

Modelo de Regressão Binária

Wesley Furriel e André F. B. Menezes

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

12 de Abril de 2018

Organização

- 1 Modelos de Regressão
Modelos de Lineares Generalizados
- 2 Modelos de Regressão Binária
Regressão Logística
- 3 Aplicações
Aborto
Corrupção
- 4 Recursos computacionais
- 5 Referências

Modelos de Regressão

Permitem a inclusão de variáveis explicativas (covariáveis) para:

- ▶ Descrever a relação entre a variável resposta e as variáveis preditoras;
- ▶ Realizar predições por meio do modelo estabelecido.

Sem perdas de generalidade podemos expressar um modelo de regressão da seguinte forma:

$$\begin{aligned} Y \mid \mathbf{X} &\sim f(\boldsymbol{\theta}) \\ Q(Y \mid \mathbf{X}) &= g(\mathbf{X} \mid \boldsymbol{\beta}) \end{aligned} \tag{1}$$

em que

- ▶ $f(\boldsymbol{\theta})$ é a f.d.p. de alguma distribuição indexada por um parâmetro $\boldsymbol{\theta}$;
- ▶ $Q(Y \mid \mathbf{X})$ é alguma quantidade de interesse (média, quantil, parâmetro) de Y condiciada as covariáveis; e
- ▶ $g(\mathbf{X} \mid \boldsymbol{\beta})$ é uma função de ligação utilizada para associar as covariáveis com a quantidade de interesse.

Definição

Dada uma amostra aleatória de n observações (y_i, \mathbf{x}_i) , em que $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$, os MLG são definidos por:

1. Componente aleatório:

$$f(y_i | \theta_i, \phi) = \exp \left\{ \frac{1}{a(\phi)} [y_i \phi + b(\theta_i)] + c(y_i, \phi) \right\} \quad (2)$$

2. Componente sistemático:

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{ou} \quad \boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta} \quad (3)$$

3. Função de ligação:

$$\eta_i = g(\mu_i) \quad (4)$$

Dados de contagem

- ▶ Modelo Binomial;
- ▶ Modelo Geométrico;
- ▶ Modelo Poisson;
- ▶ Modelo Binomial Negativo;
- ▶ Modelo Multinomial.

Dados de contínuos

- ▶ Modelo Normal;
- ▶ Modelo Gamma;
- ▶ Modelo Inversa-Gaussiana.

Modelos de Regressão Binária

Componente aleatório

Seja Y uma variável resposta binária em que seus dois resultados possíveis são 1 (“sucesso”) e 0 (“fracasso”). A função distribuição de probabilidade de Y é dada por

$$\Pr(Y = y) = \pi^y (1 - \pi)^{1-y}, \quad y = 0, 1 \quad (5)$$

em que π denota a probabilidade de sucesso.

A média e variância de Y são dadas, respectivamente, por

$$\mathbb{E}(Y) = \pi \quad \text{e} \quad \text{Var}(Y) = \pi (1 - \pi). \quad (6)$$

Componente sistemático

Sejam Y_1, \dots, Y_n variáveis aleatórias independentes, em que $Y_i \sim \mathcal{B}(\pi_i)$ para $i = 1, \dots, n$. O modelo de regressão binário é definido supondo que a média dos Y_i satisfazem a seguinte relação funcional

$$g(\pi_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} \quad (7)$$

em que $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ e $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})$ são vetores desconhecidos de dimensões $(p \times 1)$ representando, respectivamente, as covariáveis e seus respectivos parâmetros fixos e desconhecidos.

Função de ligação

- ▶ Devemos assumir que a função de ligação $g(\cdot)$ seja apropriada, isto é, uma função estritamente monótona e duas vezes diferenciável que leva $(0, 1)$ aos \mathbb{R} .
- ▶ Existem diversas possibilidades para a função de ligação. Por exemplo:

$$\text{logit} \rightarrow g(\theta) = \log \left(\frac{\theta}{1 - \theta} \right)$$

$$\text{probit} \rightarrow g(\theta) = \Phi^{-1}(\theta)$$

$$\text{clog-log} \rightarrow g(\theta) = \log(-\log(1 - \theta))$$

$$\text{cauchy} \rightarrow g(\theta) = \tan(\pi(\theta - 0.5))$$

em que $\Phi(\cdot)$ denota a função de distribuição acumulada da Normal padrão.

Verossimilhança

Independente da função de ligação adotada a função de verossimilhança para β é dada por

$$\mathcal{L}(\beta \mid \mathbf{y}, \mathbf{x}) = \prod_{i=1}^n [\pi(\mathbf{x}_i)]^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i} \quad (8)$$

em que $\pi(\mathbf{x}_i) = g^{-1}(\mathbf{x}_i^\top \beta)$.

