Modelo de Regressão Logística Ordinal com chances proporcionais

Wesley Furriel

Universidade Estadual de Maringá, Departamento de Estatística, PR, Brasil

12 de Junho de 2018

Organização

Introdução

O presente trabalho teve por objetivo investigar se os eleitores tendem a concordam ou discordam da questão **"Se um governante administra bem o país, não importa se ele é corrupto ou não"**, presente no *survey* do Datafolha em 2017. Como o interesse foi investigar essa questão sob a ótica dos modelos regressão ordinal, três variáveis sobre a opinião dos eleitores acerca do cenário político e escolaridade dos mesmos foram empregadas, com o intuito de avaliar o impacto destas, sobre a frase exposta.

Pacotes utilizados e descrição do banco

- **Rstudio 1.1.447**
  - *VGAM, dplyr, generalhoslem, ordinal, ggplot2*.
- **SAS 9.4**
  - *proc logistic*
- **Bancos de dados**
  - DataFolha
    - ⋆ Ano:2017
    - ⋆ Número de observações: 2774
    - ⋆ Número de variáveis: 124
    - ⋆ Fonte: http://datafolha.folha.uol.com.br/

Modelo de Regressão Logística Ordinal

Nos casos em que a variável de interesse apresenta mais de duas categorias, isto é r categorias $(r > 2)$ e possui ordenação, foram propostos modelos que utilizam os denominados logitos cumulativos (GIOLO, 2017). Para apresentá-lo considere $j = 1, 2, ..., r$ os índices das categorias ordenadas de Y, que segue uma distribuição multinomial,

$$\begin{aligned}
&\theta_1(\boldsymbol{x}) = p_1(x) = P(Y \leq 1|\boldsymbol{x}),\\
&\theta_2(\boldsymbol{x}) = p_1(\boldsymbol{x}) + p_2(x) = P(Y \leq 2|\boldsymbol{x}),\\
&...,\\
&\theta_r(\boldsymbol{x}) = p_1(\boldsymbol{x}) + p_2(\boldsymbol{x}) + p_{r-1}(\boldsymbol{x}) + p_r(x) = P(Y \leq r|\boldsymbol{x}),
\end{aligned}$$

em que $p_j(x) = P(Y = j|\boldsymbol{x})$ referente à probabilidade de ocorrência da j-ésima categoria de resposta para dado vetor $\boldsymbol{x}$ de p covariáveis, de forma que $\sum_{j=1}^{r} p_j(\boldsymbol{x}) = \theta_r(\boldsymbol{x}) = 1$.

Considerando as quantidades $\theta_j(\boldsymbol{x})$ correspondente as probabilidades acumuladas, tal que $\theta_1(\boldsymbol{x}) \leq \theta_2(\boldsymbol{x}) \leq ...\theta_r(\boldsymbol{x}) = 1$ definem os logitos cumulativos

$$ln\left[\frac{\theta_j(\boldsymbol{x})}{1-\theta_j(\boldsymbol{x})}\right] = ln\left[\frac{P(Y \leq j|x)}{1-P(Y \leq j|x)}\right] = ln\left[\frac{P(Y \leq j|x)}{P(Y > j|x)}\right]$$

para $j = 1, ..., r-1$ e $\boldsymbol{x}$ um de valores de p covariáveis. Partindo destas formulações podemos considerar os modelos, com chances proporcionais e não proporcionais e com chances proporcionais parciais.

Considerando a não existência de ordenação em uma variável resposta Y, toma-se o modelo logístico com categoria de referência, no qual $p_j(\boldsymbol{x})$, é a probabilidade de ocorrência da categoria $j(j = 1, 2, ..., r)$ para um dados vetor $\boldsymbol{x}$ de p covariáveis,

$$ln\left[\frac{p_1(\boldsymbol{x})}{p_r(\boldsymbol{x})}\right], ln\left[\frac{p_2(\boldsymbol{x})}{p_r(\boldsymbol{x})}\right], ..., ln\left[\frac{p_{r-1}(\boldsymbol{x})}{p_r(\boldsymbol{x})}\right]$$

em que cada nível da resposta é comparado com o nível de referência fixado. Sendo a resposta categórica ordinal a função de ligação mais popular é o de logitos cumulativos. Seja, $\theta_j(x) = P(Y \leq j)$, temos:

$$ln\left[\frac{p_1(x)}{p_2(x) + ... + p_r(x)}\right], ln\left[\frac{p_1(x) + p_2(x)}{p_3(x) + ... + p_r(x)}\right], ..., ln\left[\frac{p_1(x) + ... + p_{r-1}(x)}{p_r(x)}\right]$$

neste caso à medida que a função logit muda ordinalmente de categoria, o numerador aumenta e o denominador diminui (DERR, 2013).

Modelos com chances proporcionais

Em alguns casos a suposição de chances proporcionais equivale a supor que $\beta_j = \beta$ para todo j. Este modelo popularizado por McCullagh (1980), é dado por

$$ln\left[\frac{\theta_j(\boldsymbol{x})}{1-\theta_j(\boldsymbol{x})}\right] = ln\left[\frac{P(Y \leq j|x)}{P(Y > j|x)}\right] = \boldsymbol{\beta}_{0j} + \boldsymbol{\beta}'\boldsymbol{x}$$

para $j = 1, ..., r-1$ com $\beta = (\beta_1, ..., \beta_p)$ o modelo exposto assume que o efeito das covariáveis não diferem entre os $r-1$ logitos.

