

**UNIVERSIDADE FEDERAL DE OURO PRETO**

**AJUSTE DE UM MODELO LINEAR GENERALIZADO  
ELEIÇÕES 2020 - BELO HORIZONTE-MG**

Relatório de conclusão da disciplina de  
Modelos Lineares Generalizados, ministrada  
pela Prof.<sup>a</sup>. Dr(a). Carolina Silva Pena.

Grupo:

Diana Diniz 17.2.4348

Paulo Vitor de Souza e Silva 18.2.4090

Raphaella Vitória Fernandes da Silva 17.2.4227

**OURO PRETO - MG**

**2021**

# Sumário

1 Introdução .....	3
2 Resultados .....	3
2.1 Análise exploratória .....	4
2.2 Modelo proposto.....	6
2.2.1 Análise do ajuste do modelo .....	8
Referências .....	9

# 1 Introdução

Os modelos lineares generalizados são uma classe de modelos que aumentam as possibilidades de análises para outras distribuições além da distribuição normal. Os MLGs são utilizados quando o objetivo do estudo é fazer uma análise sobre uma variável resposta do tipo contagem, categórica ou Assimétrica (com valores positivos) e para ser um MLG é necessário que a distribuição de probabilidade seja da família exponencial.

Iremos trabalhar com uma base de dados composta por 1103 observações referentes ao resultado das eleições na cidade de Belo Horizonte - MG. Para este trabalho foi necessário juntar duas bases obtidas nos sites do TRE e TSE e em seguida foi realizada uma filtragem neste banco com objetivo de selecionar apenas os candidatos a vereador que concorreram nas eleições 2020 no município de Belo Horizonte.

Os dados podem ser obtidos em: <https://www.tre-mg.jus.br/eleicoes/eleicoes-2020/eleicoes-2020> (clicar em 'Resultados via TRE') e <https://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais> (candidatos > 2020 > Candidatos (formato ZIP) )

Sobre as variáveis presentes no banco de dados:

- Candidato: Nome do candidato que aparece na urna.
- Idade: Idade do candidato na data da posse. A idade é calculada com base na data da posse do referido candidato para o cargo e unidade eleitoral constantes no arquivo de vagas.
- Gênero: Gênero do candidato.
- Escolaridade: Grau de instrução do candidato.
- EstadoCivil: Estado civil do candidato.
- Cor\_raça: Cor/raça do candidato. (auto declaração)
- possuiBens: Indica se o candidato tem ou não bens a declarar. Pode assumir os valores: S - Sim e N - Não. Esta informação é fornecida pelo próprio candidato no momento do pedido da candidatura.
- Votos: Total de votos obtidos pelo candidato.

## 2 Resultados

Para dados de contagem, o modelo se baseia na distribuição de Poisson onde a média é igual a variância. As ligações mais comuns são: logarítmica, raiz quadrada e identidade. Nesta seção apresentaremos os resultados obtidos a partir do ajuste do modelo. Inicialmente, faremos uma análise exploratória da base de dados e posteriormente veremos o ajuste do modelo, o qual foi realizado no software R.

## 2.1 Análise exploratória

Para uma melhor visualização gráfica foi aplicada a raiz cúbica na variável resposta ‘total de votos’. Com intuito de simplificar a interpretação, durante a análise exploratória iremos tratar essa variável transformada apenas como ‘votos’, mas no ajuste do modelo será usada o valor real (total de votos).

Podemos observar na Figura 1 a presença de dois gráficos, sendo eles: (a) Gráfico de dispersão dos votos *versus* idade, onde é possível observar uma grande concentração entre os candidatos na faixa etária de 25 – 65 anos e observamos também que não há correlação significativa entre os votos e idade ( $\text{cor} = -0,0176$ ). No gráfico (b) temos um Boxplot dos votos de acordo com o gênero, onde é possível observar que os candidatos do sexo masculino têm maior quantidade de votos e maior média.