Regressão Logística

- ▶ Sem dúvidas a função de ligação logit é a mais utilizada na prática para relacionar a probabilidade π_i as variáveis explicativas \mathbf{x}_i^T .
- ▶ A função de ligação logit permite uma simples representação da razão de chances, o que auxilia na interpretação dos resultados.

Considerações

- ▶ Seja a variável resposta Y binária tal que
 - ▶ $Y_i = 1$ a ocorrência do evento de interesse (Evento);
 - ▶ $Y_i = 0$ ausência do evento de interesse (Referência).
- ▶ $\mathbf{X} = (X_1, \dots, X_k)^\top$ é um vetor de variáveis exploratórias, que podem ser discretas, contínuas ou categóricas;
- ▶ As variáveis categóricas são incorporadas ao modelo por meio de matrizes de variáveis *dummy*.

O modelo

$$\pi_i = \Pr(Y_i = 1 | X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})},$$

isto é,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$$

em que π_i denota a probabilidade de ocorrência do evento de interesse.

Suposições

- ▶ Os dados y_1, y_2, \dots, y_n são i.i.d.
- ▶ $Y_i \sim \mathcal{B}(\pi_i)$ isto é, assume a distribuição bernoulli da resposta.
- ▶ Não existe uma relação linear entre Y_i e X , mas sim, entre a função de ligação e o preditor linear.
- ▶ Não homogeneidade da variância..
- ▶ Os erros precisam ser independentes, mas não normalmente distribuídos.
- ▶ As medidas de qualidade do modelo dependem de amostras suficientemente grandes, evitando assim, subpopulações muito pequenas.

Inferência sob β

Estimação de β

- Verossimilhança

$$\mathcal{L}(\beta \mid \mathbf{y}, \mathbf{x}) = \prod_{i=1}^n [\pi(\mathbf{x}_i)]^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i}$$

- Log-Verossimilhança

$$\ell(\beta \mid \mathbf{y}, \mathbf{x}) = \sum_{i=1}^n y_i \log [\pi(\mathbf{x}_i)] + (1 - y_i) [1 - \pi(\mathbf{x}_i)]$$

- Sendo que

$$\pi_i(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \beta)}{1 + \exp(\mathbf{x}_i^\top \beta)}$$

Estimação das variâncias-covariâncias de $\hat{\beta}$

- Matriz de informação Fisher observada

$$-\frac{\partial^2 \ell}{\partial \beta_j^2} = \sum_{i=1}^n x_{ij}^2 \pi(\mathbf{x}_i) (1 - \pi(\mathbf{x}_i))$$

e

$$-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_l} = \sum_{i=1}^n x_{ij} x_{il} \pi(\mathbf{x}_i) (1 - \pi(\mathbf{x}_i))$$

para $j, l = 0, 1, 2, \dots, p$

Significância dos coeficientes estimados

- ▶ Testar hipóteses relativas a(os) parâmetro(s), isto é,

$$\mathcal{H}_0 : \beta = 0 \quad (\text{vetor } q \times 1)$$

- ▶ Teste da razão de verossimilhanças

$$S_{LR} = 2 \left[\ell(\hat{\beta} \mid \mathbf{y}, \mathbf{x}) - \ell(\beta \mid \mathbf{y}, \mathbf{x}) \right] \sim \chi_{p-q}^2$$

- ▶ Teste de Wald

$$S_W = \left(\hat{\beta} \right)^\top \left[\mathbf{I} \left(\hat{\beta} \right) \right]^{-1} \left(\hat{\beta} \right) \sim \chi_{p-q}^2$$

Regressão Logística

Interpretação dos coeficientes

- Considere um modelo de regressão logística com variável independente, x , codificada em 0 ou 1, isto é,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n.$$

- Assim, verifica-se que

$$\text{OR} = \frac{\frac{\pi(x_i = 1)}{1 - \pi(x_i = 1)}}{\frac{\pi(x_i = 0)}{1 - \pi(x_i = 0)}} = \frac{\left(\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right)}{\left(\frac{1}{1 + e^{\beta_0 + \beta_1}} \right)} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

- ▶ Os resultados do ajuste somente serão válidos se o modelo estiver adequado.
- ▶ Pode-se utilizar as estatísticas de Pearson e Deviance para avaliar a adequação do modelo
- ▶ Porém, estas estatísticas somente terão validades se $N < n$
 - ▶ N : número de possíveis combinações das covariáveis;
 - ▶ n : tamanho amostral.
- ▶ Quando esta suposição é violada, sugere-se o uso da estatística de Hosmer-Lemeshow.

