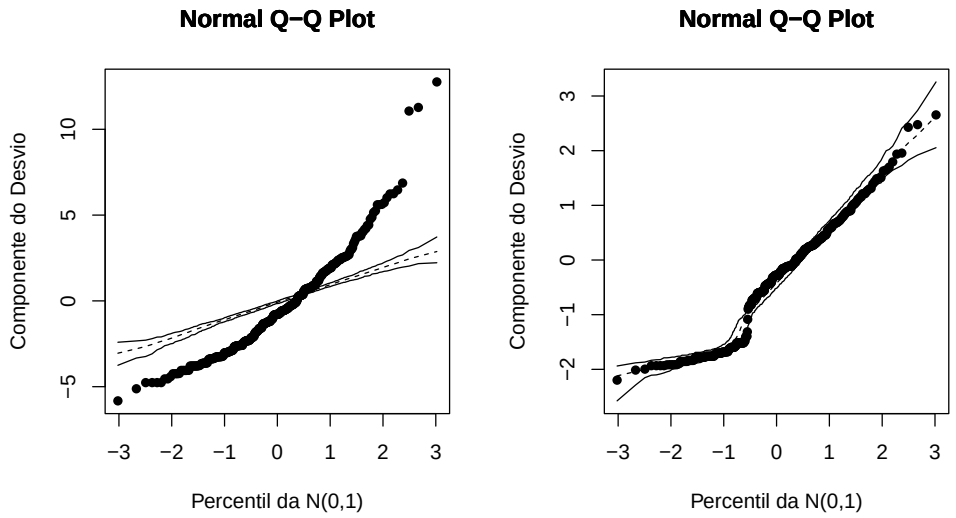

Figura 5: Envelope para modelo com resposta binomial comparado ao com resposta Poisson

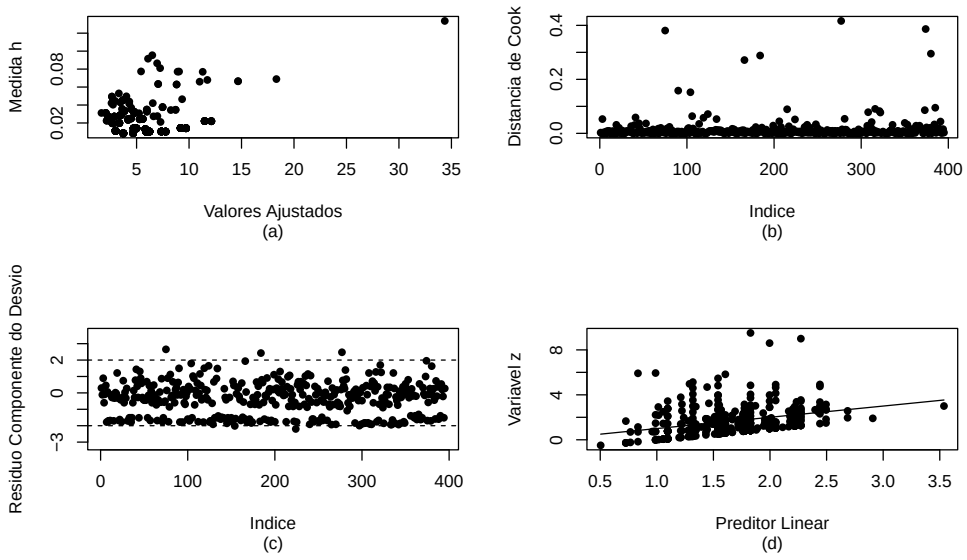

Figura 6: Diagnóstico para o modelo Binomial Negativo

Tabela 4: Dados outliers

	Dalc	sex	school	age	internet	reason	absences
248	high	M	GP	22	yes	other	16
75	low	F	GP	16	yes	home	54
166	low	M	GP	16	no	course	16
184	low	F	GP	17	yes	reputation	56
277	low	F	GP	18	yes	home	75
374	low	F	MS	17	yes	course	14
380	low	F	MS	17	yes	reputation	17
224	high	M	GP	18	yes	home	0

É possível reparar que os dados indicados, dentre eles, a observação 277,184 e 75 apresentam um alto número de faltas, valores incomuns se retomarmos a parte de análise descritiva. Uma forma de verificar o efeito da variável sobre o modelo é a remodelagem excluindo a observação em questão. Podemos observar os resultados para as estimativas dos parâmetros do modelo na Tabela 5 ao retirar cada uma das observações citadas na Tabela 4.

Tabela 5: Tabela de Estimativas retirando o valor átipico

	beta_0	beta_1	beta_2	beta_3	beta_4	beta_5	beta_6	beta_7
Original	-1.93	-0.64	0.22	-0.71	0.51	0.25	0.46	0.33
248	-2.01	-0.67	0.23	-0.72	0.51	0.26	0.46	0.33
75	-2.10	-0.65	0.23	-0.70	0.43	0.25	0.46	0.31
166	-2.09	-0.65	0.23	-0.68	0.54	0.28	0.49	0.41
184	-1.84	-0.64	0.22	-0.69	0.51	0.25	0.39	0.31
277	-1.57	-0.66	0.20	-0.66	0.45	0.24	0.46	0.31
374	-2.02	-0.68	0.23	-0.84	0.54	0.29	0.48	0.30
380	-2.01	-0.66	0.23	-0.82	0.51	0.25	0.41	0.31
224	-1.91	-0.70	0.22	-0.72	0.52	0.24	0.46	0.33

Não há presença de grandes alterações com excessão das observações 277 e 374, com a 277 se destacando de forma mais acentuada, em virtude do alto número de ausências por parte do aluno destacado.

Para esse modelo observamos os seguintes desvios padrão para esse modelo sem a observação 277:

Tabela 6: Estimativas dos parâmetros com exclusão da observação 277

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.57208	1.00913	-1.558	0.119266
Dalclow	-0.65815	0.30861	-2.133	0.032953
age	0.20237	0.05504	3.677	0.000236
schoolMS	-0.66455	0.22707	-2.927	0.003426
reasonhome	0.44660	0.16464	2.713	0.006674
reasonother	0.24492	0.24433	1.002	0.316135
reasonreputation	0.46165	0.16722	2.761	0.005767
internetyes	0.31415	0.17937	1.751	0.079870

Note porém que a ausência da observação não altera a significância de nenhum dos parâmetros estimados, todos os betas para as variáveis selecionadas permanecem informativas para o modelo.

Cortez, Paulo, e Alice Maria Gonçalves Silva. 2008. "Using data mining to predict secondary school student performance".