Relatório - Inferência Bayseana

Luca Adriano Melo Mendonça Soares

Novembro 2025

1 Introdução

Figure 1: Logo ENEM

A base de resultados do ENEM, possui um valor rico quando se trata de índice de desempenho por dentre as regiões no vestibular. Também traz algumas informacões sociais que são interessantes na análise como um todo. No discorrer desse trabalho, será feito um modelo bayseano da família das exponenciais para predizer se a escola a qual o candidato estudou é pública ou privada com base nas notas do Vestibular ENEM.

2 Explicar o algoritmo das funções utilizadas na análise bayesiana escolhida

Para realizar a inferência estatística deste projeto, utilizou-se a biblioteca **Bambi** (*Bayesian Model-Building Interface*), construída sobre o **PyMC**.

Abaixo, detalha-se o funcionamento dos algoritmos envolvidos:

2.1 O Modelo Linear Generalizado (GLM) Bayesiano

A estrutura matemática escolhida pertence à **Família Exponencial**, especificamente a distribuição **Normal (Gaussiana)**, adequada para variáveis contínuas como as notas do ENEM. A função de ligação (*Link Function*) utilizada foi a **Identidade**.

Matematicamente, o modelo é definido pelas seguintes equações:

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2) \tag{1}$$

$$\mu_i = \alpha + \beta \cdot X_i \tag{2}$$

Onde y representa a nota, α é o intercepto (média estimada da escola pública), β é o coeficiente de efeito da escola privada, e σ é o desvio padrão dos resíduos (variabilidade não explicada pelo modelo).

2.2 Definição e Justificativa das Distribuições a Priori

Para este estudo, optou-se por **Priors Fracamente Informativas**, configuradas manualmente para respeitar a escala de pontuação do ENEM (0 a 1000 pontos).

As definições foram:

- **Para o Intercepto ($\alpha \sim \mathcal{N}(500, 200)$):** O intercepto representa a nota base (escola pública).
  - A média $\mu = 500$ foi escolhida, pois nos anos anteriores a média se concentrou perto desse valor.
  - O desvio padrão $\sigma = 200$ diz respeito a flexibilidade do modelo, permitindo que a média da escola pública varie amplamente, cobrindo todos os tipos de cenários sem restringir excessivamente os dados.
- **Para o Coeficiente Beta ($\beta \sim \mathcal{N}(50, 20)$):** Este parâmetro representa a diferença de nota causada pelo tipo de escola, foi usado como exemplo a nota de redação, que teve a maior diferença de média entre escolas públicas e privadas.
  - A média $\mu = 50$ reflete uma postura que presume que a diferença média é de 50 pontos, onde assumimos a priori que há uma "pequena" diferença de desempenho até que os dados provem o contrário.
  - O desvio padrão $\sigma = 20$ atua como um mecanismo de **regularização**.

2.3 O Algoritmo MCMC (Markov Chain Monte Carlo)

Como a resolução analítica da integral do Teorema de Bayes é computacionalmente inviável para grandes volumes de dados, utilizamos métodos numéricos de aproximação denominados **MCMC**. O objetivo do algoritmo é "caminhar" pelo espaço de parâmetros e coletar amostras que representem a verdadeira distribuição de probabilidade dos coeficientes.

2.4 O Amostrador NUTS (No-U-Turn Sampler)

Especificamente, o algoritmo utilizado foi o **NUTS**, uma extensão moderna do *Hamiltonian Monte Carlo* (HMC).

- **Funcionamento:** Ao contrário de algoritmos clássicos (como Metropolis-Hastings) que realizam passos aleatórios, o NUTS utiliza a geometria da distribuição (cálculo de gradientes) para propor passos mais eficientes.
- **Eficiência:** O algoritmo simula uma trajetória física hamiltoniana e impede que o processo retorne ao ponto de origem (*No-U-Turn*), garantindo uma exploração do espaço de probabilidade muito mais rápida e com menor número de amostras, fator crucial dada a dimensão da base de dados do ENEM.

