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

Samuel M. Medeiros*

05 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 de regressão Binomial Negativo para 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

A presença regular dos alunos nas aulas é fundamental para o sucesso acadêmico e o desenvolvimento pessoal. No entanto, ausências frequentes podem prejudicar o aprendizado e o desempenho escolar. É importante, portanto, identificar as causas dessas ausências e encontrar soluções para minimizá-las.

Nesse sentido, este trabalho se propõe a investigar as causas de ausências em escolas de ensino médio de Portugal. Será utilizado o modelo de regressão Binomial Negativo para relacionar características domésticas e pessoais dos alunos com suas faltas às aulas. A análise dos resultados permitirá avaliar a influência dessas características na frequência dos alunos nas aulas e identificar possíveis soluções para melhorar a presença dos estudantes.

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/cinco níveis e regressão.

Este estudo apresenta uma metodologia baseada em modelos de regressão generalizados 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.

## Metodologia

### Modelo

O modelo de regressão binomial negativo é um tipo de modelo de regressão que é usado para prever uma variável inteira variando de 0 a infinito, ou seja, uma variável que pode ter apenas valores como contagem

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

(por exemplo, número de biscoitos em um pacote, ou número de sucessos ou falhas em um evento, etc.). Nesse modelo, a probabilidade de ocorrência da variável é relacionada a uma ou mais variáveis independentes através de uma função matemática.

No caso de uma única variável independente, a função matemática é geralmente a função logística, que mapeia a entrada da variável independente em uma probabilidade entre 0 e 1. No caso de várias variáveis independentes, o modelo de regressão binomial negativo pode ser implementado como uma regressão logística múltipla.

A função da regressão binomial negativa é dada por:

$$\log(y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

onde $y_i$ é a resposta observada a ser estudada, $x_{j,i}$ é a j-ésima variável explicativa e $\beta_j$ são os parâmetros do modelo a serem estimados. A função densidade da Binomial Negativa é dada por:

$$f(y_i|\mu_i, \phi) = \frac{\Gamma(\phi + y_i)}{\Gamma(\phi)\Gamma(y_i + 1)} \left(\frac{\mu_i}{\phi + \mu_i}\right)^{y_i} \left(\frac{\phi}{\phi + \mu_i}\right)^{\phi}, \quad y_i = 0, 1, 2, ..$$

Onde $y_i$ representa o número de faltas para cada aluno.

### Seleção de Variáveis

A análise dos dados originais foi revisada com o objetivo de selecionar as variáveis mais relevantes para explicar o número de faltas. Esta seleção foi baseada em conhecimentos prévios sobre o assunto e levou em consideração a importância de cada variável para a explicação do fenômeno estudado.

A partir desse subconjunto foi utilizado o método backward. O método de seleção de variáveis backward, também conhecido como eliminação backwards, é uma técnica de seleção de variáveis que começa com todas as variáveis potenciais incluídas no modelo e, aos poucos, remove as variáveis menos importantes. O processo é repetido até que apenas as variáveis mais significativas estejam presentes no modelo. A critério para a eliminação de uma variável pode ser baseada em estatísticas, como o p-valor, ou em outras métricas de desempenho do modelo, como o erro de predição ou o coeficiente de determinação. O método de seleção de variáveis backward é útil quando há uma grande quantidade de variáveis disponíveis e se deseja reduzir o número de variáveis sem sacrificar a capacidade de explicar a variável resposta.

Considerando o novo conjunto então seleção a partir de comparação de modelos resultou em um número ainda menor de variáveis selecionadas, sendo essas apresentadas na sessão seguinte, Banco de Dados.

## Banco de Dados

### Variáveis selecionadas

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.

Propondo um modelo com as faltas como variável resposta, a partir de conhecimentos prévios e de métodos de seleção de variáveis discutidos na seção *Metodologia*, tentamos explicar a variável dependente através das variáveis independentes selecionadas, fazendo uma filtragem em relação as variáveis originais nos resta então as apresentadas na Tabela 1.

