Perfil dos jovens de 15 a 17 anos assassinados em Minas Gerais no ano de 2013

Estudo via Regressão Logística

Fernando de Souza Bastos [1]

1 Resumo

O objetivo deste trabalho é analisar o perfil dos jovens assassinados em Minas Gerais no ano de 2013 via regressão logística binária. Os dados analisados foram coletados pelo Datasus e divulgados via Sistema de Informação de Mortalidade (SIM), disponibilizado pelo Ministério da Saúde através da Fundação Nacional de Saúde. Após o ajuste do modelo estimou-se que um jovem de 17 anos, do sexo masculino, solteiro, de cor negra e com nenhuma escolaridade tem 72,19% de probabilidade de ter sido assassinado, dado que está morto. Para o ajuste e análise do modelo foi utilizado o software R ([8]).

2 Introdução

De acordo com [3], diversos estudos têm apontado para a existência de um crescimento real da violência no Brasil, em particular das mortes por homicídios, desde o final da década de 1970. As regiões geográficas e seus respectivos municípios, principalmente as grandes cidades, apresentam um aumento na mortalidade por causas externas a partir da década de 1990.

Neste contexto, diversas são as notícias de violência cometida pelos jovens e contra os jovens no país, principalmente na faixa de 15 a 19 anos. De acordo com [1] e [9], o Brasil ocupa o terceiro lugar em relação a 85 países no ranking de mortes de adolescentes de 15 a 19 anos, perdendo apenas para México e El Salvador. São 54,9 mortes a cada 100 mil jovens. Na faixa de 16 e 17 anos, a taxa de mortalidade ficou em 54,1 homicídios por 100 mil adolescentes em 2013, um aumento de 2,7% em relação a 2012 e de 38,3% na década.

De acordo com [9], o homicídio é uma das principais causas de mortes de adolecentes de 16 a 17 anos no país e quando se faz uma análise dos dados de Minas Gerais do sistema de mortalidade do Datasus, conforme imagem (a) da figura 1, é possível notar que o maior número de assassinatos ocorre aos 17 anos de idade. Neste sentido, o presente trabalho tem a finalidade de caracterizar o perfil quanto a estado civil, escolaridade, sexo, raça/cor e idade (15 a 17 anos) dos jovens que morreram assassinados no estado de Minas Gerais no ano de 2013.

[1] Doutorando UFMG. e-mail: *fsbmat@gmail.com*

3 Material e Métodos

A informação de mortalidade é uma das mais importantes na área da saúde, pois o óbito é um evento único e seu registro obrigatório. No Brasil, o Ministério da Saúde através da Fundação Nacional de Saúde e do Datasus são responsáveis por divulgar os dados de mortalidade do país por meio do sistema de informação da mortalidade (SIM).

A fonte de informação primária da base de dados são os atestados de óbito emitidos pelos cartórios civis. Esta base contém informações sobre a data do óbito, idade, sexo, estado civil, local de ocorrência, causa de mortalidade, município de residência, ocupação e escolaridade. Apesar da enorme quantidade de informações, o banco de dados apresenta problemas sérios de preenchimento de algumas variáveis como educação, estado civil, ocupação, entre outras, que dificultam o seu uso.

Neste trabalho utilizou-se apenas as variáveis consideradas prioritárias pelo Ministério da Saúde, idade, sexo, estado cívil, escolaridade e causa de mortalidade, nas quais os valores não preenchidos foram retirados do estudo. A causa de mortalidade está codificada segundo a Classificação Internacional de Doenças através da CID10.

As descrições das codificações estão na tabela (1), abaixo:

Tabela 1: Variáveis consideradas no estudo com suas respectivas categorias

Variáveis	Categorias	Descrição
Y	0	Morte registrada como causas distintas de homicídio
	1	Morte registrada como homicídio
S	1	Masculino
	2	Feminino
R	1	Raça/Cor Branca
	2	Raça/Cor Negra
	4	Raça/Cor Parda
	5	Raça/Cor Indígena
ESC	1	Nenhum estudo
	2	1 a 3 anos de estudo
	3	4 a 7 anos de estudo
	4	8 a 11 anos de estudo
	5	12 ou mais anos de estudo
I	Idade (Contínua)	15 a 17 anos

Como havia poucos indivíduos na categoria distinta de solteiro para a variável estado civil, foram considerados somente indíviduos solteiros na análise. Não houve nenhum indivíduo de 15 a 17 anos caracterizado com a raça/cor amarela.

Foram coletados os dados de 5.418 jovens de 15 a 17 anos que morreram no ano de 2013 no estado de Minas Gerais, para a análise de regressão logística foram retirados 62 indivíduos (aproximadamente 1,15% dos dados) por não terem todas as informações completas e por terem informações cuja categoria era caracterizada como “ignorado” no banco de dados. Dessa forma, restaram 5356 indívíduos para análise. Abaixo algumas estatísticas descritivas comentadas dos dados.