Em termos de probabilidade cumulativas temos

$$\theta_j(\boldsymbol{x}) = \frac{exp(\boldsymbol{\beta}_{0j} + \boldsymbol{\beta}'\boldsymbol{x})}{1 + exp(\boldsymbol{\beta}_{0j} + \boldsymbol{\beta}'\boldsymbol{x})}, \quad j = 1, 2, ..., r-1$$

As probabilidades $p_j(x), j = 1, ..., r$, são obtidas para o modelo por meio das subtrações $p_j(x) = \theta_j(\boldsymbol{x}) - \theta_{j-1}(\boldsymbol{x})$ em que $\theta_0(\boldsymbol{x}) = 0$ e $\theta_r(\boldsymbol{x}) = \sum p_j(x) = 1$.

Considerado o modelo de chances proporcionais, a função de máxima verossimilhança é expressa por

$$L = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{r} [p_j(\boldsymbol{x_i})]^{y_{ij}} \right\} = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{r} [\theta_j(\boldsymbol{x_i}) - \theta_{j-1}(\boldsymbol{x_i})]^{y_{ij}} \right\}, \qquad (1)$$

no qual $y_{ij} = 1$ se a resposta do indivíduo $i, i = 1, .., n$, está na categoria j, $j = 1, ..., r$ e $y_{ij} = 0$, caso contrário, com $\sum_{j=1}^{r} y_{ij} = 1$. Para verificar a proporcionalidade das chances do modelo, pode-se utilizar o teste da razão de verossimilhança, em que entre o modelo sob a hipótese nula é o de proporcionalidade e o sob a alternativa é o de não proporcionalidade.

Em modelos nos quais não existem problemas de dados esparsos é possível verificar a qualidade do ajuste pelas estatísticas Q_p e Q_l dadas respectivamente por

$$Q_P = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad e \quad Q_L = 2\sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij} ln\left(\frac{n_{ij}}{e_{ij}}\right), \qquad (2)$$

em que $e_{ij} = (n_{i+})p_j(x_i), j = 1, ..., r$, a frequência esperada sob a hipótese nula de que o modelo é adequado. Essas estatísticas seguem distribuição assintótica qui-quadrado com $gl = (r-1)(s-1) - q$ em que r é o número de categorias da resposta, s o número de subpopulações e q o número de parâmetros do modelo, sem considerar β_{0j}.

A estatística Q_{HL}, de Hosmer-Lemeshow dada pela expressão

$$Q_{HL} = \sum_{i=1}^{g} \frac{[o_i - n_i \bar{p}(x_i)]^2}{n_i \bar{p}(x_i)[1 - n_i \bar{p}(x_i)]} \tag{3}$$

em que n_i é a frequência de observações no i-ésimo grupo, o_i é a frequência de resposta $Y = 1$ no i-ésimo grupo e $\bar{p}(x_i)$ é a probabilidade média da resposta $Y = 1$ estimada. Tal estatística segue aproximadamente uma X^2 com $gl = (g - 2)$.

Para o teste de Lipsitz et al.(1996) preciso atribuir uma pontuação ordinal a cada observação, utilizando pesos igualmente espaçados

$$s_i = \hat{p}_{i1} + 2\hat{p}_{i2} + ... + r\hat{p}_{ir}, \quad i = 1, ..., n.$$

as observações devem ser organizadas em g grupos baseados em um *score* ordinal.

$$I_{ik} = \begin{cases} 1, & \text{se a observação } i \text{ está no grupo } k \\ 0, & \text{caso conrário} \end{cases}$$

para $i = 1, 2, ..., n$ e $k = 1, ..., g-1$. Desse modo, é preciso ajustar um modelo incluindo o indicador I_{ik}