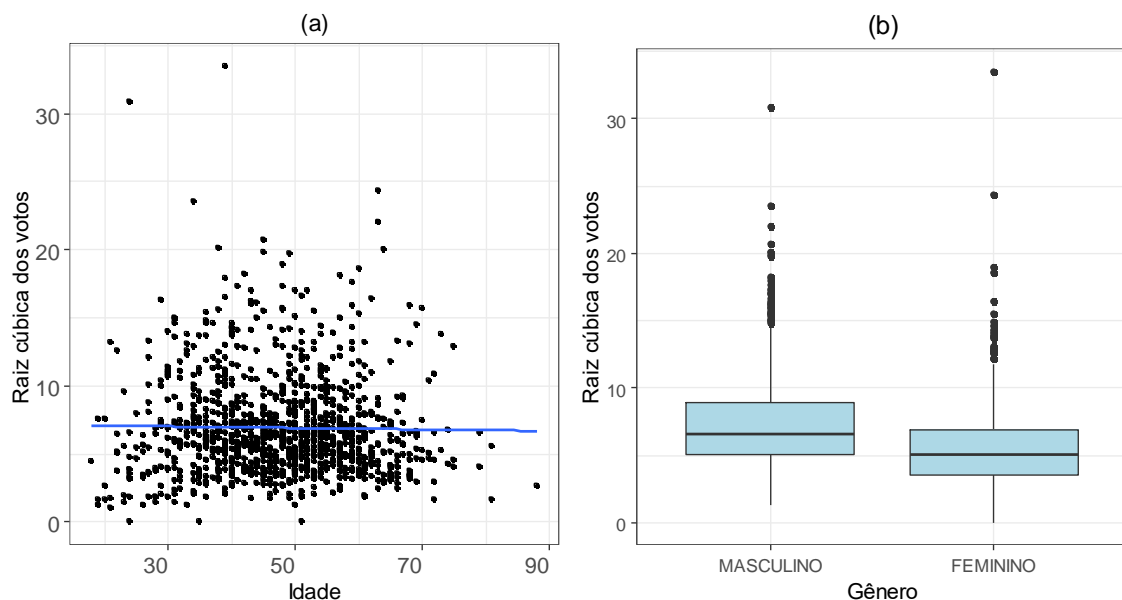


Figura 1: Análise exploratória das variáveis explicativas.

Podemos observar na Figura 2 a presença de dois gráficos. O gráfico (a) é um *Boxplot* dos votos *versus* escolaridade, onde podemos notar que a maior média é dos candidatos que possuem nível superior e superior completo. No gráfico (b) temos um *Boxplot* dos votos de acordo com Estado civil, onde podemos observar que a maior média está no grupo dos casados(as).