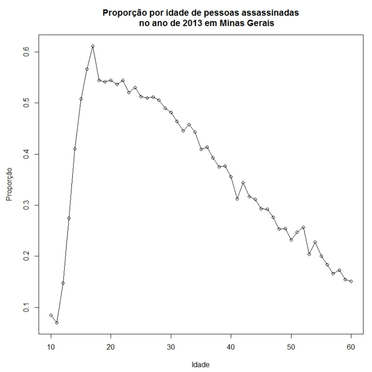
4 Resultados e Discussões

4.1 Estatística Descritiva

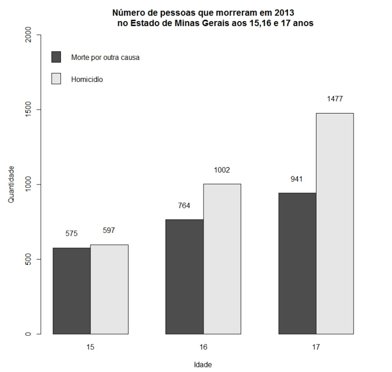
Os dados abaixo apresentam a quantidade de indívíduos em cada categoria das respectivas variáveis estudadas.

Y		S	R	ESC	I	
Min.	:0.0000	1:4666	1:1503	1: 64	Min.	:15.00
1st Qu.	:0.0000	2: 690	2: 398	2: 890	1st Qu.	:16.00
Mediana	:1.0000		4:3428	3:3030	Mediana	:16.00
Média	:0.5743		5: 27	4:1351	Média	:16.23
3rd Qu.	:1.0000			5: 21	3rd Qu.	:17.00
Max.	:1.0000				Max.	:17.00

Na imagem (a) do gráfico 1 temos a proporção de pessoas assassinadas por idade, dos 10 aos 60 anos. Note que o número de homicídios aumenta até os 17 anos e a partir dai começa a diminuir lentamente, haja visto que a morte por assassinato dos 15 aos 17 anos é superior a morte causada por outras causas, como pode ser verificado na imagem (b) do mesmo gráfico.



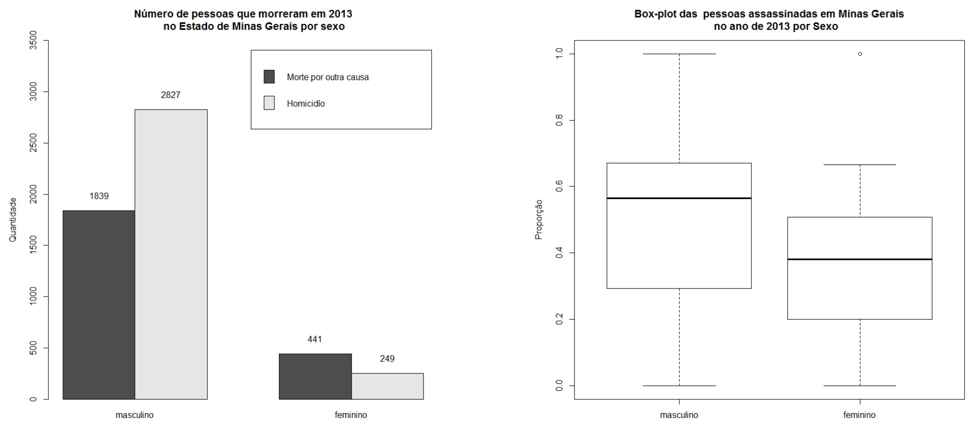
(a) Proporção por idade de pessoas, entre 10 e 60 anos, que morreram em 2013 no Estado de Minas Gerais.



(b) Número de jovens de 15 a 17 anos assassinados em 2013 no Estado de Minas Gerais.

Figura 1

A imagem (a) do gráfico 2 mostra o número de jovens de 15 a 17 anos, discriminados por sexo, que morreram por motivos distintos de homicídio e por homicídio. Note que o número de jovens, do sexo masculino, nesta faixa etária que morrem assassinados, é maior do que o número de jovens que morrem por outras causas.

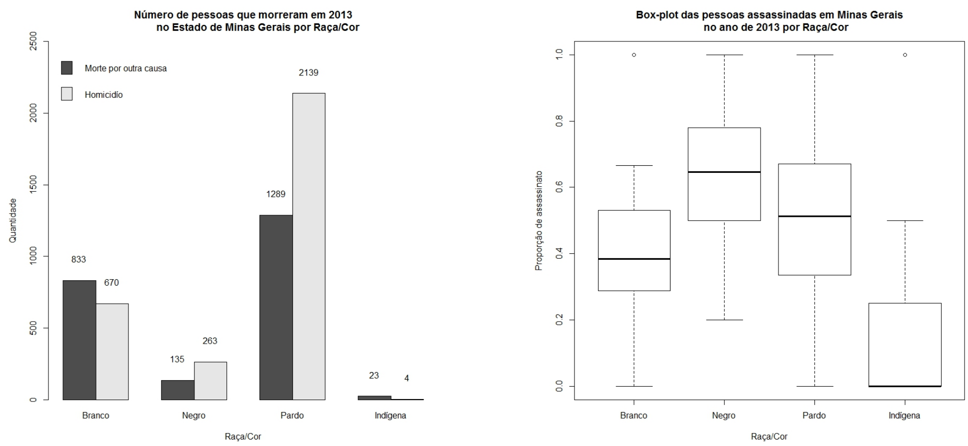


(a) Número de jovens de 15 a 17 anos que morreram em 2013 discriminado por sexo.

(b) Box- Plot da proporção de jovens de 15 a 17 anos que morreram em 2013 em relação ao total de jovens mortos discriminado por sexo.

