

Modelos Poisson com quebras estruturais

Mortes de motociclistas em São Paulo entre 2015 e 2026

Hedibert Freitas Lopes

Outubro de 2026

Este material foi produzido em colaboração com o Claude (Anthropic): os prompts, as escolhas de modelagem e a revisão final são do autor, que responde integralmente pelo conteúdo. Veja o prompt utilizado nos slides 1-3.

Insper — Doutorado em Economia dos Negócios
Advanced Bayesian Econometrics

O pedido feito ao Claude

O prompt (1/2): contexto, dados e modelos

Contexto. Vamos continuar com os dados de mortes mensais de condutores de motocicletas no município de São Paulo (jan/2015 a ago/2026, $n = 140$ meses), agora pensando em variações do modelo Poisson.

Dados. 21 28 33 38 35 39 27 24 ... 43 50 48 47 (os 140 valores, colados no prompt)

Modelos a contemplar.

- **M0:** $y_t \stackrel{iid}{\sim} \text{Poi}(\lambda)$ para a amostra toda.
- **M1:** $\text{Poi}(\lambda_1)$ para $t = 1, \dots, \tau_1$; $\text{Poi}(\lambda_2)$ para $t = \tau_1 + 1, \dots, \tau_2$; $\text{Poi}(\lambda_3)$ para $t = \tau_2 + 1, \dots, n$. $\tau_1 \in \{65, \dots, 95\}$, $\tau_2 \in \{85, \dots, 115\}$, com $\tau_2 > \tau_1$.
- **M2:** $y_t \sim \text{Poi}(\lambda(t))$, com $\log \lambda(t) = \alpha_j + \beta_j t$ no intervalo $j = 1, 2, 3$, ou seja, M1 expandido para retas com inclinação dentro de cada intervalo. Provavelmente β_1 (e talvez β_2) será predominantemente zero.

O prompt (2/2): prioris, inferência e entregável

Prioris.

- $\lambda, \lambda_1, \lambda_2, \lambda_3$: gamas com modas em 20, 30 e 50 e intervalos de credibilidade de 95% aproximadamente (10, 30), (20, 40) e (40, 60).
- (τ_1, τ_2) : uniforme discreta na região $\tau_1 + 10 < \tau_2$ (garante ao menos 10 observações no intervalo do meio).
- $\alpha_1, \alpha_2, \alpha_3$: comparáveis às prioris de $\lambda_1, \lambda_2, \lambda_3$.
- $\beta_1, \beta_2, \beta_3$: centradas em zero, desvio-padrão 10.

Inferência. Sumarize as posteriores dos parâmetros de cada modelo, compare os que são comparáveis e termine comparando M0, M1 e M2 via **fator de Bayes**. Use *sampling importance resampling*, Gibbs ou passos de Metropolis–Hastings quando necessário.

Entregável. Slides em $\text{\LaTeX}$ /beamer, editáveis. No início, inclua o próprio prompt em formato legível, para que os alunos aprendam a informar ao Claude o que realmente querem.

O que torna esse prompt eficaz?

Ingredientes presentes

- **Dados** colados e descritos (unidade, período, n).
- **Modelos** escritos em notação, sem ambiguidade de família.
- **Prioris** elicítadas de forma operacional: moda + IC 95%.
- **Suporte** das variáveis discretas e restrições explícitas.
- **Métodos** aceitáveis listados (SIR, Gibbs, MH).
- **Formato** da entrega (beamer editável) e público-alvo.