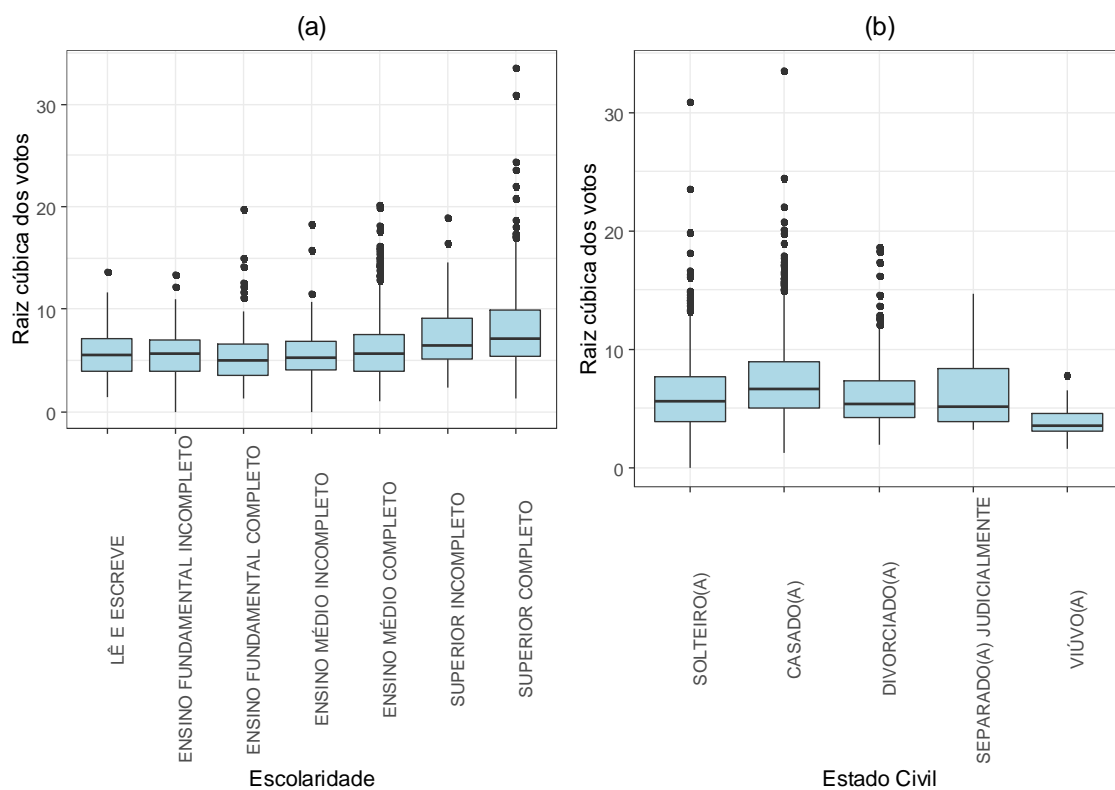


Figura 2: Análise exploratória das variáveis escolaridade e estado civil.

A Figura 3 é composta por dois gráficos, sendo eles: O gráfico (a) apresenta um Boxplot dos votos em relação a Cor/raça. Nota-se que as médias estão bem próximas, porém a cor/raça “amarela” possui maior média. Já no gráfico (b), onde também é um *Boxplot*, a maior média dos candidatos que possuem bens em seu nome.

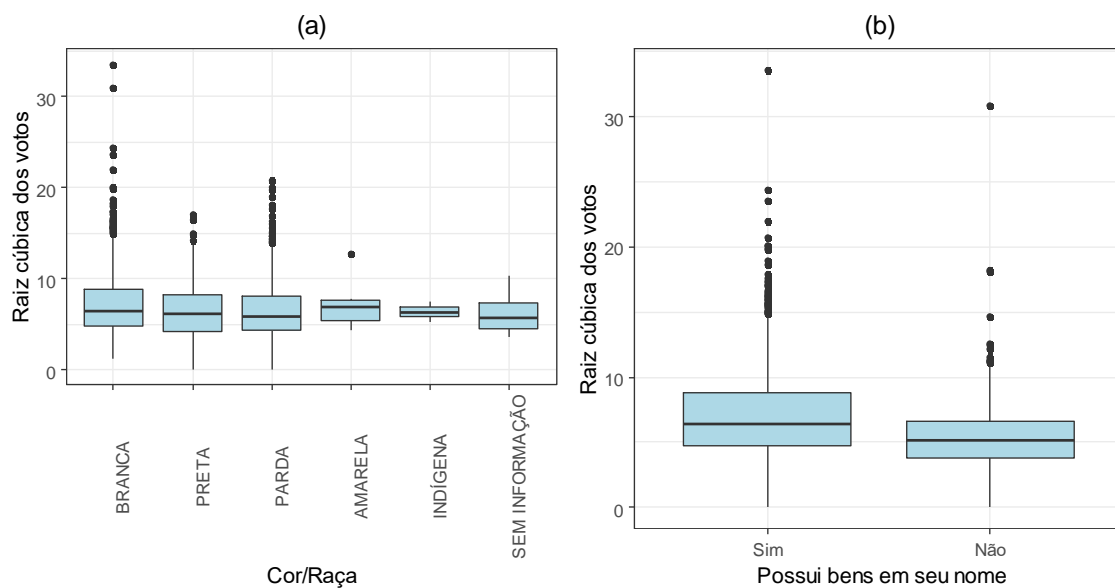


Figura 3: Análise exploratória das variáveis explicativas.

## 2.2 Modelo Proposto

Inicialmente ajustamos um modelo de Poisson pois a variável resposta é discreta e positiva. Verificamos que apesar das variáveis apresentarem coeficientes significativos, os dados exibem mais variação do que se espera em uma distribuição de Poisson. A Super dispersão ocorre quando a variância da variável resposta é maior do que a média. Nesses casos, a suposição de distribuição de Poisson para a resposta é inadequada sendo necessário o uso de modelos alternativos. Algumas das alternativas são o modelo de binomial negativa e o modelo de quase-verossimilhança que leva às mesmas estimativas do modelo de Poisson, porém corrige a variabilidade das estimativas.

Na Tabela 1 temos os valores do aic e deviance que nos ajudaram a escolher o segundo modelo como o mais adequado. Então daremos atenção especial aos modelos com resposta binomial negativa, os quais permitem uma análise mais completa dos dados do que os modelos de quase-verossimilhança propostos por Wedderburn (1974).