Figura 2

A imagem (a) do gráfico 3 mostra o número de jovens de 15 a 17 anos, discriminados por raça/cor, que morreram por motivos distintos de homicídio e por homicídio. Note que o número de jovens, de cor negra e/ou parda, nesta faixa etária que morrem assassinados, é maior do que o número de jovens que morrem por outras causas.

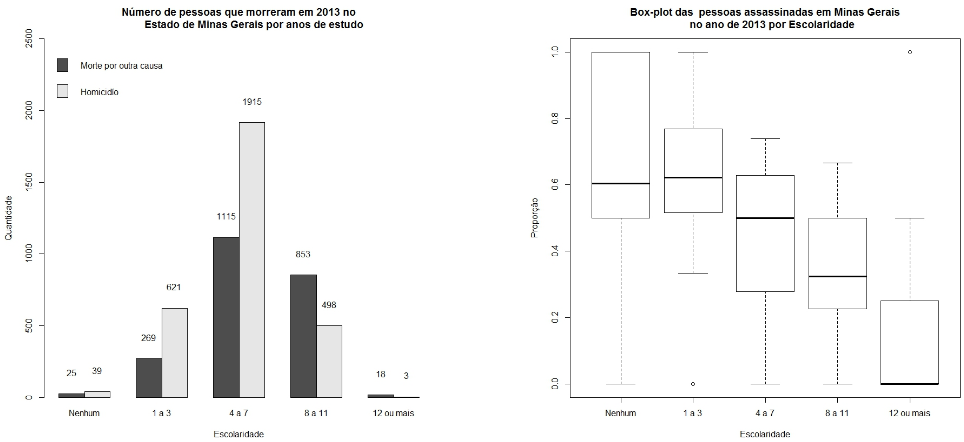


(a) Número de jovens de 15 a 17 anos que morreram em 2013 discriminado por raça/cor

(b) Box- Plot da proporção de jovens de 15 a 17 anos que morreram em 2013 em relação ao total de jovens mortos discriminado por raça/cor

Figura 3

A grande maioria dos jovens de 15 a 17 anos que morreram assassinadas em Minas Gerais no ano de 2013 tinham somente de 1 a 7 anos de estudo, como mostra a imagem (a) do gráfico 4. No box-plot da imagem (b) é possível constatar que a média da proporção de mortes por escolaridade tem uma semelhança para os indivíduos com 1 a 3 anos de estudo com os indivíduos com nenhum ano de estudo, provavelmente, o parâmetro de ESC2 não será significativo.



(a) Número de jovens de 15 a 17 anos que morreram em 2013 discriminado por escolaridade

(b) Box- Plot da proporção de jovens de 15 a 17 anos que morreram em 2013 em relação ao total de jovens mortos discriminado por escolaridade

Figura 4

4.2 Ajuste do melhor modelo de regressão logística

A construção iniciou-se com a análise univariada das variáveis explicativas. Para saber se a variável analisada era significativa ou não, foi realizado um teste baseado na estatística G para a deviance. Tal estatística possui distribuição qui-quadrado com 1 grau de liberdade. Todas as variáveis tiveram p-valor inferior a 0.05 e portanto foram selecionadas para a análise multivariada. Em seguida, foi realizada a análise indivídual dos dados e todos os modelos possiveis, com e sem interação, foram avaliados, apresentamos no apêndice todos os modelos testados, juntamente com o código explicado. Para maiores informações sobre regressão logística binária sugere-se [2], [4], [6] e [7].

O melhor modelo foi aquele que, além de estar bem ajustado, foi o mais simples, ou seja, possuía menos variável não significativa, além de possuir deviance mais próxima do número de graus de liberdade e menor AIC. Utilizou-se ainda o teste de ajuste do Qui-quadrado de Pearson e o de Hosmer-Lemeshow, ver [6] para maiores detalhes.

O teste de Hosmer-Lemeshow é um teste que avalia o modelo ajustado comparando as frequências observadas e as esperadas. O teste associa os dados as suas probabilidades estimadas da mais baixa a mais alta, então faz um teste qui quadrado para determinar se as frequências observadas estão próximas das frequências esperadas. Já o teste de Pearson, é utilizado para fazer análise dos resíduos para modelos logísticos, trata-se de uma medida útil para avaliar o quão bem o modelo selecionado ajustou-se aos dados.

As hipóteses testadas foram:

- Ho: O modelo é adequado
- H1: O modelo não é adequado

O p-valor do teste do qui-quadrado de Pearson foi 0.4151227 e do teste de Hosmer-Lemeshow 0.1109, sendo assim, os dois testes consideram o modelo que foi selecionado adequado para análise dos dados.