Diagnóstico do Modelo

- Para testar \mathcal{H}_0 : o modelo tem ajuste satisfatório , têm-se:

$$Q_P = \sum_{i=1}^r \sum_{j=1}^2 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi_m^2$$

e

$$Q_L = 2 \sum_{i=1}^r \sum_{j=1}^2 n_{ij} \log \left(\frac{n_{ij}}{e_{ij}} \right) \sim \chi_m^2$$

em que

$$e_{ij} = n_{i+} \hat{\pi}(\mathbf{x}_i), j = 1 \quad \text{e} \quad e_{ij} = n_{i+} (1 - \hat{\pi}(\mathbf{x}_i)), j = 2$$

- n_{i+} = n° de observações na i -ésima subpopulação da tabela de dados $r \times 2$.
- $\hat{\pi}(\mathbf{x}_i)$ = probabilidade predita pelo modelo ajustado.
- e_{ij} = frequências esperadas do modelo ajustado.
- m = n° de subpopulações – n° de parâmetros estimados.

- ▶ Na presença de variáveis contínuas pode-se ter frequências pequenas na maioria das r subpopulações, inviabilizando o uso de Q_P e Q_L .
- ▶ Hosmer e Lemeshow (1989) propuseram uma alternativa baseada na estatística de Pearson para uma tabela $g \times 2$.

► O teste consiste em:

1. Ordenar as n observações em ordem crescente das probabilidades $\hat{\pi}(\mathbf{x}_i)$ preditas pelo modelo
2. Divide-se as observações em g grupos, recomenda-se $g = 10$, então calcula-se a estatística definida por

$$Q_{HL} = \sum_{i=1}^g \frac{[o_i - n_i \tilde{\pi}(\mathbf{x}_i)]^2}{n_i \tilde{\pi}(\mathbf{x}_i) [1 - \tilde{\pi}(\mathbf{x}_i)]} \sim \chi_{g-2}^2$$

n_i é número de observações no grupo i ,

o_i é a frequência observada da resposta no grupo i e

$\tilde{\pi}(\mathbf{x}_i)$ probabilidade média estimada da resposta no grupo i .

Diagnóstico do Modelo

- ▶ Para avaliar o poder preditivo ou classificatório do modelo deve-se estabelecer um ponto de corte ($0 < k < 1$), tal que
 - ▶ Probabilidades preditas pelo modelo $\geq k \rightarrow Y = 1$,
 - ▶ Probabilidades preditas pelo modelo $< k \rightarrow Y = 0$;

Observado	Predito		Totais
	$Y = 1 (+)$	$Y = 0 (-)$	
$Y = 1$	a	b	$a + b$
$Y = 0$	c	d	$c + d$
Totais	$a + c$	$b + d$	n

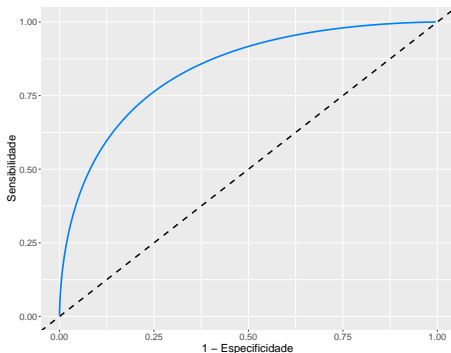
$$\text{Sensibilidade} = \frac{a}{a + b} = \text{taxa de verdadeiros +}$$

$$\text{Especificidade} = \frac{d}{c + d} = \text{taxa de verdadeiros -}$$

$$\text{Valor Preditivo} = \frac{a + d}{n} = \text{proporção geral de acertos}$$

Diagnóstico do Modelo

- ▶ Para diferentes pontos de cortes k têm-se a famigerada curva ROC
 - ▶ Plotando os pares $(x, y) = (1 - \text{especificidade}, \text{sensibilidade})$;
 - ▶ Modelo com discriminação perfeita $\rightarrow (x, y) = (0, 1)$;
 - ▶ Quanto mais próxima de 1 for a área abaixo da curva melhor o poder de predição do modelo.



Aplicações

Descrição

- ▶ **Aborto:** questionário de 2016 aplicado pelo LAPOP;
- ▶ **Corrupção:** questionário de 2017 conduzido pelo Data Folha.