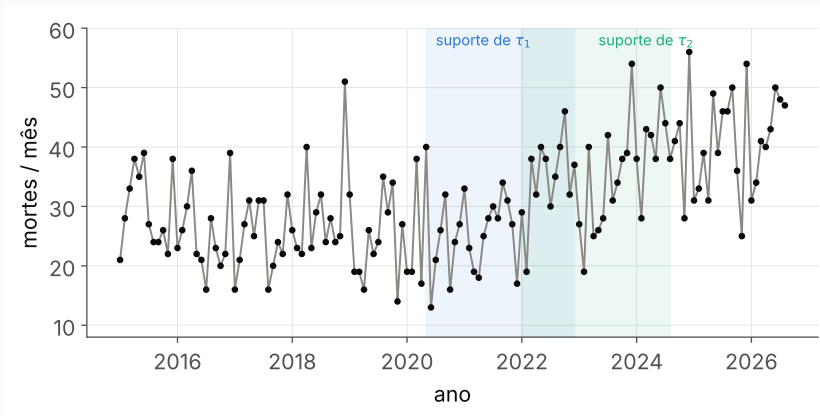
Lição: quanto mais explícito o prompt, menos decisões silenciosas. As que restarem devem ser **declaradas** na resposta.

O que o Claude teve de decidir

- “ $\text{Poi}(\lambda_1)$ ” repetido no 2º intervalo $\Rightarrow$ lido como λ_2 .
- Priori de λ em M0 não especificada $\Rightarrow$ usada a do meio (moda 30).
- t é centrado no ponto médio de cada segmento, para que e^{α_j} seja comparável a λ_j .

Dados e prioris

Os dados



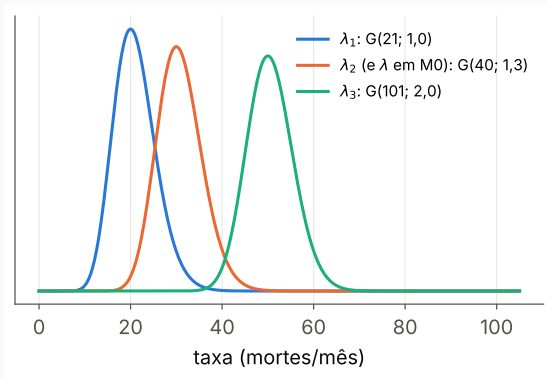
$\bar{y} = 30,9$ e $s^2 = 92,3$: forte **sobredispersão** frente a um Poisson único ($E = V$). As faixas mostram os suportes a priori de τ_1 (65–95, ago/2020–nov/2022) e τ_2 (85–115, jan/2022–jul/2024).

Elicitação das prioris gama: moda + IC 95%

$\lambda \sim \mathcal{G}(a, b)$: moda = $(a - 1)/b$. Fixada a moda, escolhemos a para aproximar o IC 95% pedido.

	$\mathcal{G}(a, b)$	moda	IC 95%
λ_1	$\mathcal{G}(21; 1,0)$	20	(13,0; 30,9)
λ_2, λ	$\mathcal{G}(40; 1,3)$	30	(22,0; 41,0)
λ_3	$\mathcal{G}(101; 2,0)$	50	(41,1; 60,8)

Em M2: $\alpha_j = \log \lambda_j^*$ com λ_j^* seguindo a mesma gama (priori log-gama). $\beta_j \sim N(0, 10^2)$.



M0 e M1

M0: Poisson–gama conjugado

$$y_t \mid \lambda \stackrel{iid}{\sim} \text{Poi}(\lambda), \quad \lambda \sim \mathcal{G}(a, b) \implies \lambda \mid y \sim \mathcal{G}\left(a + \sum_t y_t, b + n\right) = \mathcal{G}(4368; 141,3).$$

$$\lambda \mid y : \quad \text{média } 30,91, \quad \text{dp } 0,47, \quad \text{IC } 95\% (30,0; 31,8).$$

Verossimilhança marginal exata:

$$p(y \mid M_0) = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + \sum y_t)}{(b + n)^{a + \sum y_t}} \prod_t \frac{1}{y_t!}, \quad \log p(y \mid M_0) = -573,86.$$

Posteriori muito concentrada, mas o modelo ignora a tendência e a sobredispersão evidentes nos dados.