A saída do R abaixo apresenta o conjunto de variáveis que compõem o modelo ajustado junto com os valores estimados dos coeficientes do modelo, o erro padrão dos coeficientes, os quantis Z da distribuição normal padrão e o p-valor de todos os parâmetros:

```
Call: glm(formula = Y ~ S + R + ESC + I, family = binomial(link = logit),
    data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7491  -1.1673   0.8069   0.9369   2.2899

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.54342    0.66772  -5.307 1.12e-07 ***
S2          -0.71483    0.08940  -7.996 1.28e-15 ***
R2           0.64436    0.12215   5.275 1.33e-07 ***
R4           0.54040    0.06591   8.199 2.42e-16 ***
R5          -1.38578    0.55451  -2.499  0.01245 *
ESC2         0.33147    0.27559   1.203  0.22906
ESC3         0.09528    0.26815   0.355  0.72235
ESC4        -0.87388    0.27176  -3.216  0.00130 **
ESC5        -2.14116    0.68340  -3.133  0.00173 **
I            0.22666    0.03733   6.071 1.27e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 7306.3  on 5355  degrees of freedom
Residual deviance: 6744.2  on 5346  degrees of freedom
AIC: 6764.2
```

As categorias ESC2 e ESC3 foram não significativas, neste caso foi realizado o agrupamento das duas e avaliado um novo ajuste com a covariável escolaridade com somente quatro categorias, porém a categoria nova permaneceu não significativa, dessa forma, preferiu-se manter o primeiro modelo ajustado, uma vez que não houve alteração na significância dos parâmetros. Utilizando as estimativas dos parâmetros, podemos encontrar os valores da *Odds Ratio* do modelo e os respectivos intervalos de confiança dos parâmetros do modelo.

Tabela 2: Razão de Chances, parâmetros e intervalo de confiança para os parâmetros do modelo

Categorias	Parâmetros	OR	2.5 %	97.5 %
(Intercept)	-3.54	0.03	-4.85	-2.23
S2	-0.71	0.49	-0.89	-0.54
R2	0.64	1.90	0.41	0.89
R4	0.54	1.72	0.41	0.67
R5	-1.39	0.25	-2.63	-0.40
ESC2	0.33	1.39	-0.22	0.86
ESC3	0.10	1.10	-0.44	0.61
ESC4	-0.87	0.42	-1.42	-0.35
ESC5	-2.14	0.12	-3.68	-0.92
I	0.23	1.25	0.15	0.30

Após o ajuste do modelo, pode-se usá-lo para estimar a probabilidade de um indivíduo que morreu, ter sido assassinado, a partir de:

$$\hat{\pi} = \frac{exp(-3.54 - 0.72(S2) + 0.64(R2) + \cdots - 0.87(ESC4) - 2.14(ESC5) + 0.23I)}{1 + exp(-3.54 - 0.72(S2) + 0.64(R2) + \cdots - 0.87(ESC4) - 2.14(ESC5) + 0.23I)}, \quad (1)$$

Na tabela 3 apresenta-se a probabilidade de um jovem que morreu, ter sido assassinado, dado algumas características. Note que, dado que está morto, um jovem de 17 anos, solteiro, do sexo masculino, de cor negra e com nenhuma escolaridade possui 72,19% de probabilidade de ter morrido assassinado.

4.3 Qualidade do Ajuste

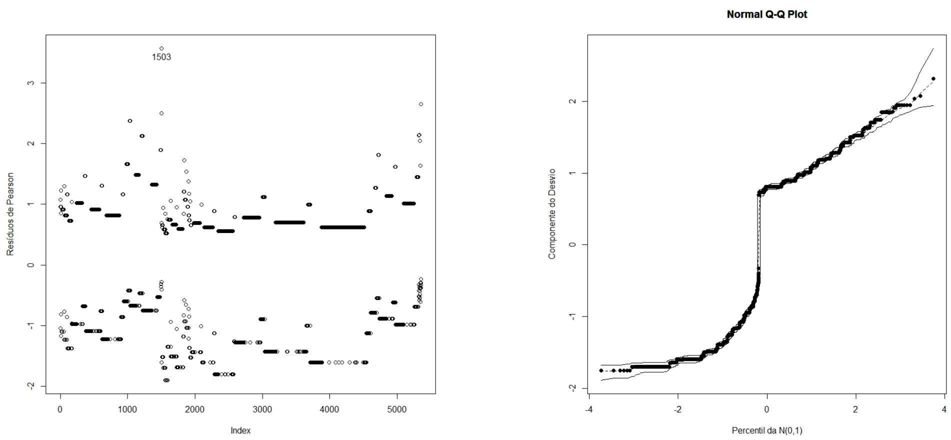
Para verificar a qualidade do modelo ajustado, foi realizada a análise gráfica dos resíduos de Pearson, o gráfico Q-Qplot dos resíduos com envelope simulado e da curva ROC, como mostra as figuras 5 e 6, respectivamente.

Na imagem (a) da figura 5, espera-se que os pontos fiquem dentro do intervalo -2.5 a 2.5, para que se possa concluir que o modelo é satisfatório. Como constatado são poucos

Tabela 3: Probabilidade de algum(a) jovem que morreu, ter sido assassinado(a), segundo algumas combinações das variáveis explicativas do modelo ajustado.

Sexo	Raça/Cor	Escolaridade	Idade	Probabilidade
Masculino	Branca	Nenhuma	17	57,68%
Masculino	Negra	Nenhum	17	72,19%
Masculino	Negra	12 ou mais anos	17	23,38%
Feminina	Branca	8 a 11 anos	16	18,16%
Feminina	Branca	8 a 11 anos	17	21,77%
Feminina	Parda	12 ou mais anos	17	11,86%

os pontos que estão fora deste intervalo, portanto apresenta um bom ajuste pela análise de resíduos de Pearson. Com relação à imagem (b) da figura 5, os pontos devem apresentar-se dentro do envelope, o que se verifica na análise, portanto apresenta-se satisfatório.