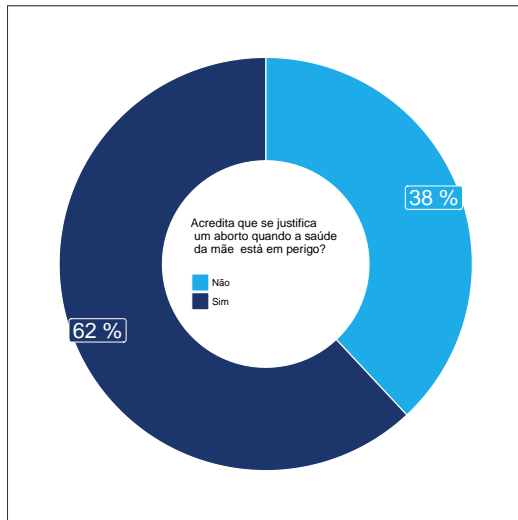
Aborto

Banco de dados

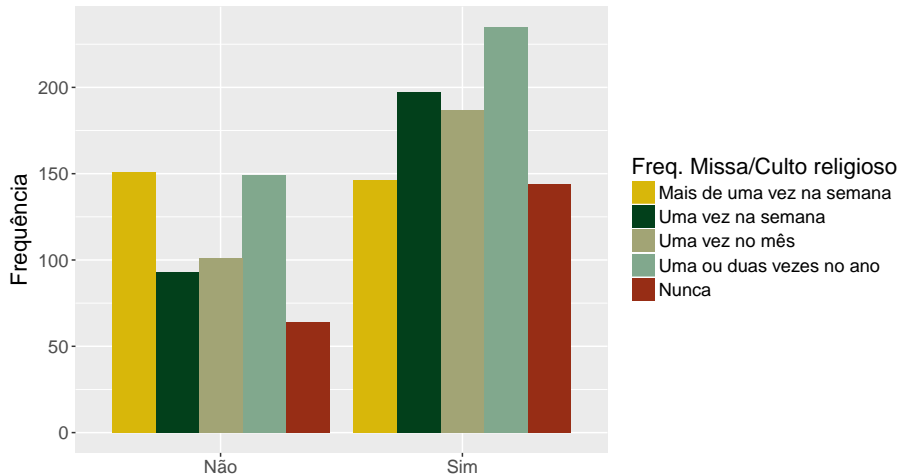
- ▶ Ano: 2016
- ▶ Número de observações: 1533
- ▶ Número de variáveis: 236
- ▶ Fonte: <https://www.vanderbilt.edu/lapop/>

Variáveis selecionadas

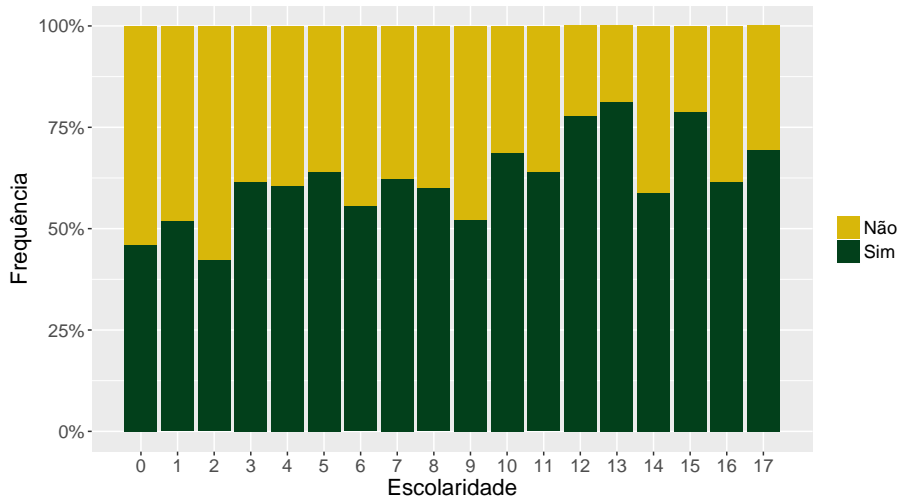
- ▶ ABORTO: “E agora, pensando em outros assuntos. O(a) sr./sra acredita que se justifica a interrupção da gravidez, ou seja, um aborto, quando a saúde da mãe está em perigo?”
- ▶ ESCOLARIDADE: “Qual foi o último ano de escola que o(a) sr./sra. terminou?”
- ▶ MISSA/CULTO: “Com que frequência o(a) sr./sra. vai à missa ou culto religioso?”



Aborto



Aborto



Modelo

$$\begin{aligned} \log \left(\frac{\pi_i}{1 - \pi_i} \right) = & \beta_0 + \beta_1 \text{Escolaridade}_i + \beta_2 \text{MissaSemana}_i \\ & + \beta_3 \text{MissaMes}_i + \beta_4 \text{MissaAno}_i + \beta_5 \text{MissaNunca}_i, \end{aligned}$$

em que π_i denota a probabilidade de achar que o aborto é justificável quando a saúde da mãe está em perigo.