*Tabela 1: Escolha do modelo adequado.*

<b>Ajuste</b>	<b>AIC</b>	<b>Deviance</b>
Modelo Poisson	1452293	1444249,188
Modelo Binomial Negativa	16091,24	1339,619
Modelo Quase-Verossimilhança		1453526,943

Ao ajustarmos o modelo com resposta binomial negativa, verificamos que a função de ligação canônica ‘log’ apresentou melhores resultados e por isso ele pode ser considerado um modelo de efeito multiplicativo. Todas as variáveis foram significativas pois pelo menos uma categoria apresentou p-valor menor que 0.05. Os coeficientes obtidos são mostrados na Tabela 2.

*Tabela 2: Modelo Binomial Negativa- Coeficientes estimados*

	<b>Estimativa</b>	<b>Erro Padrão</b>	<b>Valor Z</b>	<b>Pr(&gt; z )</b>
(Intercept)	6,5435	0,2464	26,5550	< 2e-16
idade	-0,0084	0,0036	-2,3740	0,0176
generoFEMININO	-0,5937	0,0818	-7,2540	0,0000
escolaridadeENSINO FUNDAMENTAL INCOMPLETO	0,0649	0,2390	0,2720	0,7860
escolaridadeENSINO FUNDAMENTAL COMPLETO	0,1141	0,2098	0,5440	0,5864
escolaridadeENSINO MÉDIO INCOMPLETO	0,1255	0,2332	0,5380	0,5906
escolaridadeENSINO MÉDIO COMPLETO	0,3414	0,1749	1,9520	0,0509
escolaridadeSUPERIOR INCOMPLETO	0,7144	0,2137	3,3440	0,0008
escolaridadeSUPERIOR COMPLETO	1,0746	0,1746	6,1550	0,0000

estadoCivilCASADO(A)	0,3537	0,0862	4,1020	0,0000
estadoCivilDIVORCIADO(A)	-0,0502	0,1303	-0,3860	0,6997
estadoCivilSEPARADO(A) JUDICIALMENTE	-0,1256	0,3167	-0,3970	0,6916
estadoCivilVIÚVO(A)	-1,2838	0,3181	-4,0350	0,0001
cor_raçaPRETA	-0,5789	0,1047	-5,5290	0,0000
cor_raçaPARDA	-0,2937	0,0861	-3,4100	0,0007
cor_raçaAMARELA	-0,3874	0,5083	-0,7620	0,4460
cor_raçaINDÍGENA	-0,0339	0,8763	-0,0390	0,9692
cor_raçaSEM INFORMAÇÃO	-0,7736	0,3381	-2,2880	0,0221
possuiBensNão	-0,4580	0,0953	-4,8050	0,0000

A partir dos resultados mostrados Tabela 2, podemos verificar que:

Para cada aumento de uma unidade na variável explicativa idade, o valor esperado da variável resposta será multiplicado por  $e^{\beta_1} = e^{-0,0084} = 0,9916$ . Ou ainda, há um decréscimo de 0,84% no número esperado de votos para cada aumento de um ano de vida do candidato.

Considerando que o gênero masculino é a categoria de referência, o número esperado de votos para os candidatos do gênero feminino será multiplicado por  $e^{\beta_2} = e^{-0,5937} = 0,5523$ . Ou ainda, há um decréscimo de 44,77% no número esperado de votos para aqueles que são do sexo feminino.

O número esperado de votos será multiplicado por  $e^{\beta_{35}} = e^{0,7144} = 2,043$  se compararmos a categoria de referência ‘lê e escreve’ com aqueles que possuem ensino superior incompleto e será multiplicado por  $e^{\beta_{36}} = e^{1,0746} = 2,9288$  se compararmos a categoria de referência com aqueles que possuem ensino superior completo. As demais categorias referentes a escolaridade não foram significativas (p-valor >0,05).

Considerando o estado civil ‘solteiro’ como categoria de referência: O número esperado de votos para os ‘casados’ será multiplicado por  $e^{\beta_{41}} = e^{0,3537} = 1,4243$ , ou seja, haverá um aumento de 42,43% no número esperado de votos para aqueles pertencentes a categoria casados. O número esperado de votos para os ‘viúvos’ será multiplicado por  $e^{\beta_{44}} = e^{-1,2838} = 0,27$ , ou seja, haverá um decréscimo de 72,3% no número esperado de votos para aqueles pertencentes a categoria viúvos. As demais categorias referentes ao estado civil não foram significativas (p-valor >0,05).