M1: duas quebras — inferência exata por enumeração

Dado (τ_1, τ_2) , os três segmentos são Poisson–gama independentes:

$$p(y \mid \tau_1, \tau_2, M_1) = \prod_{j=1}^3 \frac{b_j^{a_j}}{\Gamma(a_j)} \frac{\Gamma(a_j + S_j)}{(b_j + m_j)^{a_j + S_j}} \prod_{t \in \mathcal{I}_j} \frac{1}{y_t!}, \quad S_j = \sum_{t \in \mathcal{I}_j} y_t, \quad m_j = |\mathcal{I}_j|.$$

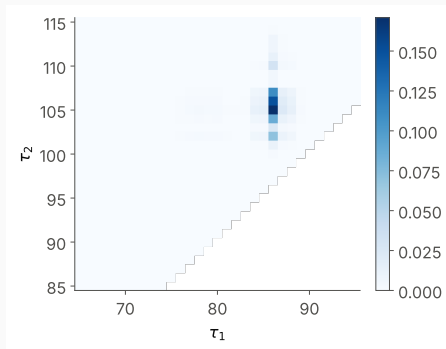
A grade $\{65 \leq \tau_1 \leq 95, 85 \leq \tau_2 \leq 115, \tau_2 > \tau_1 + 10\}$ tem $K = 730$ pares, então

$$p(\tau_1, \tau_2 \mid y) \propto p(y \mid \tau_1, \tau_2), \quad p(y \mid M_1) = \frac{1}{K} \sum_{(\tau_1, \tau_2)} p(y \mid \tau_1, \tau_2) \quad (\text{exatos}).$$

Amostras de $(\tau, \boldsymbol{\lambda}) \mid y$ por composição: $\tau \sim p(\tau \mid y)$, depois $\lambda_j \mid \tau, y \sim \mathcal{G}(a_j + S_j, b_j + m_j)$.

$$\log p(y \mid M_1) = -497,45.$$

M1: posteriori de (τ_1, τ_2)



	mediana	IC 95%
τ_1	86 (fev/2022)	[79, 88]
τ_2	105 (set/2023)	[102, 111]

Pares mais prováveis: (86, 105) 0,17; (86, 106) 0,15; (86, 107) 0,11.

A 1ª quebra é bem localizada: o regime de ≈ 26 mortes/mês termina em fev/2022. A 2ª é mais difusa, entre jun/2023 e mar/2024.

M1: posteriori das taxas

	média	dp	IC 95%
λ_1	26,18	0,55	(25,1; 27,3)
λ_2	33,48	1,46	(30,6; 36,3)
λ_3	41,44	1,09	(39,3; 43,6)
λ_2/λ_1	1,28	0,06	(1,16; 1,40)
λ_3/λ_2	1,24	0,06	(1,12; 1,37)

Cada quebra eleva a taxa em cerca de 20 a 30%, e nenhum IC das razões contém 1. Entre o 1º e o 3º regime, a taxa sobe $\approx 58\%$.

M1 via amostrador de Gibbs: checagem contra o exato

Condicionais completas

- $\lambda_j \mid \tau, y \sim \mathcal{G}(a_j + S_j, b_j + m_j)$
- $(\tau_1, \tau_2) \mid \lambda, y$: discreta nos 730 pares, $\propto \prod_j \lambda_j^{S_j} e^{-m_j \lambda_j}$ (bloco)

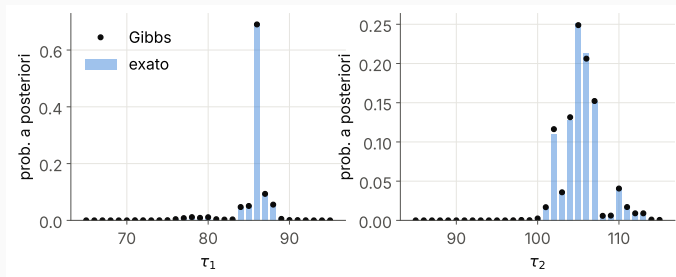
10.000 iterações após
aquecimento de 2.000.

$E(\tau_1 \mid y)$: 85,69 (Gibbs) vs 85,71
(exato)

$E(\tau_2 \mid y)$: 105,37 vs 105,38

$E(\lambda_2 \mid y)$: 33,46 vs 33,48

Distância de variação total: 0,03



M2: log-taxas lineares por segmento

M2: especificação

$$y_t \sim \text{Poi}(\lambda(t)), \quad \log \lambda(t) = \alpha_j + \beta_j (t - c_j), \quad t \in \mathcal{I}_j,$$

com c_j o **ponto médio** do segmento j . Assim, e^{α_j} é a taxa no meio do segmento, comparável a λ_j de M1, e α_j e β_j ficam quase ortogonais a posteriori.