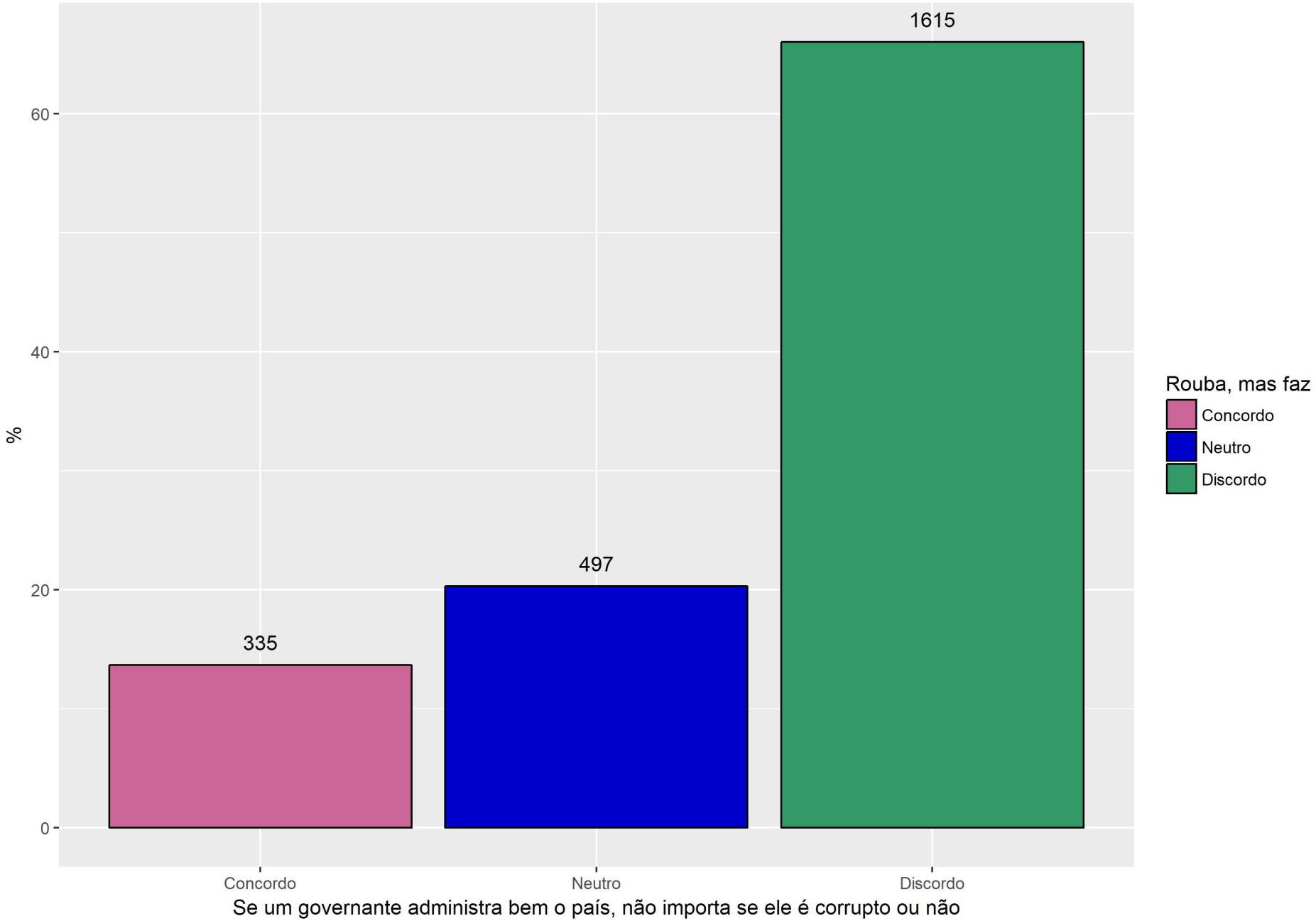
$$ln\left[\frac{\theta_j(\boldsymbol{x})}{1-\theta_j(\boldsymbol{x})}\right] = \beta_{0j}+\beta_1 x_1+...+\beta_2 x_2+\gamma_1 I_1+...+\gamma_{g-1} I_{g-1}, j = 1, ..., r-1$$

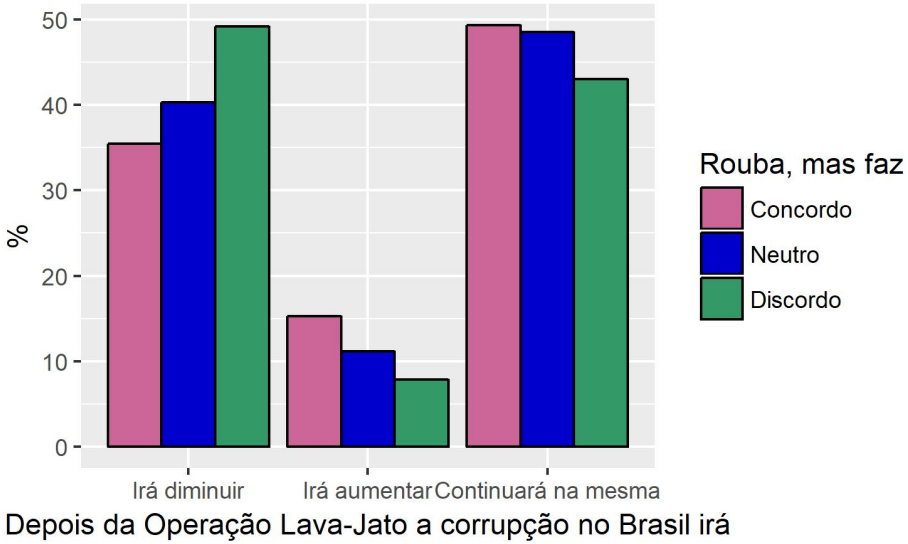
para conferir se o modelo foi corretamente ajustado, $\gamma_1 + ... + \gamma_{g-1} = 0$. Partindo disso, realiza-se o TRV $-2(L_1 - L_0) \sim X^2_{g-1}$, em que L_1 é o modelo padrão e L_0 o modelo com indicador binário(FAGERLAND & HOSMER, 2013).