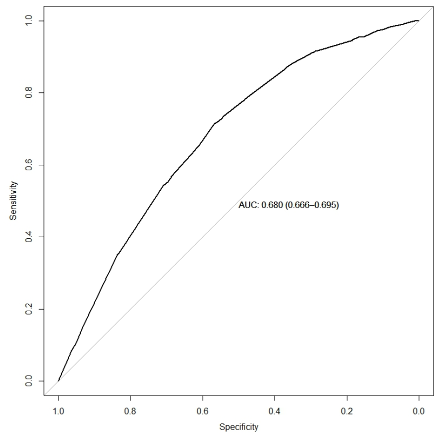


(a) Resíduos de Pearson

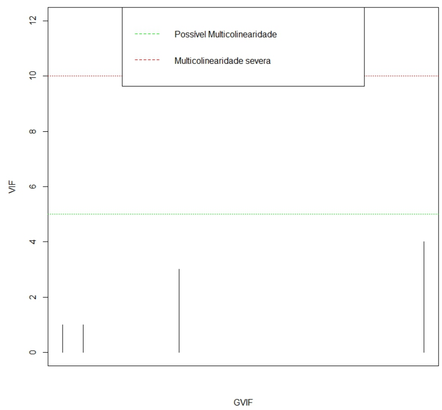
(b) Q-Qplot com envelope simulado Deviance

Figura 5

O valor correspondente à área abaixo da curva ROC foi de 0.6802, que de acordo com os níveis conhecidos indicam que o modelo detém uma capacidade de discriminação aceitável. Com o calculo do fator de inflação da variância (VIF), observamos que nenhum VIF foi maior que 2, portanto, não houve problemas de multicolinearidade.



(a) Curva ROC



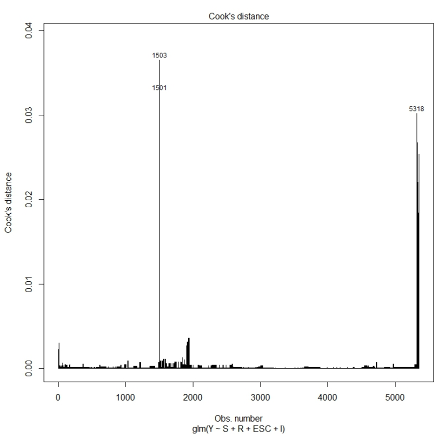
(b) Fator de Inflação da variância

Figura 6

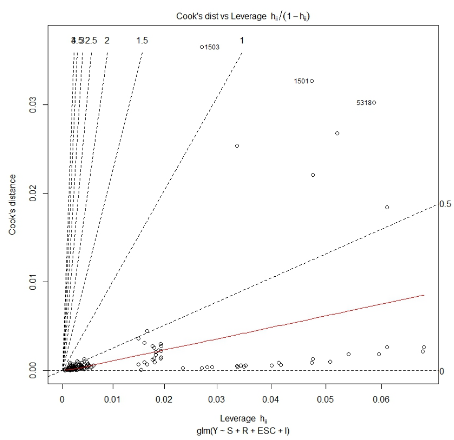
Assim, pelos gráficos apresentados, conclui-se que o modelo ajustado é satisfatório. Foi realizado também o teste de Durbin-Watson para não indepêndencia dos erros, encontramos p-valor menor que 0,05 e portanto rejeita-se H0 e os erros são independentes.

4.4 Pontos Influentes

Foi realizada ainda a análise gráfica da distância de Cook's e do Leverage, as análises estão apresentadas no gráfico 7. Em todos existem poucos pontos que justificam uma análise, porém como não há erros no banco de dados e todas as observações dos gráficos estão dentro de limites aceitáveis, considera-se no trabalho o ajuste com todas as observações influentes, nenhuma observação foi retirada do banco de dados.



(a) Distância de Cook's



(b) Distância de Cook's versus Leverage

Figura 7

Foi realizada também uma análise para a presença de outliers. Alguns valores foram identificados graficamente como outliers, utilizou-se o teste de Bonferroni para alguns, como por exemplo 1501, 1503, 5318 e todos foram considerados outliers. Porém, como os dados são de indivíduos reais e não havia nada que os inviabilizassem quanto a erros de descrição, eles foram mantidos no banco de dados. Abaixo segue o teste de Bonferroni:

- Ho: As observações são outliers.
- H1: Não H0

Tabela 4: Teste de Bonferroni

Observação	Resíduo	P-valor
1501	2.286602	0.022219
1503	2.366189	0.017972
5318	2.539078	0.011114

5 Conclusões

Este trabalho teve como objetivo gerar um modelo logístico capaz de caracterizar o perfil, quanto a estado civil, escolaridade, sexo, idade e raça/cor dos jovens de 15 a 17 anos que morreram assassinados no estado de Minas Gerais em 2013. A partir da Regressão Logística pode-se observar que um jovem solteiro de 17 anos, do sexo masculino, de cor negra e com nenhuma escolaridade do estado de Minas Gerais, dado que esta morto, tem probabilidade de 0,72 de ter morrido assassinado. Pode-se verificar ainda que: [5]

- A chance de um indivíduo negro ser assassinado é 90% maior que a chance de um indivíduo branco;
- A chance de um indivíduo pardo ser assassinado é 72% maior que a chance de um indivíduo branco;
- A chance de um indivíduo indígena ser assassinado é 75% menor que a chance de um indivíduo branco;
- A chance de um indivíduo que tem de 8 a 11 anos de estudo ser assassinado é 58% menor que a chance de um indivíduo com nenhum estudo;
- A chance de um indivíduo que tem 12 ou mais anos de estudo ser assassinado é 88% menor que a chance de um indivíduo com nenhum estudo;
- A chance de um indivíduo do sexo feminino ser assassinado é 52% menor que a chance de um indivíduo do sexo masculino;
- A mudança de uma unidade na idade altera em 25% o logito do modelo.