Tabela 1: Variaveis selecionadas para analise

| Codigo | Variavel | Classificacao |
|---|---|---|
| school | Escola | Categorica |
| 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 variáveis categóricas. Podemos identificar para algumas variáveis que temos uma concentração maior de dados, como a categoria "Mãe" para a variável "Reason".

```
## [1] 15
```

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 | Perto de casa | 145 | 0.37 |
| | home | Reputacao | 109 | 0.28 |
| | other | Preferencia de curso | 36 | 0.09 |
| | reputation | Outro | 105 | 0.27 |
| Dalc | 1 | muito baixo | 276 | 0.7 |
| | 2 | baixo | 75 | 0.19 |
| | 3 | moderado | 26 | 0.07 |
| | 4 | alto | 9 | 0.02 |
| | 5 | muito alto | 9 | 0.02 |
| 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). Levando esse fato em consideração, dois modelos foram analisados, com e sem a variável 'School', sendo o resultado dos dois modelos apresentados no trabalho. Observe a afirmação na Tabela 2.

A Figura 2 mostra a distribuição de faltas em relação à variável acesso a internet, e pode ser visto que a tendência é semelhante para ambas as classificações. A maior porcentagem de dados se concentra em casos onde o responsável pelo aluno é a mãe, enquanto há uma distribuição mais equilibrada para quando o responsável é o pai ou outra pessoa. No entanto, ao considerarmos a quantidade de dados disponíveis, 68% dos dados da variável "responsável" são classificados como "mãe", o que justifica o comportamento observado. Podemos observar tambem que a variável sexo não interfere tanto na distribuição da variável uso de internet e número de faltas.

