Regressão beta bayesiana

Abstenção no 2º turno de 2026 × voto em Lula no 1º turno

Hedibert Freitas Lopes

Advanced Econometrics
Mestrado Profissional em Economia – Insper

30 de outubro de 2026

Slides preparados com assistência de IA (Claude, Anthropic) e revisados pelo autor. Dados: TSE via imprensa; abstenção do 2º turno por UF *estimada*.

Roteiro

Pergunta e dados

Modelo

MCMC

Resultados

Discussão

A pergunta

- A abstenção caiu de 21,1% (1º turno) para 19,2% (2º turno), mas de forma muito desigual entre as UFs: subiu em todo o Norte, caiu no Sudeste e no Nordeste.
- **Pergunta:** a abstenção do 2º turno em cada UF está associada ao apoio a Lula no 1º turno, depois de controlar pela região?
- A variável resposta é uma proporção $y_i \in (0, 1)$:
  - MQO sobre y_i ignora os limites e a heterocedasticidade típica de proporções;
  - logit/probit para dados agregados exigiria trabalhar com as contagens;
  - **regressão beta** (Ferrari & Cribari-Neto, 2004) modela y_i diretamente.
- $n = 27$ UFs $\Rightarrow$ amostra pequena: um bom caso para discutir o papel das prioris.

Dados (aba `Por UF`)

- y_i = % de abstenção no 2º turno na UF i.
- x_i = % de votos em Lula no 1º turno, sobre o comparecimento:

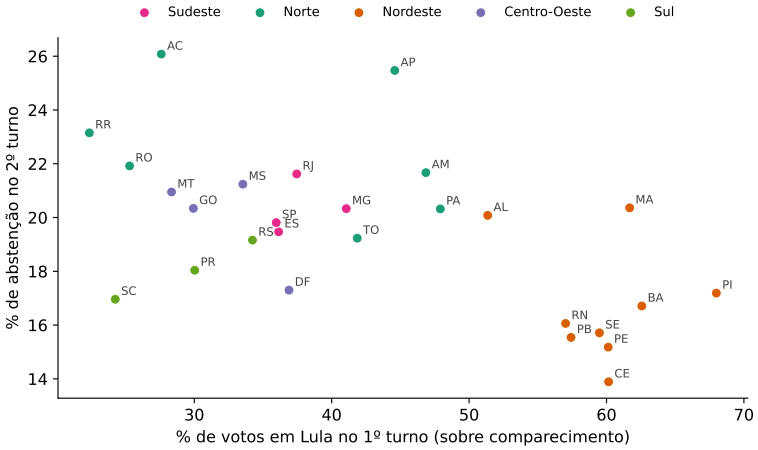
$$x_i = 100 \times \frac{\text{votos Lula}_{1T}}{\text{aptos} - \text{abstenção}_{1T}}$$

  (inclui brancos e nulos no denominador; os votos válidos não estão na planilha).
- Dummies regionais: Norte, Nordeste, Centro-Oeste, Sul; **Sudeste = base**.

Cuidados com os dados

- Aptos e abstenção do 1º turno: TSE.
- Abstenção do 2º turno: estimada como taxa do 1º turno + variação em p.p. publicada pela imprensa. Confere com as taxas divulgadas para CE, PE, SP, BA, AC, AP, RR e RO.
- Votos de Lula: recorte do TSE de 25/10/2026, 23h08 (BA com 99,99% das seções).

Os dados



Regiões quase separadas em x: o Nordeste fica todo à direita e o Sul à esquerda. $\Rightarrow \beta_1$ e as dummies competem pela mesma variação.

Regressão beta (Ferrari & Cribari-Neto, 2004)

Reparametrização da Beta(a, b) por média μ e precisão ϕ: $a = \mu\phi$, $b = (1-\mu)\phi$,

$$f(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\,\Gamma((1-\mu)\phi)}\, y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1,$$

$$E(y) = \mu, \qquad V(y) = \frac{\mu(1-\mu)}{1+\phi}.$$

- A variância depende da média: heterocedasticidade "de graça".
- ϕ grande $\Rightarrow$ y concentrado em torno de μ.
- Função de ligação logit: $\mathrm{logit}(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$.
- No artigo original a estimação é por máxima verossimilhança (pacote `betareg` no R). Aqui: **inferência bayesiana via MCMC**.