Resumo inferências do modelo (9)

Parâmetro	$\hat{\beta}$	EP $\left(\hat{\beta}\right)$	$\exp\left(\hat{\beta}\right)$	I.C. 95%
Intercepto	-0.3587	0.1681	0.6986	(0.5019, 0.9706)
Escolaridade	0.0398	0.0148	1.0406	(1.0108, 1.0714)
MissaSemana	0.4594	0.1571	1.5831	(1.1642, 2.1557)
MissaMes	0.6146	0.1703	1.8489	(1.3259, 2.5860)
MissaAno	0.8188	0.1905	2.2677	(1.5659, 3.3069)
MissaNunca	0.7629	0.1717	2.1445	(1.5343, 3.0093)

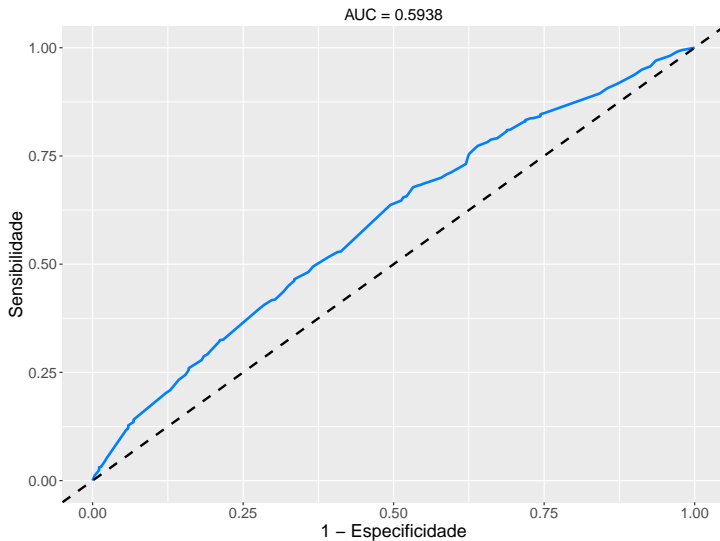
* Referência: mais de uma vez na semana.

Estatísticas de qualidade de ajuste do modelo (9)

Critério	Valor	G.L.	valor- p
Pearson	98.2827	84	0.1366
Deviance	86.1203	84	0.4154
Hosmer-Lemeshow	6.8375	8	0.5543

Aborto

- Curva ROC

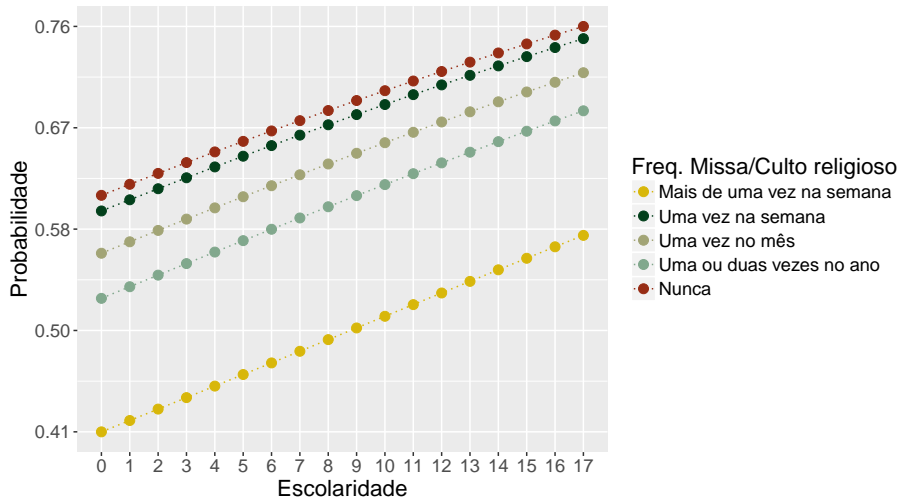


Interpretações dos resultados

- ▶ Escolaridade $\left[\exp \left(\hat{\beta}_1 \right) = 1.0406 (1.0108, 1., 0714) \right]$: para o aumento de um ano na escolaridade a **chance** da pessoa acreditar que o aborto seja justificado **aumenta em 4,06%**
- ▶ MissaAno $\left[\exp \left(\hat{\beta}_4 \right) = 2.2677 (1.5659, 3.3069) \right]$: pessoas que não frequentam missa/culto religiosos possuem **2.2677 vezes mais chance** de acreditar que o aborto seja justificado ao comparado com pessoas que frequentam duas ou mais vezes na semana.

