Perfil dos jovens de 15 a 17 anos assassinados em Minas Gerais no ano de 2013
Estudo via Regressão Logística

Fernando de Souza Bastos

Doutorando em Estatística, Universidade Federal de Minas Gerais
fsbmat@gmail.com

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, [5].

Metodologia

A fonte de informação primária da base de dados são os atestados de óbito emitidos pelos cartórios civis. Esta 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. Os valores não preenchidos foram retirados do estudo. A causa de mortalidade no banco de dados está codificada segundo a Classificação Internacional de Doenças através da CID10.

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 indivíduos para análise.

As descrições das codificações estão na tabela abaixo:

Variáveis	Categorias	Descrição
Y	0	Não homicídio
	1	homicídio
S	1	Masculino
	2	Feminino
R	1	R/C Branca
	2	R/C Negra
	4	R/C Parda
	5	R/C Indígena
ESC	1	Nenhum
	2	1 a 3 anos
	3	4 a 7 anos
	4	8 a 11 anos
	5	$\geq$ 12 anos
I	Idade (Cont.)	15 a 17 anos

Tabela 1: Variáveis consideradas no estudo com suas respectivas categorias

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

Como a variável resposta é binária (0 ou 1), considerou-se a distribuição binomial como componente aleatório. O componente sistemático é dado pela combinação linear das variáveis explicativas e para função de ligação considerou-se as funções: logit, probit, complemento log-log e cauchit. Após a descrição e exploração dos dados foi feita a seleção das variáveis que melhor explicam a variável dependente. A seleção foi realizada pelo algoritmo *stepwise* considerando como critério de seleção o AIC (Critério de Informação de Akaike). Considerou-se como modelo completo o modelo aditivo com todos os efeitos principais e todas as interações.

Os modelos especificados com as diferentes funções de ligação apresentaram um comportamento muito parecido. Todos os gráficos de resíduos simulados apresentaram resíduos dentro dos intervalos simulados. Em relação ao *AIC*, *Deviance*, *pseudo-*R^2 e curva *ROC*, todos tiveram desempenho semelhante. Em relação ao teste de *Hosmer-Lemeshow* e de *Pearson* somente o modelo *Clog-log* não foi adequado. Logo, pela magnitude das medidas comparativas e pela vantagem interpretativa da especificação *logit*, esta foi a função de ligação considerada para o ajuste e análise. Para maiores informações sobre regressão logística binária sugere-se [1], [2], [3] e [4].

Modelo Ajustado

$$Y_i \sim Binomial(n, \widehat{\pi}_i)$$

$$g(\widehat{\pi}_i) = \widehat{\beta_0} + \widehat{\beta_1} S_i + \widehat{\beta_2} R_i + \widehat{\beta_3} ESC_i + \widehat{\beta_4} I_i$$

Resumo da saída do R:

	Valor Est.	Erro Padrão	P-valor
Intercepto	-3.54342	0.66772	1.12e-07*
S2	-0.71483	0.08940	1.28e-15*
R2	0.64436	0.12215	1.33e-07*
R4	0.54040	0.06591	2.42e-16*
R5	-1.38578	0.55451	0.01245*
ESC2	0.33147	0.27559	0.22906
ESC3	0.09528	0.26815	0.72235
ESC4	-0.87388	0.27176	0.00130*
ESC5	-2.14116	0.68340	0.00173*
I	0.22666	0.03733	1.27e-09*

Tabela 2: Resultado da função glm do R

As categorias ESC2 e ESC3 foram não significativas, assim realizou-se o agrupamento destas e avaliou-se um novo ajuste com a covariável escolaridade com quatro categorias, porém a categoria nova permaneceu não significativa, logo, manteve-se o primeiro modelo. Após o ajuste, encontrou-se:

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

Tabela 3: OR e IC dos parâmetros do modelo

Resultados

Na tabela abaixo apresenta-se a probabilidade de um jovem que morreu, ter sido assassinado, dado algumas características.

Sexo	R/C	Esc.	Idade	Prob.
Masc.	Branca	-	17	0,58
Masc.	Negra	-	17	0,72
Masc.	Negra	$\geq$12	17	0,23
Fem.	Branca	8 a 11	16	0,18
Fem.	Branca	8 a 11	17	0,22
Fem.	Parda	$\geq$12	17	0,12

Tabela 4: Probabilidade de um jovem que morreu, ter sido assassinado, dado algumas características.

Algumas interpretações do Odds Ratio:

- A chance de um indivíduo negro ser assassinado é 90% 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 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.

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. Para visualizar os gráficos e a análise completa deste trabalho, o script do R e o banco de dados para a reprodução dos resultados pode-se acessar a página `fsbmat.github.io`.

Agradecimento

O autor agradece a Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG) pelo apoio no desenvolvimento deste trabalho.

Referências

[1] Christopher R Bilder and Thomas M Loughin. *Analysis of categorical data with R*. CRC Press, 2014.

[2] Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2008.

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

[4] Gilberto Alvarenga Paula. *Modelos de regressão: com apoio computacional*. IME-USP São Paulo, 2004.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.