Prioris: $e^{\alpha_j} \sim \mathcal{G}(a_j, b_j)$ (as mesmas de M1), $\beta_j \sim N(0, 10^2)$ e (τ_1, τ_2) uniforme na grade.

Por que a escala importa: β_j é a variação de $\log \lambda$ por mês. Assim, $100(e^{12\beta_j} - 1)$ é a variação percentual ao ano. Um dp de 10 equivale a admitir a priori que a taxa seja multiplicada por $e^{10} \approx 22.000$ **a cada mês**.

M2: algoritmo (Laplace + IS + SIR)

Dado (τ_1, τ_2) , os segmentos são independentes. Para cada segmento (s, e) distinto (792 no total):

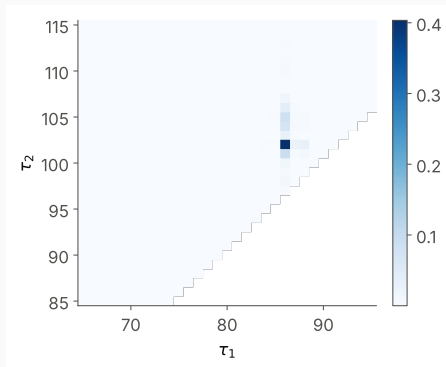
1. **Laplace:** moda $\hat{\theta}$ de $\log p(y_{s:e} | \theta)p(\theta)$ por Newton (função côncava) e $V = -H^{-1}$.
2. **Amostragem por importância:** $\theta^{(i)} \sim t_5(\hat{\theta}, 1, 2^2 V)$, com $i = 1, \dots, 6000$, e

$$\hat{p}(y_{s:e}) = \frac{1}{N} \sum_i \frac{p(y_{s:e} | \theta^{(i)})p(\theta^{(i)})}{q(\theta^{(i)})}.$$

3. $p(y | \tau, M_2) = \prod_j \hat{p}(y_{\mathcal{I}_j})$, depois $p(\tau | y, M_2)$ e $p(y | M_2)$ por enumeração (como em M1).
4. **SIR:** $\tau \sim p(\tau | y, M_2)$; (α_j, β_j) reamostrados dos $\theta^{(i)}$ com pesos w_i .

Diagnósticos: tamanho efetivo de amostra mínimo de 4.854 em 6.000. $\log p(y | M_2)$ vale $-519,02$ por IS e $-519,03$ por Laplace; 5 réplicas de IS variam em $\pm 0,01$.

M2: posteriori de (τ_1, τ_2)

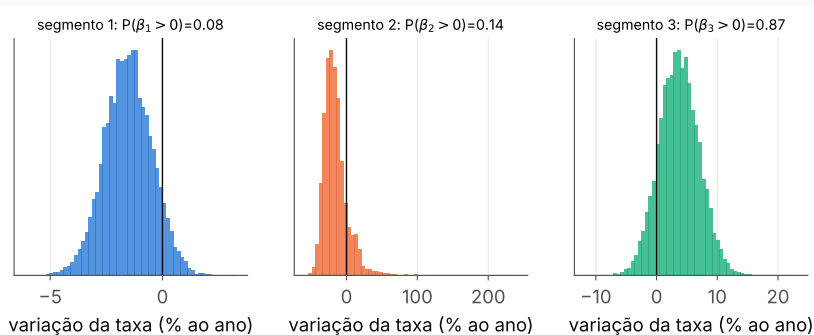


	mediana	IC 95%
τ_1	86 (fev/2022)	[71, 88]
τ_2	102 (jun/2023)	[94, 111]

Par modal (86, 102) com probabilidade 0,40.

A inclinação dentro dos segmentos “absorve” parte da mudança de nível, por isso os τ ficam mais incertos que em M1, sobretudo τ_2 .