Aborto

• Predição



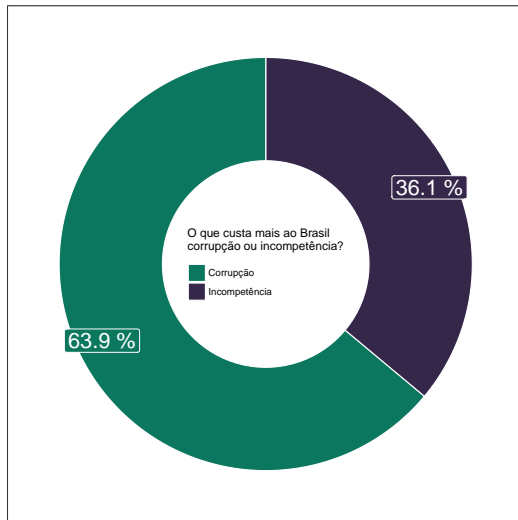
Corrupção

Banco de dados

- ▶ Ano: 2017
- ▶ Número de observações: 2772
- ▶ Número de variáveis: 130
- ▶ Fonte: https://www.cesop.unicamp.br/eng/banco_de_dados/v/4253

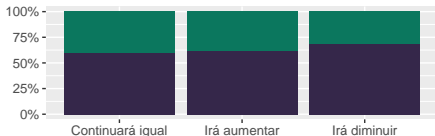
Variáveis selecionadas

- ▶ CORRUPCAO: “O que custa mais ao Brasil corrupção ou incompetência”
- ▶ SEXO: “Sexo do entrevistado”
- ▶ IDADE: “Idade do entrevistado”
- ▶ RENDAF: “Renda familiar”
- ▶ ESCOLA: “Escolaridade do entrevistado”
- ▶ LJATO: “Depois da Operação Lava-Jato a corrupção no Brasil irá”
- ▶ PARTIDO: “Você tem preferência por algum partido”
- ▶ AVALTEMER: “Avaliação do governo Temer”
- ▶ DENUNCIATEMER: “Os deputados federais deveriam autorizar a segunda denúncia do Ministério Público contra Michel Temer”



Corrupção

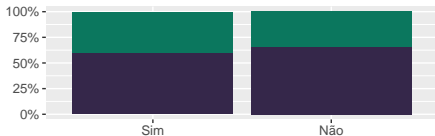
Após a Lava-Jato a corrupção no Brasil



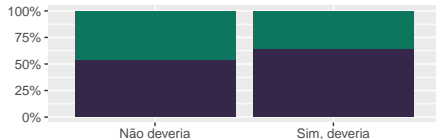
Sexo



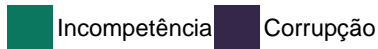
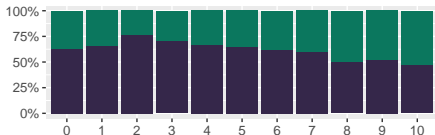
Simpatiza com algum partido



Denúncia contra Temer

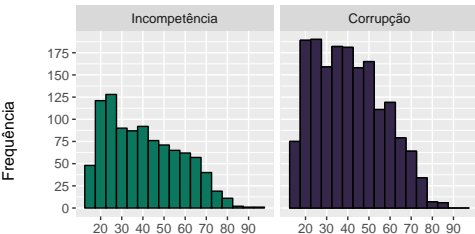




Avaliação governo Temer



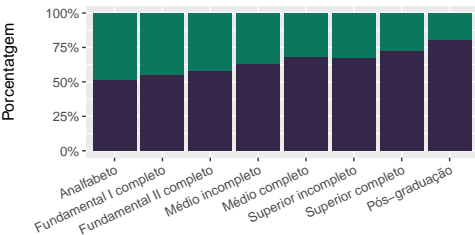
Corrupção

Idade

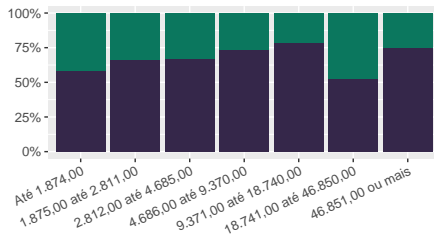


 Incompetência  Corrupção

Escolaridade



Renda



Foram considerados os seguintes modelos:

$$m_0 : \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{renda}_i + \beta_3 \text{escola}_i + \beta_4 \text{Ljato}_i + \beta_5 \text{partido2}_i + \\ \beta_6 \text{LjatoAumentar}_i + \beta_7 \text{LjatoDiminuir}_i + \beta_8 \text{DenúnciaTemerSim}_i + \\ \beta_9 \text{sexo}_i$$

$$m_1 : \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{escola}_i + \beta_3 \text{LjatoAumentar}_i + \beta_4 \text{LjatoDiminuir}_i + \\ \beta_5 \text{DenúnciaTemerSim}_i$$

$$m_2 : \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{LjatoAumentar}_i + \beta_3 \text{LjatoDiminuir}_i + \beta_4 \text{DenúnciaTemerSim}_i$$

$$m_3 : \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{escola}_i + \beta_3 \text{LjatoAumentar}_i + \beta_4 \text{LjatoDiminuir}_i$$

$$m_4 : \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{escola}_i + \beta_3 \text{DenúnciaTemerSim}_i$$

$$m_5 : \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{escola}_i + \beta_2 \text{LjatoAumentar}_i + \beta_3 \text{LjatoDiminuir}_i + \beta_4 \text{DenúnciaTemerSim}_i$$