Análise descritiva

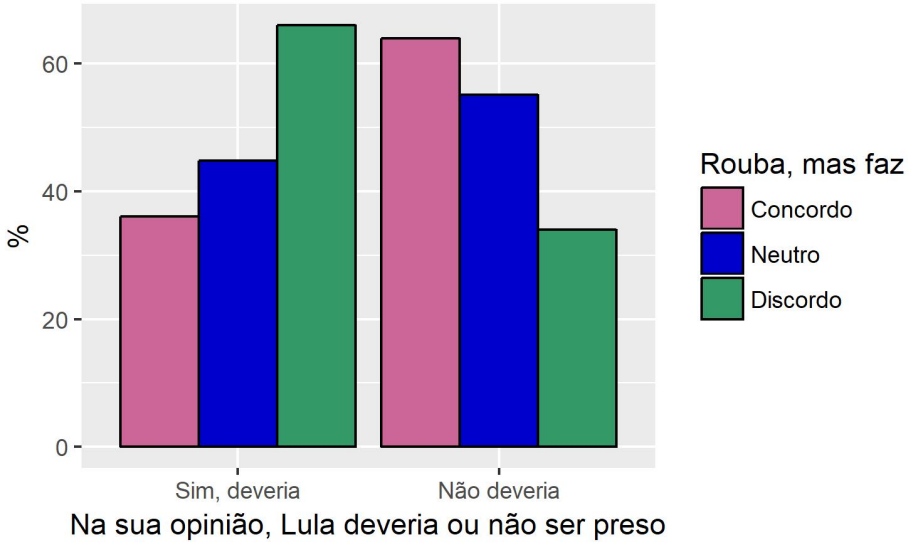
Variáveis utilizadas na investigação

p20a	Se um governante administra bem o país, não importa se ele é corrupto ou não	1. Condorda totalmente 2. Não concorda, nem discorda 3. Discorda totalmente
p15	Na sua opinião, depois da Operação Lava-Jato a corrupção no Brasil irá diminuir, aumentar ou continuará na mesma proporção	1. Irá diminuir 2. Irá aumentar 3. Continuará na mesma proporção de sempre
p16	Considerando o que foi revelado pela Operação Lava-Jato e seus desdobramentos até o momento, na sua opinião, Lula deveria estar preso?	1. Sim, deveria 2. Não deveria
p21a	Gostaria que você me dissesse com qual dessas três afirmações você concorda mais	1. A democracia é melhor que qualquer outra forma de governo; 2. Em certas circunstâncias, é melhor uma ditadura do que democracia; 3. Tanto faz se o governo é uma democracia ou uma ditadura?
escola	Até que ano da escola você estudou?	

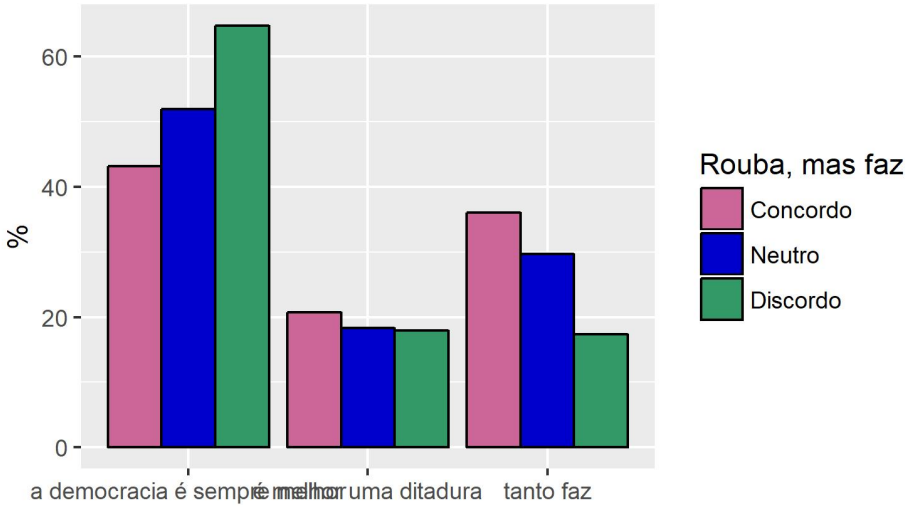

1615
497
335
60
40
20
0
%
Concordo
Neutro
Discordo
Se um governante administra bem o país, não importa se ele é corrupto ou não
Rouba, mas faz
Concordo
Neutro
Discordo



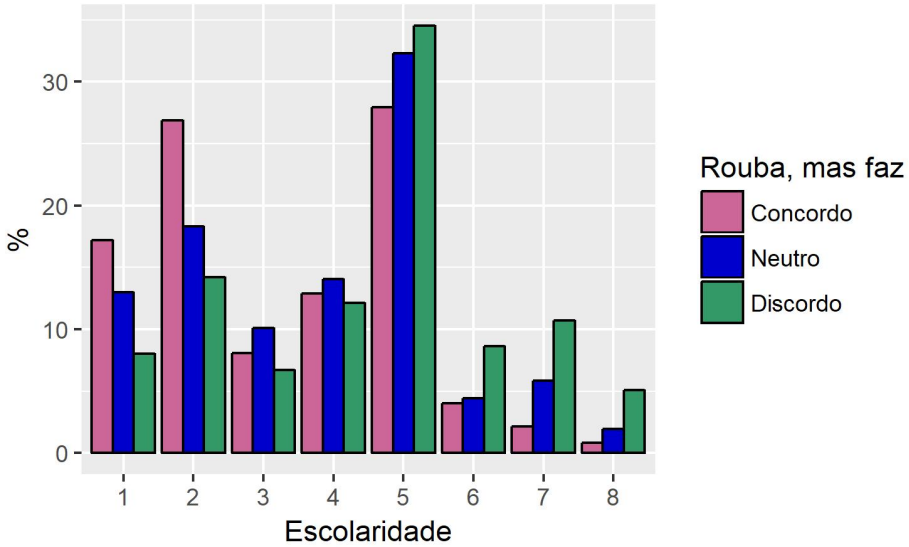
Depois da Operação Lava-Jato a corrupção no Brasil irá