Especificação

$$y_i \mid \mu_i, \phi \sim \text{Beta}\big(\mu_i\phi,\ (1-\mu_i)\phi\big), \qquad i = 1, \ldots, 27$$

$$\text{logit}(\mu_i) = \beta_0 + \beta_1(x_i - \bar{x}) + \gamma_N N_i + \gamma_{NE}\, NE_i + \gamma_{CO}\, CO_i + \gamma_S S_i$$

- $\bar{x} = 43{,}0$ p.p. (média das 27 UFs). Centrar x dá a β_0 uma interpretação:

$$\text{expit}(\beta_0) = \text{abstenção esperada no Sudeste com } x = \bar{x}.$$

- É isso que permite pôr uma priori substantiva no intercepto ("em torno de 20%").
- Sem centrar, β_0 seria a abstenção numa UF com 0% de votos em Lula: sem sentido.
- Atenção: no Sudeste x varia de 36% a 41%, então $x = \bar{x}$ já é uma pequena extrapolação para essa região.

Prioris

Parâmetro	Priori	Comentário
β_0	$\text{expit}(\beta_0) \sim N(0{,}20;\ 0{,}03^2)$	Gaussiana na escala de probabilidade
β_1	$N(0, 100^2)$	vaga
$\gamma_N, \gamma_{NE}, \gamma_{CO}, \gamma_S$	$N(0, 5^2)$	escala logit
ϕ	Gamma$(0{,}01;\ 0{,}01)$	vaga (escolha nossa)

- Priori induzida em β_0 (jacobiano):
  $p(\beta_0) = N\big(\text{expit}(\beta_0); 0{,}2; 0{,}03^2\big)\ \text{expit}(\beta_0)\{1 - \text{expit}(\beta_0)\}$.
- Pelo método delta: $\beta_0 \approx N\big(\text{logit}\, 0{,}2;\ (0{,}03/0{,}16)^2\big) = N(-1{,}39;\ 0{,}19^2)$. Por simulação: média $-1{,}40$, dp $0{,}19$.
- $N(0, 5^2)$ na escala logit é praticamente plana: ± 10 no logit corresponde a quase todo o intervalo $(0, 1)$.
- O enunciado não especifica a priori de ϕ: fica como ponto para discussão.

Posteriori e algoritmo

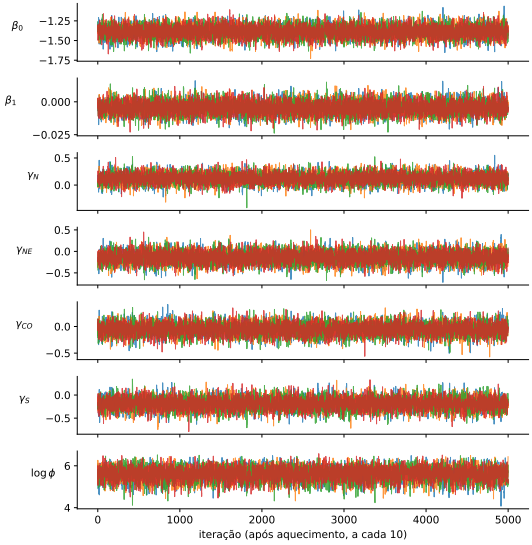
$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \;\propto\; \prod_{i=1}^{27} f(y_i \mid \mu_i(\boldsymbol{\beta}), \phi)\; p(\boldsymbol{\beta})\, p(\phi), \qquad \boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}, \log \phi)$$

Metropolis passeio aleatório em $\boldsymbol{\theta} \in \mathbb{R}^7$:

1. Calcule a moda a posteriori $\hat{\boldsymbol{\theta}}$ e a hessiana $\boldsymbol{H}$ (aproximação de Laplace).
2. Proposta: $\boldsymbol{\theta}^\star = \boldsymbol{\theta}^{(t)} + \boldsymbol{\epsilon}$, com $\boldsymbol{\epsilon} \sim N\big(\mathbf{0},\ \frac{2{,}38^2}{7}\boldsymbol{H}^{-1}\big)$ (Roberts, Gelman & Gilks, 1997).
3. Aceite com probabilidade $\min\{1,\ p(\boldsymbol{\theta}^\star \mid \boldsymbol{y}) / p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})\}$.