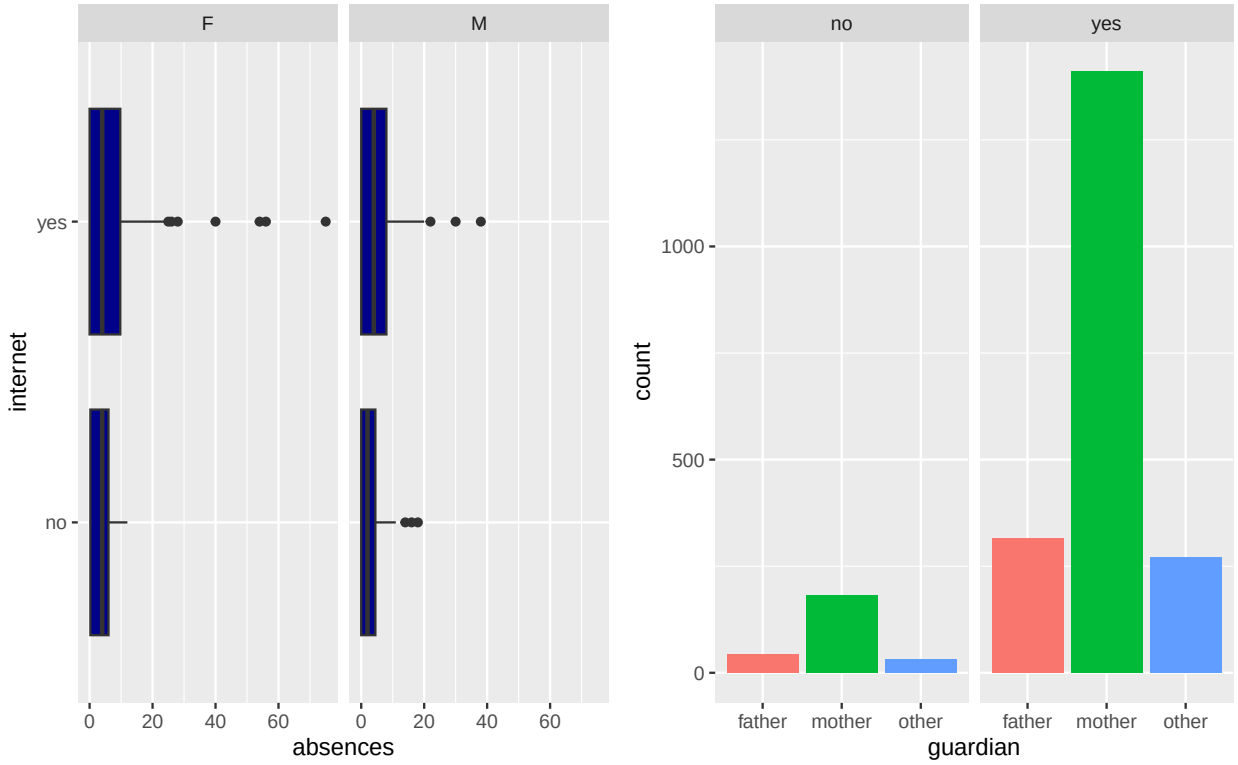


Figura 1: Número total de faltas por responsável e acesso a internet.

É possível olha a distribuição de faltas absoluta na Figura 2, onde identificamos a presença de possíveis três ou quatro outliers, apresentados na Tabela 3, onde identificamos um possível padrão de sexo, escola e acesso a internet, porém como já discutido o fator escola se dá pela quantidade de dados apresentados para a categoria em questão.

Tabela 3: Possiveis Outliers

| | school | sex | reason | Dalc | age | internet |
|---|---|---|---|---|---|---|
| 75 | GP | F | home | low | 16 | yes |
| 184 | GP | F | reputation | low | 17 | yes |
| 277 | GP | F | home | low | 18 | yes |

## Aplicação do modelo

A análise dos dados realizada na sessão anterior permitiu identificar o impacto de cada variável em relação ao número de faltas dos estudantes. Para isso, foram utilizados dois modelos diferentes: o primeiro analisou o efeito puro de cada variável, sem considerar suas interações, enquanto o segundo modelo considerou o efeito combinado de todas as variáveis.

Ao avaliar o modelo completo, foram realizadas análises ANOVA e testes de valor-p para identificar as covariáveis mais relevantes. O resultado apontou que a variável sexo não possui um impacto significativo na quantidade de faltas dos estudantes, hipótese essa testada pela razão de verossimilhança. Dessa forma, a variável sexo não foi mantida no modelo final.