Na sua opinião, Lula deveria ou não ser preso



Com qual dessas três afirmações você concorda mais



Modelo de chances proporcionais

Como observado a variável resposta é composta por 3 níveis, isto é, $r = 3$, dessa forma, os logitos cumulativos são definidos por

$$ln\left[\frac{\theta_1(x)}{1-\theta_1(x)}\right] = ln\left[\frac{P(Y \leq 1|x)}{P(Y > 1|x)}\right] = ln\left[\frac{p_1}{p_2 + p_3}\right],$$
$$ln\left[\frac{\theta_2(x)}{1-\theta_2(x)}\right] = ln\left[\frac{P(Y \leq 2|x)}{P(Y > 2|x)}\right] = ln\left[\frac{p_1 + p_2}{p_3}\right]$$

Para avaliar a suposição de chances proporcionais para as covariáveis, foram realizados testes da razão de verossimilhança entre os modelos de chances proporcionais e não proporcionais

$$ln\left[\frac{\theta_j(x)}{1-\theta_j(x)}\right] = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3 + \beta_{4j}x_4 + \beta_{5j}x_5 + \beta_{6j}x_6,$$

$$ln\left[\frac{\theta_j(x)}{1-\theta_j(x)}\right] = \beta_{0j} + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6$$

Sendo a hipótese nula do teste baseada na proporcionalidade das chances, isto é, $H_0 : \beta = \beta_j$, $j = 1, 2,$.

Teste da Razão de Verossimilhança

Modelos	TRV	gl	valor-p
Completo	3.5497	6	0.7374
p15	0.5792	2	0.7486
p16	0.0013	1	0.9711
p21a	1.0235	2	0.5994
escola	2.8775	1	0.0898

Existem evidência a favor do modelo de chances proporcionais, dado por

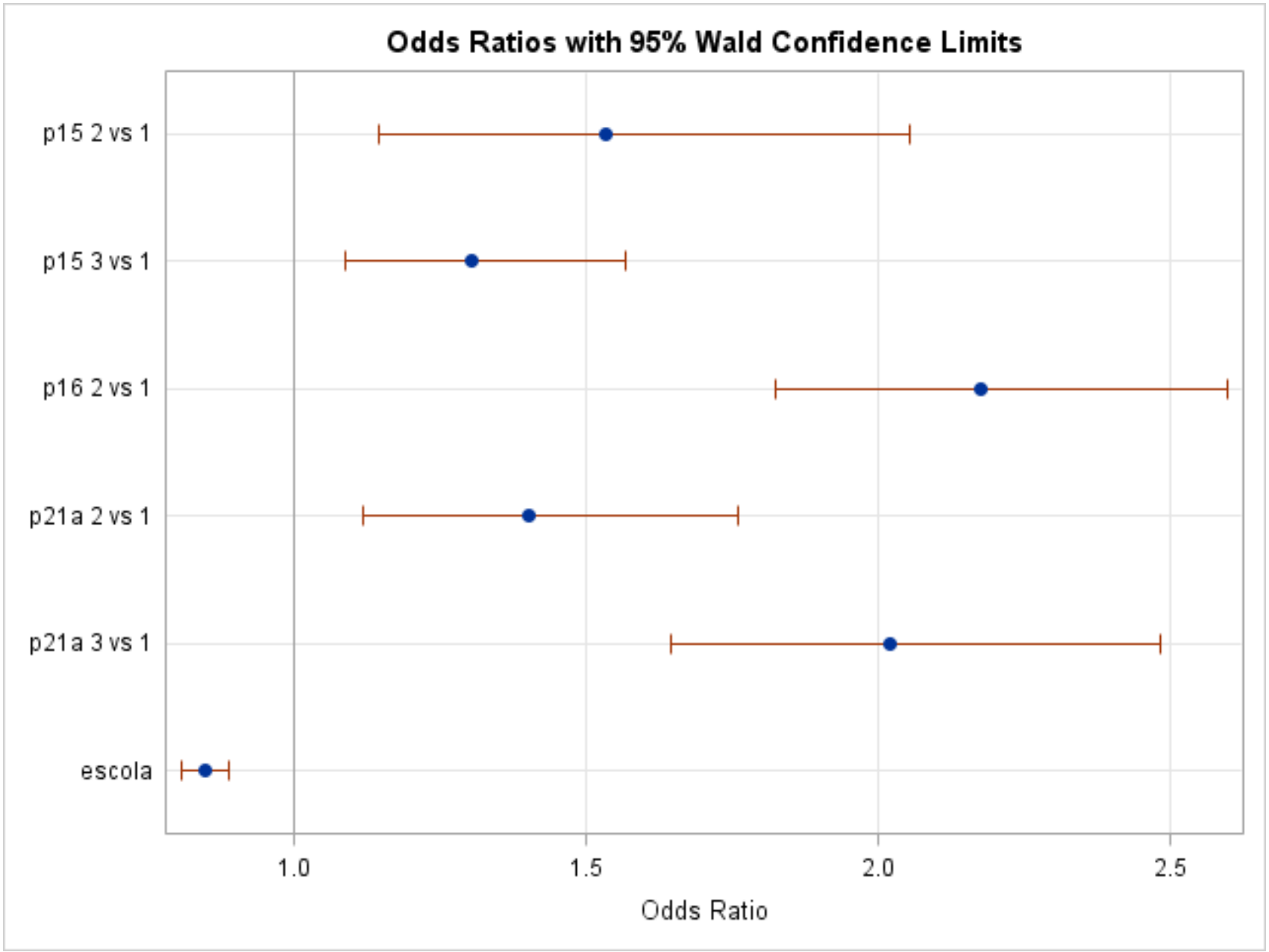
$$ln\left[\frac{\hat{\theta}_j(x)}{1-\hat{\theta}_j(x)}\right] = \hat{\beta}_{0j} + \hat{\beta}_1 p152 + \hat{\beta}_2 p1532 + \hat{\beta}_3 p162 + \hat{\beta}_4 p21a2 + \hat{\beta}_5 p21a3 + \hat{\beta}_6 escola$$