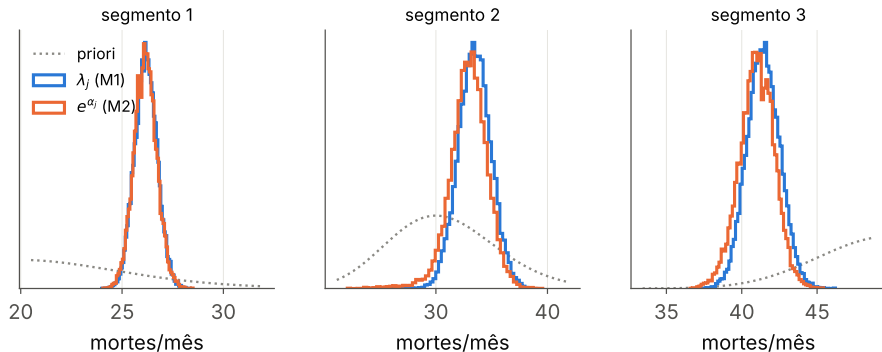
M2: posteriori das inclinações



	β_j (média)	variação % a.a. (IC 95%)	$P(\beta_j > 0 \mid y)$
segmento 1	-0,0013	-1,5 (-3,6; 0,5)	0,08
segmento 2	-0,0154	-15 (-39; 26)	0,14
segmento 3	0,0030	3,7 (-2,7; 10,3)	0,87

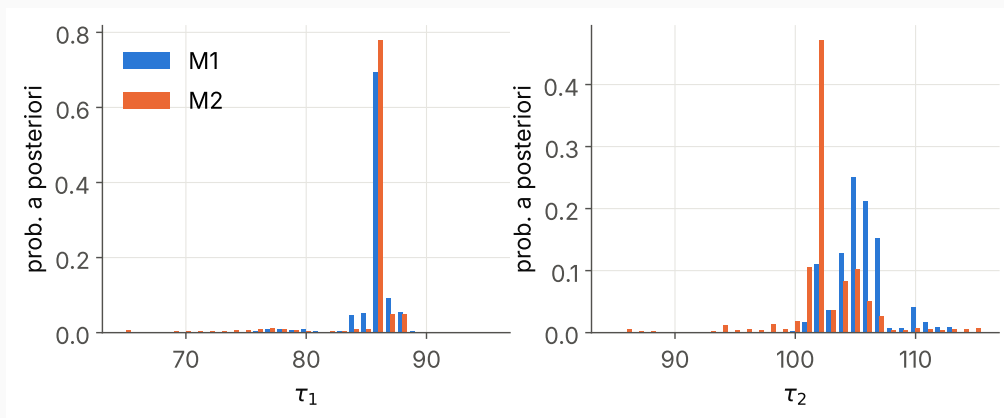
Comparações

Parâmetros comparáveis: λ_j (M1) vs e^{α_j} (M2)



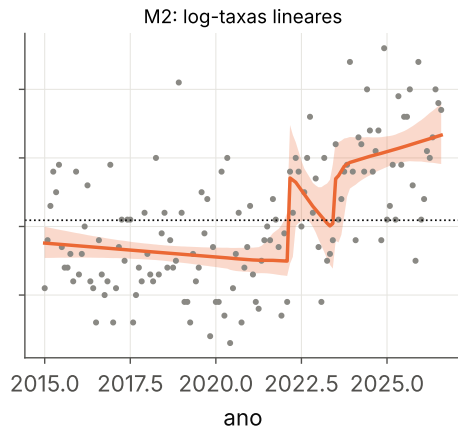
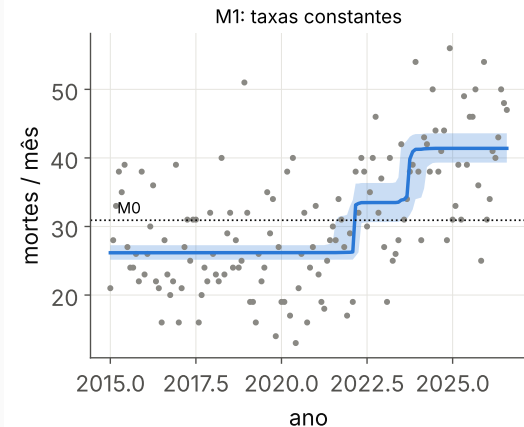
	segmento 1	segmento 2	segmento 3
λ_j (M1)	26,2 (25,1; 27,3)	33,5 (30,6; 36,3)	41,4 (39,3; 43,6)
e^{α_j} (M2)	26,2 (25,1; 27,2)	32,9 (29,4; 36,0)	41,0 (38,6; 43,2)