Referências

[1] Homicídio é principal causa de mortes de jovens de 16 e 17 no país. http://g1.globo.com/politica/noticia/2015/06/homicidio-e-principal-causa-de-mortes-de-jovens-de-16-e-17-no-pais.html. acessado em 11/11/2015.

[2] C. R. Bilder and T. M. Loughin. *Analysis of categorical data with R*. CRC Press, 2014.

[3] R. F. C. da Trindade, F. A. d. M. M. Costa, G. B. Caminiti, C. B. dos Santos, et al. Mapa dos homicídios por arma de fogo: perfil das vítimas e das agressões. *Revista da Escola de Enfermagem da USP*, 49(5):748–755, 2015.

[4] A. J. Dobson and A. Barnett. *An introduction to generalized linear models*. CRC press, 2008.

[5] J. R. Giovanni. *A conquista da matemática: teoria, aplicação, 6a. série*. FTD, 1985.

[6] D. W. Hosmer. Lemeshow. 1989. applied logistic regression. *Ed. John Wolfley y Sons*, pages 8–20, 81.

[7] G. A. Paula. *Modelos de regressão: com apoio computacional*. IME-USP São Paulo, 2004.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[9] J. J. Waiselfisz. Mapa da violência 2015: adolescentes de 16 e 17 anos do brasil. 2015.