Resultados do modelo

Estimativas do modelo ordinal

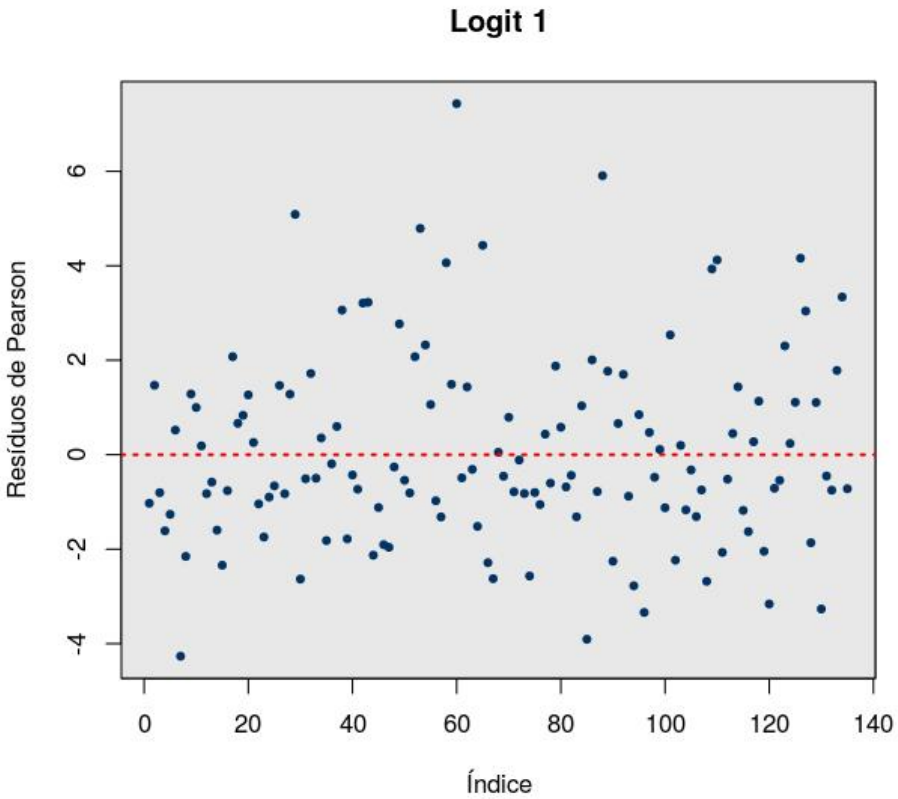
Parâmetros	Estimativas	E.P.	Z	Valor-p
β_{01}	-1.9969	0.1478	-13.5141	$<$0.0001
β_{02}	-0.7144	0.1414	-5.0533	$<$0.0001
β_1p15(aumentar)	0.4201	0.1471	2.8562	0.0043
β_2p15(na mesma)	0.2674	0.0927	2.8845	0.0039
β_3p16(não deveria)	0.7623	0.0896	8.5079	$<$0.0001
β_4p21a(ditadura)	0.3319	0.1150	2.8861	0.0039
β_5p21a(tanto faz)	0.6880	0.1044	6.5913	$<$0.0001
$\beta_6$escola	-0.1685	0.0242	-6.9721	$<$0.0001