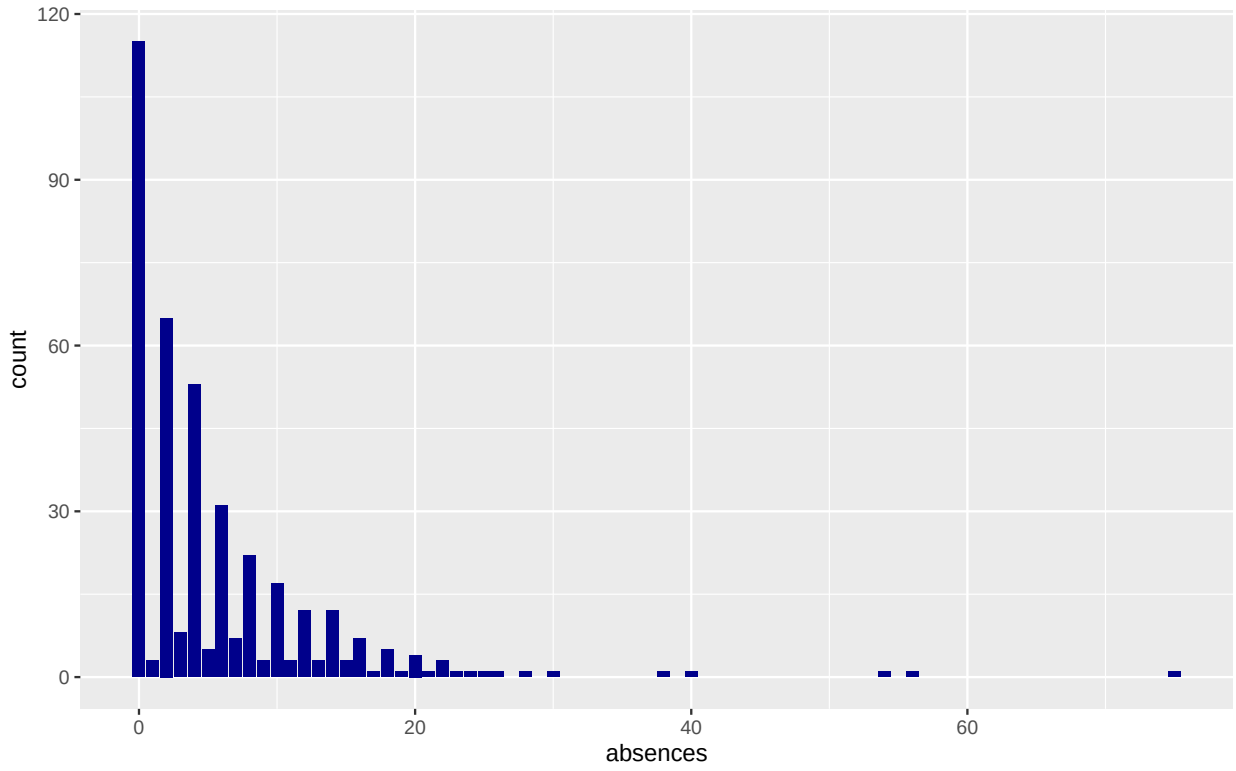


Figura 2: Frequencia de faltas

Além disso, ao comparar o modelo inteirativo com o modelo de efeito puro, pode-se perceber que a inclusão da interação idade e acesso a internet agregou informações ao modelo, tornando-o mais preciso. Esse resultado indica que a combinação entre a idade do estudante e o acesso à internet exercem uma influência relevante sobre a variável resposta, ou seja, a quantidade de faltas.

Com essa alteração o efeito puro das duas variáveis em questão se torna pouco significativo para o modelo, tendo um deviance Nulo parecido para ambos os dois modelos. Em deterimento disso o modelo com interação foi o selecionado como mais adequado a explicação dos dados.

Outro importante resultado vem da variável "Dalc" consumo de alcool. Foi identificado que apenas um dos fatores era significante, logo, uma alternativa razoável foi a alteração de 5 níveis de fator para apenas 2, "high"- alto consumo de alcool e "low"- baixo consumo de alcool. A variável 'Medu', grau de estudo da Mãe, teve que ser retirada do banco de dados devido a incapacidade de inversão da matriz de covariáveis com a presença da covariável.

Por fim, é importante destacar que a seleção de variáveis é um passo crucial na modelagem de dados, pois permite identificar quais variáveis são realmente relevantes para explicar o fenômeno em questão. Onde obtivemos o modelo a seguir:

$$log(Absences) = \beta_0 + \beta_1 Reason_h + \beta_2 Reason_o + \beta_3 Reason_r + \beta_4 Dalc_low + \beta_5 Age : Internet_n + \beta_6 Age : Internet_s$$

As estimativas para cada um dos $\beta_i$ bem como os desvios, níveis de significancia e Deviânce do modelo podem ser observados abaixo.

```
##
## Call:
```

```
## glm.nb(formula = absences ~ reason + school + Dalc + age:internet,
##     data = mat, link = log, init.theta = 0.6743475915)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1258  -1.5108  -0.2756   0.2535   2.6414
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.64615    0.99139  -1.660 0.096825 .
## reasonhome        0.50800    0.16541   3.071 0.002132 **
## reasonother       0.24586    0.24651   0.997 0.318582
## reasonreputation  0.45151    0.16876   2.675 0.007463 **
## schoolMS         -0.69861    0.22890  -3.052 0.002273 **
## Dalclow          -0.64157    0.31126  -2.061 0.039285 *
## age:internetno    0.20442    0.05542   3.688 0.000226 ***
## age:internetyes   0.22551    0.05553   4.061 4.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.6743) family taken to be 1)
##
##     Null deviance: 493.83  on 394  degrees of freedom
## Residual deviance: 449.31  on 387  degrees of freedom
## AIC: 2168
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.6743
##           Std. Err.:  0.0614
##
##  2 x log-likelihood:  -2149.9890
```

Apresentando um deviance final de 449.31 para 384 graus de liberdade.