- Amostrar $\log \phi$ (e não ϕ) evita a restrição $\phi > 0$; inclui-se o jacobiano ϕ.
- 4 cadeias com inícios dispersos, 60.000 iterações cada, 10.000 de aquecimento $\Rightarrow$ 200.000 draws.
- Taxa de aceitação $\approx 0{,}33$ em todas as cadeias.
- $\widehat{R}$ (*split*) $\leq 1{,}001$ e ESS entre 6.400 e 8.500 para todos os parâmetros.

Traços (4 cadeias)

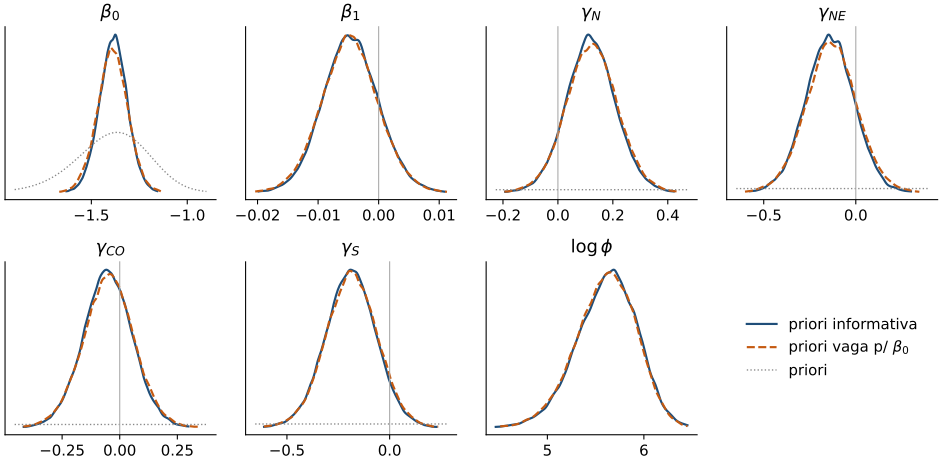


Estimativas: MV $\times$ posteriori

	Máx. verossimilhança		Posteriori (priori informativa)				
Parâmetro	Est.	EP	Média	DP	2,5%	97,5%	$P(>0 \mid y)$
β_0	$-1{,}390$	0,059	$-1{,}386$	0,074	$-1{,}533$	$-1{,}242$	0,00
β_1	$-0{,}005$	0,003	$-0{,}005$	0,005	$-0{,}014$	$0{,}005$	0,16
γ_N	$0{,}124$	0,069	$0{,}120$	0,091	$-0{,}059$	$0{,}303$	0,91
γ_{NE}	$-0{,}138$	0,102	$-0{,}141$	0,134	$-0{,}407$	$0{,}119$	0,14
γ_{CO}	$-0{,}049$	0,081	$-0{,}052$	0,108	$-0{,}267$	$0{,}163$	0,31
γ_S	$-0{,}184$	0,092	$-0{,}189$	0,126	$-0{,}438$	$0{,}056$	0,06
$\log \phi$	$6{,}21$	0,27	$5{,}61$	0,31	$4{,}95$	$6{,}18$	–

- Médias a posteriori quase iguais às de MV para $\boldsymbol{\beta}$, mas os desvios-padrão são 25–40% maiores.
- Motivo: o EMV de ϕ é viesado para cima em amostras pequenas ($\hat{\phi}_{MV} \approx 500$ contra mediana a posteriori de 279). Com ϕ menor, a incerteza sobre $\boldsymbol{\beta}$ aumenta. Esse viés é conhecido na literatura de regressão beta (Ospina, Cribari-Neto & Vasconcellos, 2006).

Posterioris marginais



Linha contínua: priori $\text{expit}(\beta_0) \sim N(0{,}2; 0{,}03^2)$. Linha tracejada: $\beta_0 \sim N(0, 100^2)$. Pontilhada: prioris. Os dados dominam a priori do intercepto, porque 0,20 já é a média observada.

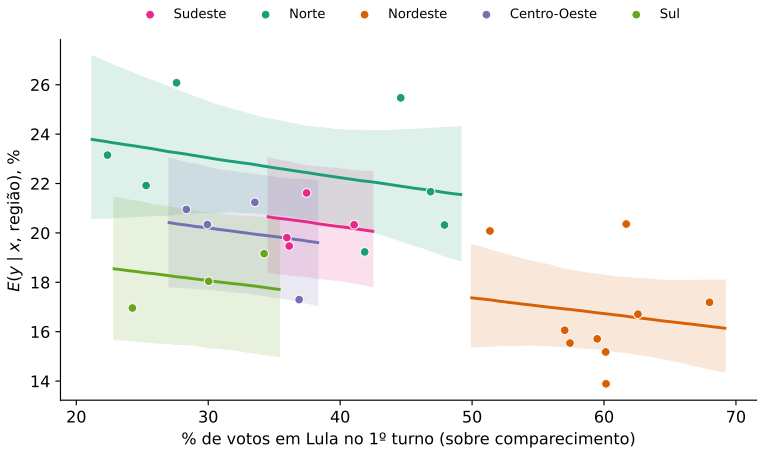
Na escala que interessa: pontos percentuais