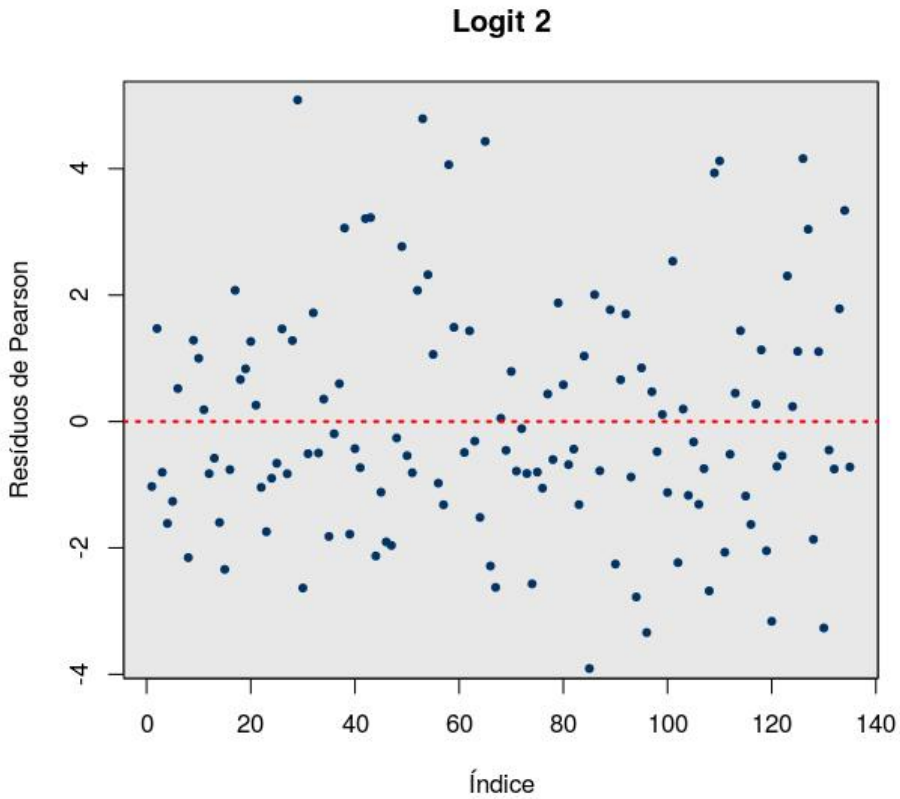
A Tabela 3 apresenta as estimativas dos parâmetros do modelo de chances proporcionais, primeiramente é preciso ressaltar a existência de dois parâmetros de intercepto.


Odds Ratios with 95% Wald Confidence Limits
p15 2 vs 1
p15 3 vs 1
p16 2 vs 1
p21a 2 vs 1
p21a 3 vs 1
escola
1.0
1.5
2.0
2.5
Odds Ratio

- No que diz respeito a variável p15 referente a opinião sobre a Lava-jato temos que os eleitores que acreditam que a operação irá aumentar a corrupção quando comparados a quem acha que ela irá diminuir, possuem 1.54 mais chances de discordar com a frase "Se um governante administra bem o país, não importa se ele é corrupto ou não". No caso em que o eleitor acredita que a após a Lava-jato a corrupção continuará mesma coisa, quando comparado a quem acha que irá diminuir, as chances de discordar da frase são 1.31 maiores.
- Quanto a p21a nota-se que os eleitores que apresentam menor afinidade com a democracia, tem mais chances de discordar da frase investigada. Neste caso, eleitores que não veem diferença entre democracia e ditadura, tem 2 vezes mais chances de discordar da frase exposta como resposta, do que os eleitores que acreditam que a democracia é a melhor forma de governo.

Pelo teste de Lipsitz, $Q_{Lip} = 13.54(gl = 9, valor-p = 0.1396)$, indicando evidências a favor do modelo ajustado.