Seleção do modelo

Teste da Razão de Verossimilhança

Modelos	ℓ_1	ℓ_2	χ^2	Valor- p
$m_0 - m_1$	-1446.6	-1450.1	6.972	0.1374
$m_1 - m_2$	-1450.1	-1483.9	67.565	<0.0001
$m_1 - m_3$	-1450.1	-1454.5	8.772	0.0030
$m_1 - m_4$	-1450.1	-1463.3	26.398	<0.0001
$m_1 - m_5$	-1450.1	-1459.2	18.271	0.0001

O modelo selecionado foi o m_1 .

Modelo

$$\begin{aligned} \log \left(\frac{\pi_i}{1 - \pi_i} \right) &= \beta_0 + \beta_1 \text{Idade}_i + \beta_2 \text{Escolaridade}_i + \beta_3 \text{LjatoAumentar}_i \\ &+ \beta_4 \text{LjatoDiminuir}_i + \beta_5 \text{DenúnciaTemerSim}_i, \quad i = 1, \dots, n. \end{aligned}$$

em que π_i denota a probabilidade de considerar a corrupção como o problema que mais custa ao Brasil.

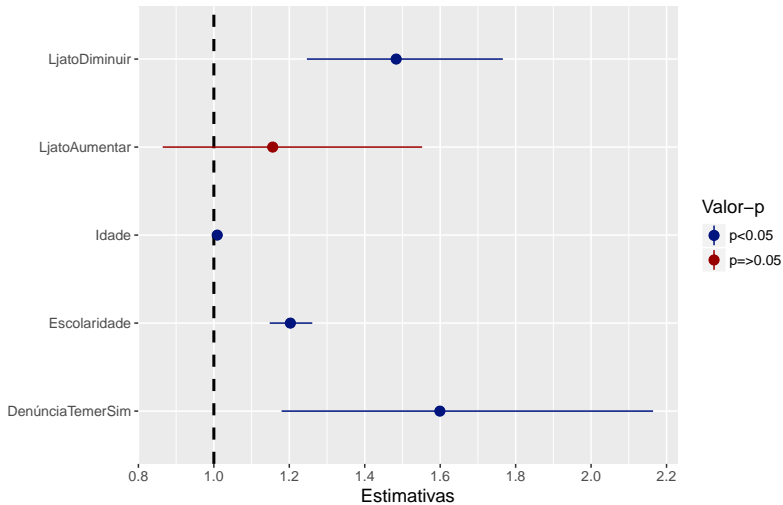
Resumo inferências do modelo (11)

Parâmetro	$\hat{\beta}$	EP $\left(\hat{\beta}\right)$	$\exp\left(\hat{\beta}\right)$	I.C. 95%
Intercepto	-1.1685	0.2344	0.3108	(0.1959, 0.4913)
Idade	0.0089	0.0027	1.0089	(1.0036, 1.0143)
Escolaridade	0.1847	0.0239	1.2029	(1.1481, 1.2609)
LjatoAumentar ¹	0.1449	0.1492	1.1560	(0.8644, 1.5522)
LjatoDiminuir ¹	0.3943	0.0889	1.4833	(1.2465, 1.7660)
DenúnciaTemerSim ²	0.4695	0.1546	1.5992	(1.1796, 2.1643)

¹Referência: Corrupção após Lava-Jato continuará igual.