### Resultados e Diagnóstico

A escolha pelo modelo de regressão binomial negativo foi baseada na sua adequação para o conjunto de dados em questão. Essa adequação foi verificada pela análise da Figura 3, que mostrou que a função de ligação parece estar correta para o banco de dados. Além disso, a análise dos diagnósticos do modelo (Figura 4) permitiu identificar a presença de 3 outliers, que foram mencionados previamente na análise exploratória dos dados. Além disso, a Figura 4 também permite verificar a adequação do modelo, bem como a suposição de linearidade para a variável resposta.

## Conclusões

A análise dos dados coletados aponta para uma relação positiva entre o número de faltas e o consumo de álcool. Alunos que apresentam um alto nível de consumo de álcool tendem a ter um número maior de faltas, o que também pode ser observado na comparação entre as escolas. Aqueles que estudam na escola "MS" tendem a ter um número menor de faltas, possivelmente devido à falta de dados iguais para as duas escolas.

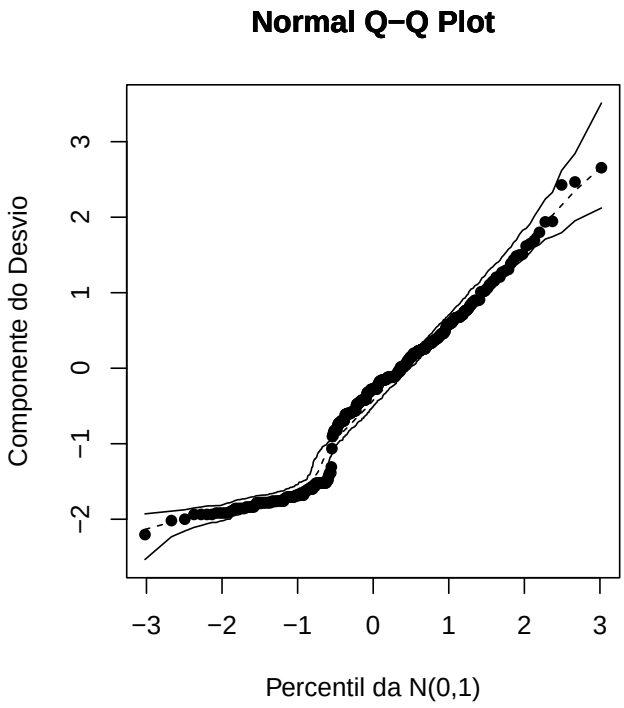


Figura 3: Envelope do modelo binomial negativo

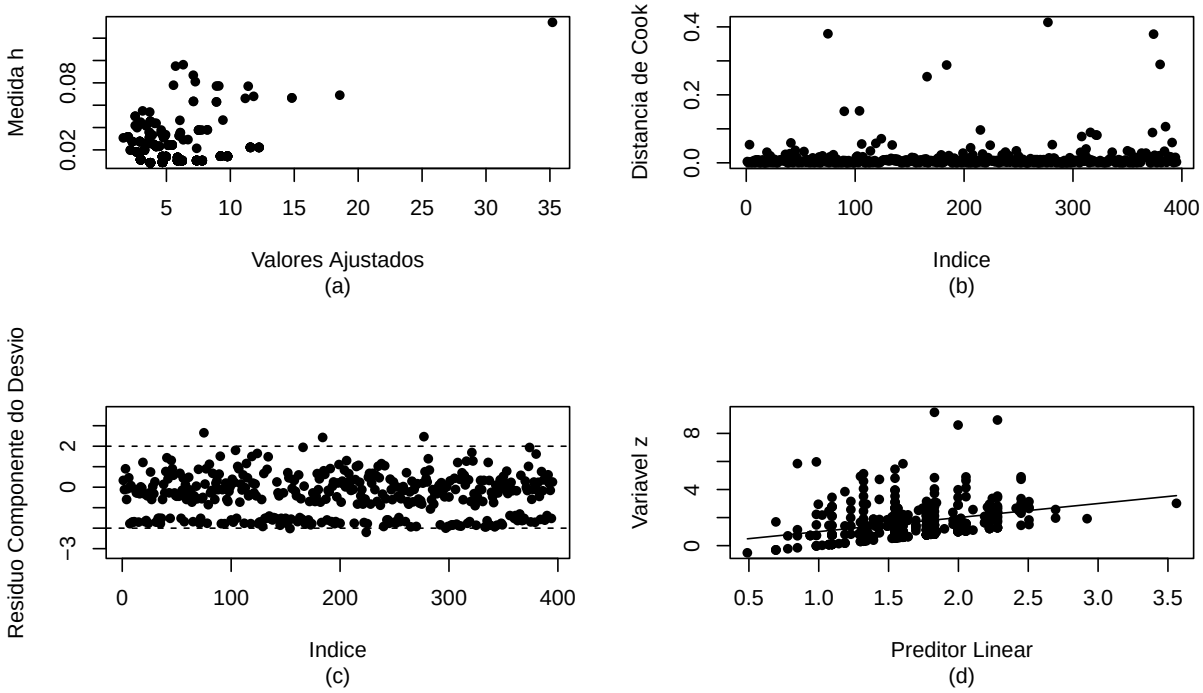


Figura 4: Diagnostico do modelo

Neste contexto, iniciativas de prevenção ao uso excessivo de álcool nas escolas são uma alternativa eficaz para reduzir o número de faltas dos estudantes. Campanhas e ações de prevenção podem ser benéficas para educar os estudantes sobre os riscos e as consequências do consumo excessivo de álcool, além de ajudar a diminuir o número de faltas.

Esse fato também é estudado e apresentado no trabalho citado como referência de estudo. A variável tempo livre fora da escola não se mostrou tão eficaz mas possívelmente seria capaz de descrever uma relação entre o consumo de alcool dos alunos. Um estudo em cima das motivações para a fuga e a embriagues excessíva são sugestões para estudo da melhor forma de combater essas ausências. Vemos também que alunos com acesso a internet tem uma tendência maior a números altos de faltas. Controle do acesso pelos pais são outras alternativas capazes de contornar melhor o problema.