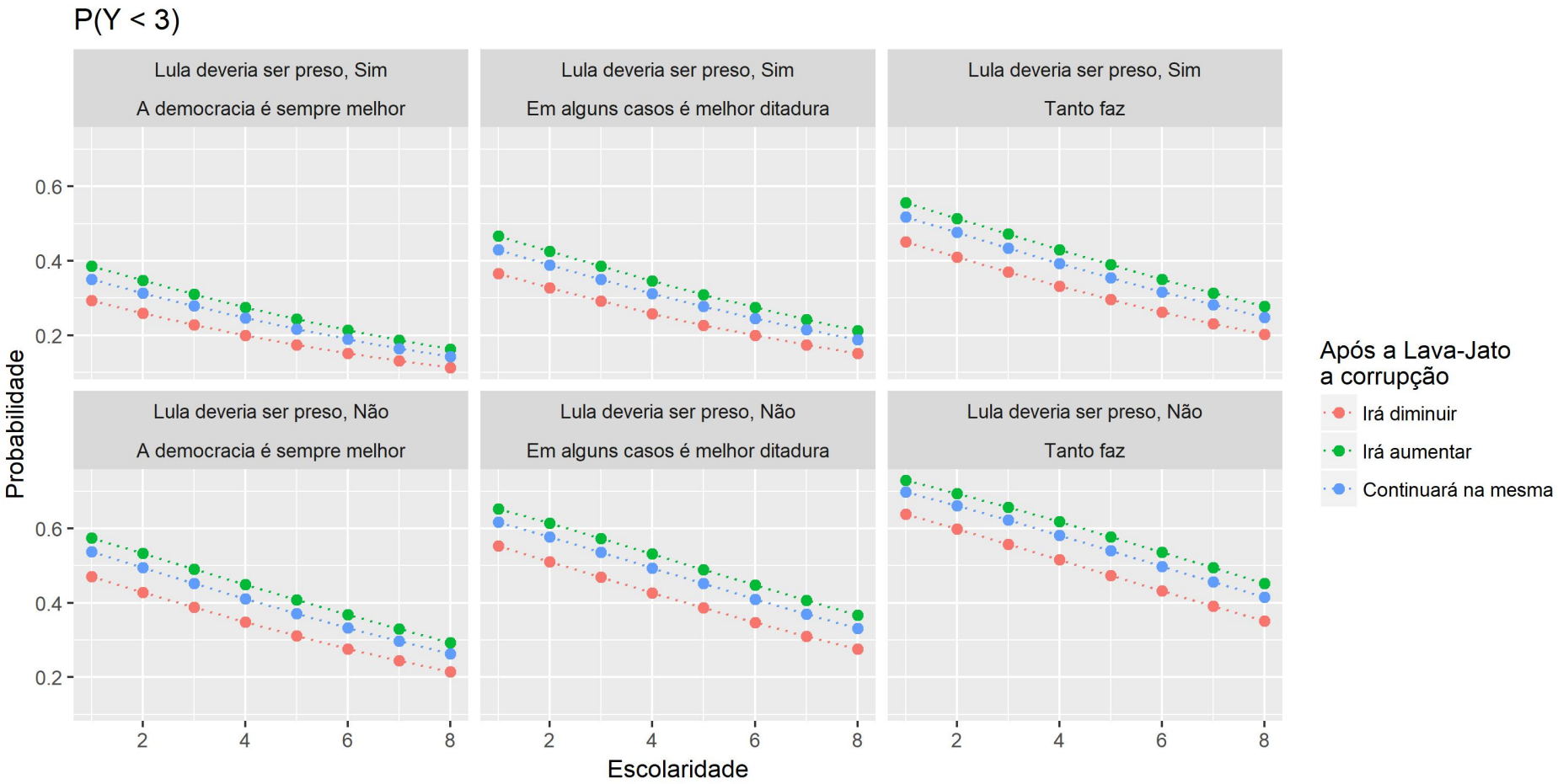




Probabilidades não cumulativas do modelo ordinal

id	escola	p15	p16	p21a	$\hat{p}_1(x)$	$\hat{p}_2(x)$	$\hat{p}_3(x)$
1	1	1	1	1	0.1029	0.1897	0.7074
2	2	1	1	1	0.0884	0.1706	0.7410
3	3	1	1	1	0.0757	0.1523	0.7720
4	4	1	1	1	0.0647	0.1350	0.8003
5	5	1	1	1	0.0552	0.1189	0.8259
6	6	1	1	1	0.0471	0.1041	0.8488
7	7	1	1	1	0.0401	0.0907	0.8692
8	8	1	1	1	0.0341	0.0788	0.8872
.	.	.	.	.	.	.	.
137	1	3	2	3	0.3899	0.3075	0.3026
138	2	3	2	3	0.3506	0.3100	0.3393
139	3	3	2	3	0.3133	0.3086	0.3780
140	4	3	2	3	0.2782	0.3034	0.4184
141	5	3	2	3	0.2457	0.2944	0.4599
142	6	3	2	3	0.2158	0.2823	0.5019
143	7	3	2	3	0.1887	0.2674	0.5439
144	8	3	2	3	0.1642	0.2505	0.5853

Para obter as probabilidades cumulativas, deve-se considerar $\hat{p}_1(x) = \hat{\theta}_1(x)$ e para $P(Y \leq 2|x) = \hat{\theta}_2(x) = \hat{p}_1(x) + \hat{p}_2(x)$.


P(Y < 3)
Lula deveria ser preso, Sim
A democracia é sempre melhor
Lula deveria ser preso, Sim
Em alguns casos é melhor ditadura
Lula deveria ser preso, Sim
Tanto faz
Lula deveria ser preso, Não
A democracia é sempre melhor
Lula deveria ser preso, Não
Em alguns casos é melhor ditadura
Lula deveria ser preso, Não
Tanto faz
Probabilidade
0.2
0.4
0.6
2
4
6
8
Escolaridade
Após a Lava-Jato a corrupção
Irá diminuir
Irá aumentar
Continuará na mesma

Referências

- Giolo SR. Introdução a análise de dados categóricos. Curitiba, Universidade Federal do Paraná-Departamenmto de Estatística. 2006.
- Agresti A. Categorical data analysis. John Wiley & Sons; 2003 Mar 31.
- Christensen RH, Christensen MR. Package 'ordinal'. Stand. 2015 Jun 28;19:2016.
- Derr, B., 2013. Ordinal response modeling with the LOGISTIC procedure. In SAS Global Forum pp. 1-20.
- Fagerland, M.W. and Hosmer, D.W., 2013. A goodness-of-fit test for the proportional odds regression model. Statistics in medicine, 32(13), pp.2235-2249.

Muito obrigado!