Quantidades derivadas, calculadas draw a draw (mediana e IC 95%):

Quantidade	p.p.
Abstenção esperada no Sudeste, $x = \bar{x}$: $100\ \mathrm{expit}(\beta_0)$	20,0 (17,8; 22,4)
Norte − Sudeste	$+2{,}0$ ($-1{,}0$; 5,0)
Nordeste − Sudeste	$-2{,}2$ ($-6{,}2$; 1,9)
Centro-Oeste − Sudeste	$-0{,}8$ ($-4{,}1$; 2,6)
Sul − Sudeste	$-2{,}9$ ($-6{,}4$; 0,9)
$+10$ p.p. de voto em Lula (Sudeste)	$-0{,}7$ ($-2{,}1$; 0,8)

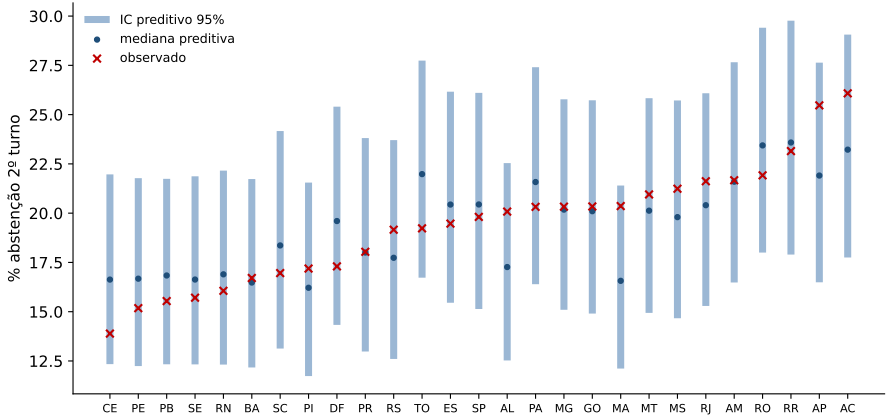
- Nenhum intervalo de 95% exclui zero, mas $P(\gamma_N > 0 \mid y) = 0{,}91$ e $P(\gamma_S < 0 \mid y) = 0{,}94$.
- Uma vez controlada a região, o voto em Lula no 1º turno pouco explica a abstenção no 2º turno.

Curvas ajustadas por região (mediana e banda de 95%)



Inclinações paralelas na escala logit (modelo sem interação). As faixas de x quase não se sobrepõem entre regiões, o que limita a identificação de β_1 separadamente das dummies.

Checagem preditiva a posteriori



$y_i^{rep} \sim \text{Beta}(\mu_i\phi, (1-\mu_i)\phi)$ para cada draw. As 27 UFs ficam dentro do intervalo preditivo de 95%. CE (abaixo) e AP, AC (acima) estão perto das bordas: são as UFs com histórias específicas.

Robustez: x = votos em Lula / aptos

Usando a coluna `% Lula/aptos 1º T` da planilha:

	x sobre comparecimento			x sobre aptos		
Parâmetro	Média	IC 95%	$P(>0)$	Média	IC 95%	$P(>0)$
β_0	−1,386	(−1,533; −1,242)	0,00	−1,395	(−1,545; −1,246)	0,00
β_1	−0,005	(−0,014; 0,005)	0,16	−0,007	(−0,018; 0,004)	0,11
γ_N	0,120	(−0,059; 0,303)	0,91	0,125	(−0,052; 0,301)	0,92
γ_{NE}	−0,141	(−0,407; 0,119)	0,14	−0,111	(−0,384; 0,164)	0,21
γ_{CO}	−0,052	(−0,267; 0,163)	0,31	−0,056	(−0,269; 0,153)	0,30
γ_S	−0,189	(−0,438; 0,056)	0,06	−0,194	(−0,436; 0,042)	0,05

- Aqui $x_i = (\text{Lula/comparecimento}) \times (1 - \text{abst}_{1T})$: a abstenção do 1º turno entra mecanicamente no regressor e, em princípio, β_1 poderia captar persistência da abstenção.
- Na prática as conclusões quase não mudam: β_1 fica um pouco mais negativo ($P(\beta_1 > 0 \mid y)$ cai de 0,16 para 0,11), e as dummies ficam praticamente iguais.
- Lição: definir o denominador faz parte da especificação, e vale a pena verificar.