# Appendix

## Códigos

```
############ funcoes ####################

envelope.poi <- function(fit.model){
par(mfrow=c(1,1))
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
w <- fit.model$weights
W <- diag(w)
H <- solve(t(X)%*%W%*%X)
H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h <- diag(H)
td <- resid(fit.model,type="deviance")/sqrt((1-h))
e <- matrix(0,n,100)
#
for(i in 1:100){
  nresp <- rpois(n, fitted(fit.model))
  fit <- glm(nresp ~ X, family=poisson)
  w <- fit$weights
  W <- diag(w)
  H <- solve(t(X)%*%W%*%X)
  H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
  h <- diag(H)
  e[,i] <- sort(resid(fit,type="deviance")/sqrt(1-h))}
#
e1 <- numeric(n)
e2 <- numeric(n)
#
for(i in 1:n){
  eo <- sort(e[i,])
  e1[i] <- (eo[2]+eo[3])/2
  e2[i] <- (eo[97]+eo[98])/2}
#
med <- apply(e,1,mean)
faixa <- range(td,e1,e2)
par(pty="s")
```

```
qqnorm(td,xlab="Percentil da N(0,1)",
       ylab="Componente do Desvio", ylim=faixa, pch=16)
par(new=T)
#
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2) }
envelope.bn <- function(fit.model){
  # par(mfrow=c(1,1))
  X <- model.matrix(fit.model)
  n <- nrow(X)
  p <- ncol(X)
  fi <- fit.model$theta
  w <- fi*fitted(fit.model)/(fi + fitted(fit.model))
  W <- diag(w)
  H <- solve(t(X)%*%W%*%X)
  H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
  h <- diag(H)
  td <- resid(fit.model,type="deviance")/sqrt(1-h)
  fi <- fit.model$theta
  e <- matrix(0,n,100)
  #
  for(i in 1:100){
    resp <- rnegbin(n, fitted(fit.model),fi)
    fit <- glm.nb(resp ~ X, control = glm.control(maxit = 50))
    w <- fit$weights
    W <- diag(w)
    H <- solve(t(X)%*%W%*%X)
    H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
    h <- diag(H)
    e[,i] <- sort(resid(fit,type="deviance")/sqrt(1-h))}
  #
  e1 <- numeric(n)
  e2 <- numeric(n)
  #
  for(i in 1:n){
    eo <- sort(e[i,])
    e1[i] <- (eo[2]+eo[3])/2
    e2[i] <- (eo[97]+eo[98])/2}
  #
  med <- apply(e,1,mean)
  faixa <- range(td,e1,e2)
  par(pty="s")
  qqnorm(td,xlab="Percentil da N(0,1)",
         ylab="Componente do Desvio", ylim=faixa, pch=16)
  par(new=T)
  #
  qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1)
  par(new=T)
  qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1)
  par(new=T)
```

```
  qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2)
}
diagnostico.bn <- function(fit.model){
  X <- model.matrix(fit.model)
  n <- nrow(X)
  p <- ncol(X)
  fi <- fit.model$theta
  w <- fi*fitted(fit.model)/(fi + fitted(fit.model))
  W <- diag(w)
  H <- solve(t(X)%*%W%*%X)
  H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
  h <- diag(H)
  ts <- resid(fit.model,type="pearson")/sqrt(1-h)
  td <- resid(fit.model,type="deviance")/sqrt(1-h)
  di <- (h/(1-h))*(ts^2)
  par(mfrow=c(2,2))
  a <- max(td)
  b <- min(td)
  plot(fitted(fit.model),h,xlab="Valores Ajustados", ylab="Medida h",
       pch=16)
  identify(fitted(fit.model), h, n=5)
  title(sub="(a)")