Pontos de quebra: M1 vs M2



Os dois modelos concordam em $\tau_1 = 86$ (fev/2022). M2 desloca τ_2 para mais cedo e o torna mais difuso.

Taxas ajustadas (mediana e banda de 95%)



A linha pontilhada é M0 ($\hat{\lambda} \approx 30,9$).

Seleção de modelos

Fatores de Bayes

	$\log p(y \mid M_k)$	$P(M_k \mid y)$
M0: Poisson único	-573,86	$\approx 10^{-33}$
M1: 3 regimes constantes	-497,45	≈ 1
M2: 3 regimes com tendência	-519,02	$\approx 4 \times 10^{-10}$

	$\log \text{FB}$	$2 \log \text{FB}$
FB_{10}	76,4	152,8
FB_{21}	-21,6	-43,1
FB_{20}	54,8	109,7

Prioris iguais nos modelos. Na escala de Kass & Raftery (1995), $2 \log \text{FB} > 10$ é evidência “muito forte”.

Sensibilidade: o paradoxo de Lindley em ação

dp de β_j	$\log p(y \mid M_2)$	$\log \text{FB}_{21}$
10	-519,0	-21,6
1	-512,1	-14,7
0,1	-505,2	-7,8
0,01	-499,5	-2,1

- Dividir o dp por 10 soma $\approx \log 10 = 2,3$ **por segmento** a $\log p(y \mid M_2)$, ou seja, $\approx 6,9$ no total. É o fator de Occam: $\text{dp}_{\text{post}}/\text{dp}_{\text{priori}}$.
- Prioris vagas para parâmetros extras penalizam o modelo maior, mesmo que a posteriori quase não mude.
- Mesmo com $\text{dp} = 0,01$ ($\approx \pm 12\%$ a.a.), M1 continua preferido.
- M0 é descartado sob qualquer especificação.

- Há evidência esmagadora de **mudança de regime**: M1 e M2 dominam M0 ($\log FB > 54$).
- M1 é o modelo preferido. Há três patamares, $\approx 26 \rightarrow 33 \rightarrow 41$ mortes/mês, com quebras em fev/2022 e por volta de set/2023.
- Dentro dos segmentos, as inclinações são pequenas ou mal identificadas. A tendência visível na série é capturada por *saltos* de nível, não por retas.
- O fator de Bayes é sensível à priori de parâmetros adicionais. Elicite β_j na escala certa (% ao ano), não com “dp = 10” por hábito.
- Ainda há sobredispersão dentro dos regimes. Próximos passos: Poisson–gama (binomial negativa), sazonalidade e número de quebras desconhecido.

Apêndice: reprodutibilidade

Todo o cálculo está em `analise_poisson_quebras.py` (Python 3 com `numpy`, `scipy` e `matplotlib`), com semente fixa. O script gera as figuras em `figs/` e os números em `resultados.json`.

```
# log p(y_seg) Poisson-gama (M1, exato)
def logml_pg(S, m, L, a, b):
    return a*log(b) - gammaln(a) + gammaln(a+S) - (a+S)*log(b+m) - L

# M1: soma sobre os 730 pares (tau1, tau2) da grade
l1 = [logml_pg(seg1) + logml_pg(seg2) + logml_pg(seg3) for (tau1, tau2) in grid]
log_py_M1 = logsumexp(l1) - log(K)
```

Tempo total de execução: cerca de 1 minuto num laptop.