Script do R

```
rm(list=ls())
gc(reset=TRUE)

#Os dados encontram-se em pasta pública do Dropbox e podem ser acessados via link
#abaixo:

setwd("https://www.dropbox.com/sh/k4vrb56qmzi88gv/AADePL9ijO46Tdz3mFYmIoF6a?dl=0")

require(gdata)
require(foreign)
require(ggplot2)
require(MASS)
library(betareg)
require(Epi)
library(car)
library(mlbench)
library(effects)
library(np)
library(biglm)
cat("\014")


#update.packages(checkBuilt=TRUE)

dados=read.table("15_17.txt",header=TRUE)
attach(dados)
#S=sexo, R=Raça/cor, ESC=Escolaridade, I=Idade,
#Y=1: Morrer assassinado e 0: morrer de outra forma
dados=transform(R=factor(R),ESC=factor(ESC),S=factor(S),dados)
head(dados)

#Estatística Descritiva dos dados
summary(dados)

#Gráfico que mostra por idade (15 a 17 anos) a proporção do
#número de pessoas assassinadas em relação ao número de
#pessoas que morreram em 2013 no estado de Minas Gerais.
```

```
par(mfrow=c(1,1))

a=table(I,Y)
prop=a[,2]/(a[,1]+a[,2])
Idade <- seq(15, 17, 1)
plot(Idade,prop,ylab="Proporção", main="Proporção por idade
de pessoas assassinadas em 2013")
lines(Idade,prop)

#Note que as mortes por assassinato superam as mortes por
#outras causas

bp=barplot(table(Y,I), beside=T, leg=c("Morte por outra causa",
"Homicidío" ),
           args.legend = list(x = "topleft",bty = "n"),
ylim=c(0, 2000), ylab="Quantidade", xlab="Idade",
main="Número de pessoas que morreram em 2013
           no Estado de Minas Gerais aos 15,16 e 17 anos")
text(bp, table(Y,I)+100, table(Y,I))



#Outros gráficos interessantes para analise

bp=barplot(table(Y,S),ylim=c(0, 3500), beside=T,
leg=c("Morte por outra causa", "Homicidío" ),
           ylab="Quantidade", xlab="",
names.arg = c("masculino","feminino"),
main="Número de pessoas que morreram em 2013
           no Estado de Minas Gerais por sexo")
text(bp, table(Y,S)+150, table(Y,S))

bp=barplot(table(Y,R),args.legend = list(x = "topleft",bty = "n"),
 ylim=c(0, 2500), beside=T,
leg=c("Morte por outra causa", "Homicidío" ),
           ylab="Quantidade", xlab="Raça/Cor",
names.arg = c("Branco","Negro","Pardo","Indígena"),
main="Número de pessoas que morreram em 2013
           no Estado de Minas Gerais por Raça/Cor")
```

```
text(bp, table(Y,R)+150, table(Y,R))

bp=barplot(table(Y,ESC), args.legend = list(x = "topleft",
bty = "n"), ylim=c(0, 2500), beside=T,
leg=c("Morte por outra causa", "Homicidío" ),
           ylab="Quantidade", xlab="Escolaridade",
names.arg = c("Nenhum","1 a 3","4 a 7","8 a 11","12 ou mais"),
main="Número de pessoas que morreram em 2013 no
           Estado de Minas Gerais por anos de estudo")
text(bp, table(Y,ESC)+150, table(Y,ESC))

#A função aggregate() é usada para encontrar o
#número de sucessos e o número de fracassos para cada "cenário":
w <- aggregate(Y ~ S+R+ESC+I, data=dados,FUN=sum)
n <- aggregate(Y ~ S+R+ESC+I,data=dados,FUN=length)
success=w$Y; failure=n$Y
w.n <- data.frame(S=w$S,R=w$R,ESC=w$ESC,I=w$I, success=w$Y,
failure=n$Y , proportion = round(w$Y/n$Y,6))
head(w.n)


#Número de pessoas assassinadas por idade
plot(w.n[,4],w.n[,5],ylab="Quantidade",xlab="Idade",
main="Quantidade de pessoas assassinadas por Idade
     em Minas Gerais no ano de 2013
     a partir dos dez anos de idade")

#Box-plot da Proporção de pessoas assassinadas por Raça/Cor
plot(w.n[,2],w.n[,7],ylab="Proporção de assassinato",xlab="Raça/Cor",
names=c("Branco","Negro","Pardo","Indígena"),
main="Box-plot das pessoas assassinadas em Minas Gerais
     no ano de 2013 por Raça/Cor")

#Box-plot da Proporção de pessoas assassinadas por Escolaridade
plot(w.n[,3],w.n[,7],ylab="Proporção",xlab="Escolaridade",
names=c("Nenhum","1 a 3","4 a 7","8 a 11","12 ou mais"),
main="Box-plot das  pessoas assassinadas em Minas Gerais
     no ano de 2013 por Escolaridade")
```

```
#Box-plot da Proporção de pessoas assassinadas por Sexo
bx=plot(w.n[,1],w.n[,7],ylab="Proporção",xlab="",
names= c("masculino","feminino"), main="Box-plot das
pessoas assassinadas em Minas Gerais
        no ano de 2013 por Sexo")


#Análise visual do impacto das covariáveis em Y

Y1=factor(Y)
layout(matrix(1:2, ncol = 2))
cdplot(Y1 ~ I, data = dados)
cdplot(Y1 ~ R, data = dados)
cdplot(Y1 ~ ESC, data = dados)
cdplot(Y1 ~ S, data = dados)

####################################
#Possíveis modelos
###################################

mod1<- glm(formula = Y ~ 1, family=binomial(link=logit), data = dados)
mod2<- glm(formula = Y ~ S, family=binomial(link=logit), data = dados)
mod3<- glm(formula = Y ~ R, family=binomial(link=logit), data = dados)
mod4<- glm(formula = Y ~ ESC, family=binomial(link=logit), data = dados)
mod5<- glm(formula = Y ~ I, family=binomial(link=logit), data = dados)
mod6<- glm(formula = Y ~ S+R, family=binomial(link=logit), data = dados)
mod7<- glm(formula = Y ~ S+ESC, family=binomial(link=logit), data = dados)
mod8<- glm(formula = Y ~ S+I, family=binomial(link=logit), data = dados)
mod9<- glm(formula = Y ~ R+ESC, family=binomial(link=logit), data = dados)
mod10<- glm(formula = Y ~ R+I, family=binomial(link=logit), data = dados)
mod11<- glm(formula = Y ~ ESC+I, family=binomial(link=logit), data = dados)
mod12<- glm(formula = Y ~ S+R+ESC, family=binomial(link=logit), data = dados)
mod13<- glm(formula = Y ~ S+R+I, family=binomial(link=logit), data = dados)