  #
  plot(di,xlab="Indice", ylab="Distancia de Cook", pch=16)
  identify(di,n=3)
  title(sub="(b)")
  #
  plot(td,xlab="Indice", ylab="Residuo Componente do Desvio",
       ylim=c(b-1,a+1), pch=16)
  abline(2,0,lty=2)
  abline(-2,0,lty=2)
  identify(td,n=1)
  title(sub="(c)")
  #
  eta = predict(fit.model)
  z = eta + resid(fit.model, type="pearson")/sqrt(w)
  plot(predict(fit.model),z,xlab="Preditor Linear",
       ylab="Variavel z", pch=16)
  lines(smooth.spline(predict(fit.model), z, df=2))
  title(sub="(d)")
}

########## IMPORTACAO E PACOTES ##############

library(corrplot)
library(dplyr)
library(ggplot2)

###############################################

mat <- read.csv('dados/Maths.csv')
# selecao manual de variaveis #################
mat |> summary()
```

```r
mat |> colnames()
mat[,mat |> sapply(is.character) | mat|>sapply(is.integer)] |> colnames()
mat$reason |> unique()
mat[sapply(mat, is.character)] <- lapply(mat[sapply(mat, is.character)],
                                         as.factor)

y <- c('absences')
x <- c('school','sex','age','famsize','famsize','Medu','Fedu','reason','guardian','traveltime',
       'failures','schoolsup','famsup','activities','higher','internet','romantic','famrel','freetime',
       'Dalc')
mat <- mat[,c(x,y)]
mat[,'Dalc'] <- mat$Dalc |> as.character()
mat[mat$Dalc >3,'Dalc'] <- 'high'
mat[mat$Dalc <=3,'Dalc'] <- 'low'
str(mat)
cols <- c('famrel','freetime','goout','Dalc','Medu','Fedu')
mat[,cols] <- lapply(mat[,cols],as.factor)
##  exploratoria ##############################

str(mat)

mat |> ggplot() +
  aes(x = absences, y = internet) +
  geom_boxplot(fill = "darkblue") +
  theme_minimal()
mat |> ggplot() +
  aes(x = Medu) +
  geom_boxplot(fill = "darkblue") +
  theme_minimal()
mat |> ggplot() +
  aes(x = absences, y = internet) +
  geom_boxplot(fill = "darkblue") +
   facet_wrap(vars(sex))
ggplot(mat) +
  aes(x = internet, weight = absences) +
  geom_bar(fill = "#112446") +
  theme_minimal() +
  facet_wrap(vars(sex))
mat |> ggplot() +
    aes(school, fill = school) + geom_bar(show.legend = F) + facet_wrap(vars(sex))

mat |> ggplot() +
  aes(age) + geom_bar() + facet_wrap(vars(failures))

mat |> ggplot() +
  aes(guardian, weight =absences, fill = guardian) + geom_bar(show.legend = F) + facet_wrap(vars(interne

mat |> ggplot() +
  aes(guardian) + geom_bar()
mat |> ggplot() +
  aes(x = absences, y = guardian) +
  geom_boxplot(fill = "darkblue") +
  theme_minimal()
```

```r
mat |> ggplot() +
  aes(x = reason, fill = absences) +
  geom_bar() +
  scale_fill_gradient() +
  theme_minimal()

 mat |> ggplot() +
  aes(school) + geom_bar(fill ='#112350') +
   ggtitle( 'Frequencia por Escola')

table(mat$school)
# NAO SEI SE VALE A
#PENA INCLUIR ESCOLA POR CAUDA DA DIFERENCA
#DE OBSERVACOES PRA CADA UMA
### selecao de variaveis pelo step ################################
glm(absences ~. ,mat, family='poisson') |> step(direction = 'backward')