Considerando a raça ‘Branca’ como categoria de referência: O número esperado de votos para os que se declaram ‘pretos’ será multiplicado por  $e^{\beta_{51}} = e^{-0,5789} = 0,5605$ , ou seja, haverá uma redução de 43,95% no número esperado de votos para aqueles pertencentes a categoria preto. Para aqueles que se declararam ‘pardos’ espera-se uma redução de 25,45% no

número esperado de votos e para aqueles que não declararam raça, espera-se uma redução de 53,87%. As categorias ‘amarelo’ e ‘indígena’ não foram significativas.

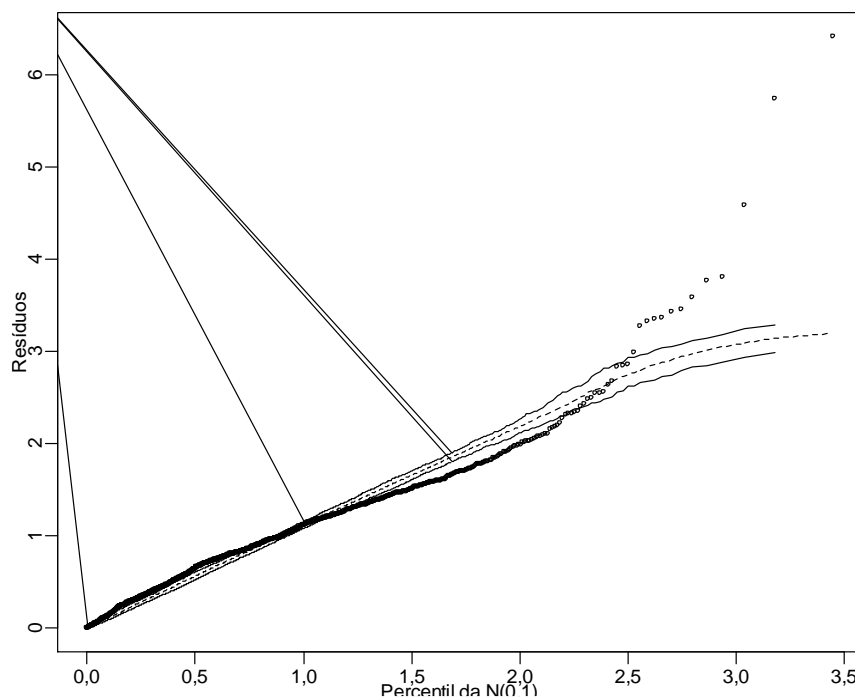
O número esperado de votos para os candidatos que responderam ‘não’ quando perguntados se possuíam bens em seu nome será multiplicado por  $e^{\beta_6} = e^{-0,4580} = 0,6325$ . Ou seja, espera-se uma redução de 36,75% no número esperado de votos em relação àqueles que responderam ‘sim’.

### 2.2.1 Análise do ajuste do modelo

*Tabela 3: Análise do ajuste do modelo - Deviance.*

Deviance	Graus de Liberdade	P-valor
1339,619	1085	0

De acordo com o p-valor observado na Tabela 3, rejeitamos a hipótese de que o modelo está bem ajustado. Entretanto, a partir da aproximação assintótica, ainda que a amostra em estudo seja consideravelmente maior que o número de possíveis combinações entre as covariáveis, não é razoável dizer ainda que o modelo esteja mal ajustado, sendo necessário também a análise dos resíduos para verificar a adequabilidade do modelo.



*Figura 4: Gráfico de envelope dos resíduos.*



Pelo gráfico de envelope, Figura 4, construído a partir da distribuição empírica dos resíduos, obtida sob 100 simulações Monte Carlo, observa-se que os resíduos do modelo não estão completamente contidos dentro do intervalo simulado.

Mesmo utilizando os métodos de super dispersão não foi possível obter um modelo adequado utilizando as técnicas de modelo linear generalizado, porém existem outras possibilidades como os modelos não lineares e os modelos não paramétricos, que não são foco dessa disciplina.

## **Referências**

PAULA, Gilberto A. Modelos de regressão com apoio computacional. São Paulo: IME-USP, 2004. 245p. Disponível em: [https://www.ime.unicamp.br/~cnaber/Livro\\_MLG.pdf](https://www.ime.unicamp.br/~cnaber/Livro_MLG.pdf) . Acesso em 05/12/2021.

ROSSI, Alfredo; PORTELA, Cayan. Modelos Lineares Generalizados. LAMFO, 2018. Disponível em: <https://lamfo-unb.github.io/2018/09/29/MLG/>. Acesso em: 05/12/2021.

Wedderburn, R.W.M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method". *Biometrika* 61 (3): 439–447.