mod14<- glm(formula = Y ~ R+ESC+I, family=binomial(link=logit), data = dados)
mod15<- glm(formula = Y ~ S+R+ESC+I, family=binomial(link=logit), data = dados)
mod16<- glm(formula = Y ~ S+R+ESC+I+I*S, family=binomial(link=logit), data = dados)
mod17<- glm(formula = Y ~ S+R+ESC+I+I*R, family=binomial(link=logit), data = dados)
mod18<- glm(formula = Y ~ S+R+ESC+I+I*ESC, family=binomial(link=logit), data = dados)
mod19<- glm(formula = Y ~ S+R+ESC+I+I*S+I*R, family=binomial(link=logit),
```

```
 data = dados)
mod20<- glm(formula = Y ~ S+R+ESC+I+I*S+I*ESC, family=binomial(link=logit),
 data = dados)
mod21<- glm(formula = Y ~ S+R+ESC+I+I*R+I*ESC, family=binomial(link=logit),
 data = dados)
mod22<- glm(formula = Y ~ S+R+ESC+I+I*S*R, family=binomial(link=logit),
 data = dados)
mod23<- glm(formula = Y ~ S+R+ESC+I+I*S*ESC, family=binomial(link=logit),
 data = dados)
mod24<- glm(formula = Y ~ S+R+ESC+I+S*R, family=binomial(link=logit),
 data = dados)
mod25<- glm(formula = Y ~ S+R+ESC+I+S*R*ESC, family=binomial(link=logit),
 data = dados)

#Teste para comparação dos modelos
anova(mod1,mod2,mod3,mod4,mod5,mod6,mod7,mod8,mod9,mod10,mod11,mod12,mod13,mod14,
mod15,mod16,mod17,mod18,mod19,mod20,mod21,mod22,mod23,mod24,mod25,test="Chisq")

#Por este teste foi possível avaliar os diversos modelos acima,
#foram significativos os modelos mod2 ao mod15

anova(mod1,mod2,mod3,mod4,mod5,mod6,mod7,mod8,mod9,mod10,mod11,mod12,mod13,
mod14,mod15,test="Chisq")
#O melhor modelo que possui deviance mais próxima do número de graus de liberdade é:
#Model 15: Y ~ S + R + ESC + I

#Vejamos métodos de seleção de covariáveis

#Pacote e códigos para escolha do melhor modelo

library (glmulti)
search.1.aicc<-glmulti(y=Y ~.,data=dados, fitfunction="glm",level=1,method="h",
crit="aicc", family=binomial(link="logit"))
print(search.1.aicc)
aa<-weightable(search.1.aicc)
cbind(model=aa[1:5,1],round(aa[1:5,2:3],digits=3))

#Por meio do código acima, o melhor modelo foi "Y ~ 1 + S + R + ESC + I"
```

```
#Por meio do método de seleção abaixo podemos tentar achar um melhor modelo
#Foward selection
empty.mod<-glm(formula = Y ~ 1, family=binomial(link=logit), data = dados)
full.mod<-glm(formula = Y ~ ., family=binomial(link=logit), data = dados)
forw.sel<-step(object=empty.mod, scope=list(upper=full.mod), direction="forward",
 k=log(nrow(dados)), trace=TRUE)

#Melhor modelo pelo foward: Y ~ ESC + S + R + I

#Backward selection
back.sel<-step(object=full.mod,scope=list(lower=empty.mod), direction="backward",
k=(nrow(dados)),trace = TRUE )

#Melhor modelo pelo backward: Y ~ S

##Stepwise selection
step(glm(formula = Y ~ ., family=binomial(link=logit), data = dados),
direction = "both")

#Melhor modelo Stepwise: Y ~ S + R + ESC + I

#Outro método de seleção individual

search.1.bic<-glmulti(y=Y ~., data=dados, fitfunction="glm", level=1, method="h",
 crit="bic", family=binomial(link="logit"))
head(weightable(search.1.bic))
plot(search.1.bic,type="w")
#Coeficientes do melhor modelo pela função glmulti
parms<- coef(search.1.bic)
# Renaming columns to fit in book output display
colnames(parms)<-c("Estimate","Variance","n.Models","Probability","95%CI+/-")
round(parms,digits=3)
parms.ord<-parms[order(parms[,4],decreasing=TRUE),]
ci.parms<-cbind(lower=parms.ord[,1]-parms.ord[,5],upper=parms.ord[,1]+
parms.ord[,5])
round(cbind(parms.ord[,1],ci.parms),digits = 3)

#Melhor modelo:  Y ~ 1 + S + R + ESC + I
```

```
#Razão de chances
round(exp(cbind(OR=parms.ord[,1], ci.parms))[-1,],digits=2)


bestfit=glm(formula = Y ~ S+R+ESC+I, family=binomial(link=logit), data = dados)
anova(bestfit,test="Chisq")
summary(bestfit)

library(ResourceSelection)
#Teste de Hosmer e Lemeshow
#H0: O modelo está bem ajustado
#H1: O modelo não está bem ajustado
hoslem.test(bestfit$y, fitted(bestfit))

#Resíduo de Pearson
rp=residuals(bestfit,type="pearson")

#H0:o modelo ajustado está correto
#Ha:Não H0
1-pchisq(sum(rp^2),bestfit$df.residu)
#O modelo está correto

library(pROC)
roc(bestfit$y,bestfit$fitted.values,plot=T,ci=T,identity=TRUE,print.auc=TRUE)

#Gráfico para o VIF
v=vif(bestfit)
plot(v,type="h",ylim=c(0,ifelse(max(v)<10,12,max(v)*1.1)),xaxt="n",ylab="VIF")
abline(h=c(5,10),col=c("green","red"),lty=3)
legend("top",legend=c("Possível Multicolinearidade","Multicolinearidade severa"),
col=c("green","red"),lty=2)
axis(1,at=1:length(names(v)),names(v))


#Não independência dos erros
# Teste para indepedência dos erros
a=durbinWatsonTest(bestfit)
```

```
plot(residuals(bestfit,type="pearson"),ylab="Resíduos de Pearson")
identify(1:length(Y), residuals(bestfit,type="pearson"))

outlierTest(bestfit) # Bonferonni p-value for most extreme obs

#Distância de Cook's
plot(bestfit, which=4)
#Distância de Cook's vs Leverage
plot(bestfit, which=6)

fit.model=bestfit
source("envel_bino.txt")


cbind(Par=coef(bestfit),OR=exp(coef(bestfit)), confint(bestfit))

S2=1
R2=0
R4=1
R5=0
ESC2=0
ESC3=0
ESC4=0
ESC5=1
I=17

hat_pi = exp(-3.54342-0.71483*(S2)+0.64436*(R2)+0.54040*(R4)-1.38578*(R5)+
0.33147*(ESC2) +0.09528*(ESC3)-0.87388*(ESC4)-2.14116*(ESC5)+0.22666*I)/
(1+exp(-3.54342-0.71483*(S2)+ 0.64436*(R2)+0.54040*(R4)-1.38578*(R5)+
0.33147*(ESC2)+0.09528*(ESC3)-0.87388*(ESC4)-2.14116*(ESC5)+0.22666*I))
hat_pi
```