Para discutir em aula

1. **Falácia ecológica**: a associação entre UFs não diz se eleitores de Lula se abstêm mais ou menos.
2. **Papel da priori**: a priori do intercepto quase não muda nada aqui. Por quê? Experimente $\text{expit}(\beta_0) \sim N(0{,}10;\, 0{,}01^2)$.
3. **Centragem**: a mesma priori "20% $\pm$ 3%" com x não centrado seria uma afirmação sobre uma UF com 0% de votos em Lula.
4. ϕ **e incerteza**: por que os EPs de MV são otimistas com $n = 27$? Compare com a correção de viés de Ospina et al. (2006) ou com o `betareg(type="BC")`.
5. **Extensões**: precisão variável $\log \phi_i = \boldsymbol{z}_i'\boldsymbol{\delta}$; interação $x \times$região; modelo hierárquico para os efeitos regionais; abstenção do 1º turno como covariável.
6. **Pesos**: SP tem 34 milhões de aptos e RR tem 0,4 milhão. Deveriam pesar igual? Um binomial com sobredispersão (beta-binomial) usaria as contagens.
7. **Dados**: a abstenção do 2º turno é estimada. Como esse erro de medição em y afeta a inferência?

Código: verossimilhança e priori (Python)

```python
def loglik(th):
    beta, phi = th[:p], np.exp(th[p])
    mu = expit(X @ beta)
    a, b = mu * phi, (1 - mu) * phi
    return np.sum(gammaln(phi) - gammaln(a) - gammaln(b)
                  + (a - 1) * np.log(y) + (b - 1) * np.log1p(-y))

def logprior(th):
    b0, b1, g, lphi = th[0], th[1], th[2:p], th[p]
    m  = expit(b0)                                  # expit(b0) ~ N(.20, .03^2)
    lp = stats.norm.logpdf(m, 0.20, 0.03) + np.log(m * (1 - m))   # + jacobiano
    lp += stats.norm.logpdf(b1, 0, 100)             # vaga
    lp += np.sum(stats.norm.logpdf(g, 0, 5))        # dummies regionais
    lp += stats.gamma.logpdf(np.exp(lphi), a=0.01, scale=100) + lphi
    return lp

# Metropolis: prop = th + L @ z,  L = chol(2.38^2/7 * inv(H)),  aceita se log(u) < lpp - lp
```

Script completo: `beta_reg_mcmc.py` (MV, MCMC, $\widehat{R}$, ESS) e `figures.py`.

Referências

- Ferrari, S. L. P. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815.
- Cribari-Neto, F. & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34(2).
- Ospina, R., Cribari-Neto, F. & Vasconcellos, K. L. P. (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics & Data Analysis*, 51(2), 960–981.
- Branscum, A. J., Johnson, W. O. & Thurmond, M. C. (2007). Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, 49(3), 287–301.
- Roberts, G. O., Gelman, A. & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1), 110–120.
- Gamerman, D. & Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd ed. Chapman & Hall/CRC.

Apêndice: o prompt usado para gerar esta análise

Na aba "Por UF", faça uma regressão beta (como em Ferrari and Cribari-Neto) que explica o percentual de abstenções do 2º turno como função do percentual de votos em Lula no 1º turno + dummies para as regiões do país (usando o SUDESTE como baseline). Para o intercepto da regressão beta use uma priori Gaussiana centrada em 20% e desvio padrão igual a 3%. Para as dummies use normais com média 0 e desvio-padrão 5. Finalmente, para o percentual de votos em Lula no 1º turno use uma priori vaga. Obtenha as posterioris dos parâmetros do modelo via MCMC e prepare slides em latex/beamer [...]

- Decisões que o prompt deixou em aberto e foram tomadas na análise: denominador de x (comparecimento), centragem de x, escala da priori do intercepto (probabilidade), priori de ϕ, algoritmo de MCMC.
- Exercício: quais dessas escolhas mudam as conclusões?