fit1 <- glm(formula = absences ~ school + sex + age + famsize + Medu +
      reason + guardian + traveltime + schoolsup + higher + internet +
      romantic + famrel + freetime + Dalc, family = "poisson",
    data = mat)

envelope.poi(fit1)

######### modelo binomial negativo ##############################
library(MASS)
str(mat)
fit2 <- glm.nb(formula = absences ~ school + sex + age + famsize + Medu +
      reason + guardian + traveltime + schoolsup + higher + internet +
      romantic + famrel + freetime + Dalc,
    data = mat)

fit2 |> summary()

glm.nb(absences ~ ., data = mat ) |> step(direction = 'backward')

fit3 <- glm.nb(absences ~ school + sex + age + Medu + reason +
         internet + famrel + Dalc, data = mat,
         init.theta = 0.7000362531,
       link = log)

fit3 |> summary()

fit3.inter <- glm.nb(absences ~ (school + sex + age + Medu + reason +
                internet + famrel + Dalc)^2, data = mat,
                init.theta = 0.7000362531,
               link = log)
fit3.inter |> summary()
fit4 <- glm.nb(absences ~ school + sex + age + Medu + reason +
                internet  + Dalc, data = mat,
               init.theta = 0.7000362531,
               link = log)
fit4 |> summary()
```

```
anova(fit4,fit3,test = 'LR')

fit5 <-  glm.nb(absences ~   sex + age + Medu + reason +
                  internet  + Dalc, data = mat,
                init.theta = 0.7000362531,
                link = log)
fit5 |> summary()
anova(fit5,fit4)

fit4.int   <- glm.nb(absences ~ (school + sex + age + Medu + reason +
                             internet  + Dalc)^2, data = mat, init.theta = 0.7000362531,
                           link = log)
fit4.int |> summary()

fit4.int   <- glm.nb(absences ~ school + sex + age + Medu + reason +
                           internet  + Dalc + age*internet, data = mat,
                     init.theta = 0.7000362531,
                           link = log)
fit4.int |> summary()

anova(fit4,fit4.int,test = 'Chisq')

fit5 <- glm.nb(absences ~ school + sex  + Medu + reason +
                   Dalc + age:internet, data = mat, init.theta = 0.7000362531,
                link = log)
fit5 |> summary()
fit4 |> summary()
fit5 |> diagnostico.bn()
par(mfrow = c(1,2))
fit5 |> envelope.bn()
fit4 |> envelope.bn()
anova(fit4,fit5,test = 'Chisq')
fit6 <- glm.nb(absences ~  sex  + Medu + reason +
                 Dalc + age:internet, data = mat, init.theta = 0.7000362531,
                link = log)
fit7 <- glm.nb(absences ~  sex  + Medu + reason + freetime +
                 Dalc + age:internet, data = mat, init.theta = 0.7000362531,
                link = log)
fit6 |> summary()
fit8 <- glm.nb(absences ~    reason + school +
                 Dalc + age:internet, data = mat, init.theta = 0.7000362531,
                link = log)
  anova(fit8,fit6,test = 'Chisq')
fit8 |> summary()
```

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