²Referência: Deputados deveriam autorizar denúncia contra Temer. Sim

Corrupção



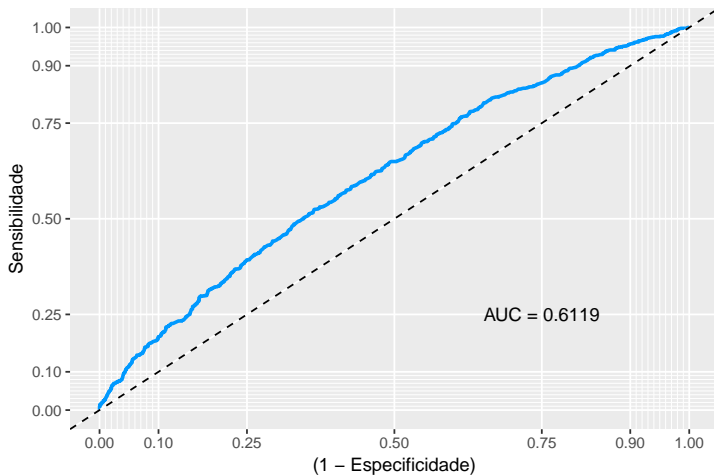
Interpretações dos resultados

- ▶ Idade: $\left[\exp \left(\hat{\beta}_2 \right) = 1.0089 (1.0036, 1.0143) \right]$: para o aumento de um nível na escolaridade a **chance** da pessoa acreditar que a corrupção custa mais ao Brasil **aumenta em 20,29%**
- ▶ Ljato: $\left[\exp \left(\hat{\beta}_4 \right) = 1.4833 (1.2465, 1.7660) \right]$: pessoas crentes de que a Lava Jato irá diminuir a corrupção no Brasil tem **1.4833 vezes mais chance** de acreditar que a corrupção custa mais ao Brasil quando comparadas as que acreditam que a incompetência custa mais.

Estatísticas de qualidade de ajuste do modelo (11)

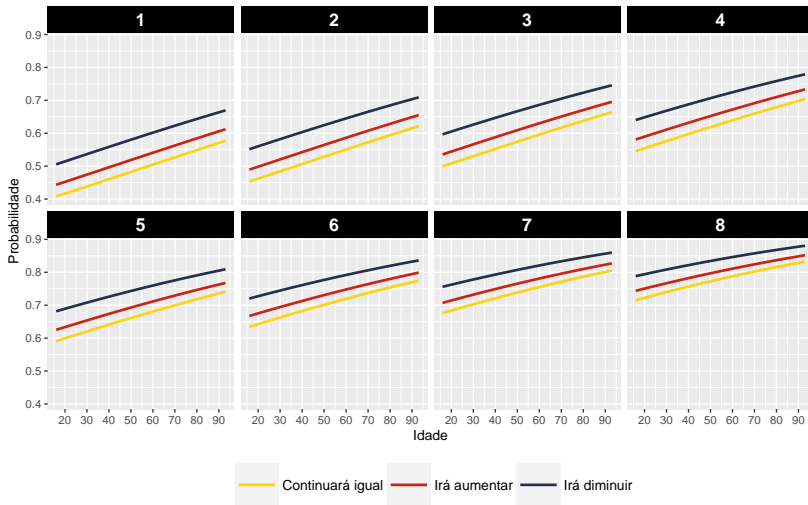
Critério	Valor	G.L.	valor- p
Pearson	1045.8403	1003	0.1690
Deviance	1313.5410	1003	<0.0001
Hosmer-Lemeshow	2.7110	8	0.9512

- Curva ROC



Corrupção

Predição quando Deputados deveriam autorizar denúncia contra Temer? = Sim



Recursos computacionais

Recursos computacionais

Software SAS® 9.4

- ▶ PROC GENMOD
- ▶ PROC LOGISTIC
- ▶ PROC CATMOD

Software R, versão 3.4.3

- ▶ Função glm

Software Python, versão 3.5

- ▶ Função sm.Logit

• PROC LOGISTIC

```
1 proc logistic data = dados plots = roc;  
2   class freqmissa(ref = "More than once a week") / param = ref;  
3   model aborto(event = "Sim") = escolaridade freqmissa /  
4   link = logit expb ctable lackfit  
5   aggregate = (freqmissa escolaridade) scale = none;  
6 quit;
```

• PROC GENMOD

```
1 proc genmod data = dados;  
2   class freqmissa(ref = "More than once a week") / param = ref;  
3   model aborto = escolaridade freqmissa / dist = bin  
4   link = logit lrci;  
5 quit;
```

- glm

```
1 mod.aborto <- glm(aborto ~ escolaridade + freqmessa,  
2   data = dados, family = binomial(link = "logit"))  
3 summary(mod.aborto)  
4 exp(cbind(coef(mod.aborto), confint(mod.aborto)))
```

• Logit

```
1 import scipy
2 import pandas as pd
3 import statsmodels.api as sm
4 cols = ["idade1", "escola", "Ljato_lra_aumentar", "
        Ljato_lra_diminuir", "denunciaTemer_Sim_deveria"]
5 X = dados[cols]
6 y = dados["corrupcao_Corrupcao"]
7 logit_model = sm.Logit(y, X)
8 result = logit_model.fit()
9 print(result.summary())
```

- [1] Agresti, A., 1990. **Categorical Data Analysis**. Wiley: New York.
- [2] Agresti, A., 2007. **An Introduction to Categorical Data Analysis**. John Wiley & Sons, Inc: New Jersey.
- [3] Hosmer, D. W.; Lemeshow, S.; Sturdivant, R. X. **Applied Logistic Regression**. Third. John Wiley & Sons, Inc, 2013.
- [4] Stokes, M. E.; Davis, C. S.; Koch., G. G. **Categorical Data Analysis using the SAS System**. Second edition Cary, NC: SAS Institute Inc, 2000.

Muito obrigado!