3 Avaliar o modelo e interpretar os resultados das estimativas

A avaliação da qualidade do ajuste do modelo foi realizada através de diagnósticos de convergência e análise visual das distribuições a posteriori.

3.1 Convergência das Cadeias (R-hat)

```
--- DIAGNÓSTICO DE CONVERGÊNCIA (R-HAT) ---
Valores ideais devem ser iguais a 1.0 (ou < 1.05)

> Disciplina: NATUREZA
                r_hat
escola_privada    1.0
Intercept         1.0
------------------------------
> Disciplina: HUMANAS
                r_hat
escola_privada    1.0
Intercept         1.0
------------------------------
> Disciplina: LINGUAGENS
                r_hat
escola_privada    1.0
Intercept         1.0
------------------------------
> Disciplina: MATEMÁTICA
                r_hat
escola_privada    1.0
Intercept         1.0
------------------------------
> Disciplina: REDAÇÃO
                r_hat
escola_privada    1.0
Intercept         1.0
------------------------------
```

Figure 2: Output - R-hat

Para todas as disciplinas modeladas, o estatístico **R-hat** ($\hat{R}$) apresentou valor igual a **1.0**. Isso indica que as cadeias de Markov convergiram perfeitamente para uma distribuição estacionária, ou seja, o algoritmo estabilizou e as amostras coletadas são confiáveis para a inferência estatística.

3.2 Priori vs. Posteriori (O Aprendizado do Modelo)

Definiu-se uma distribuição **Priori Vaga** para o coeficiente β ($\mathcal{N}(0, 100)$), indicando ceticismo inicial quanto à magnitude do efeito, mas permitindo que os dados conduzissem a estimativa.

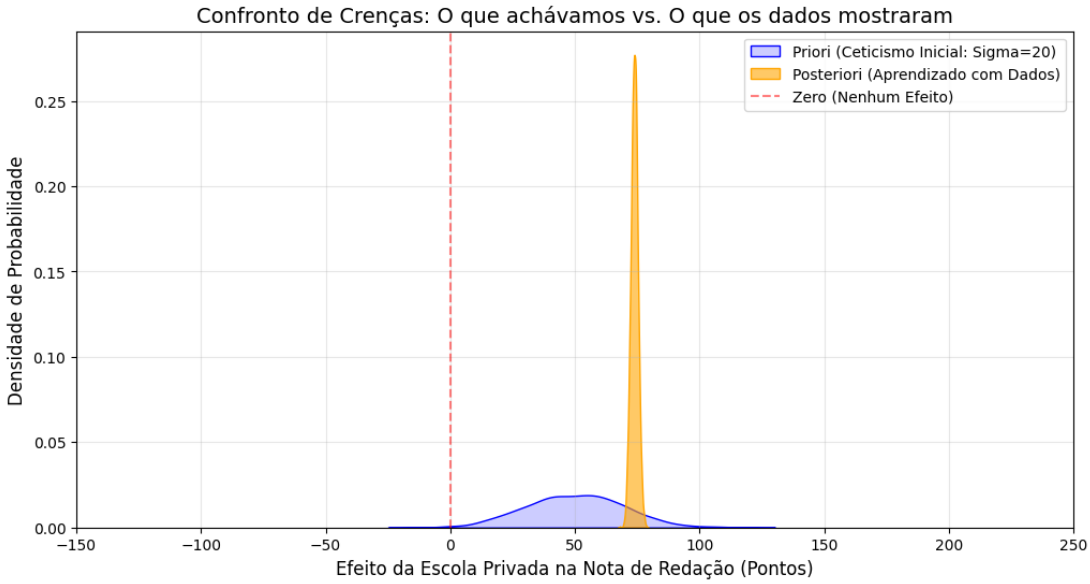

Figure 3: Output - Curvas Priori x Posteriori

- **Resultado Visual:** Ao comparar os gráficos de densidade, observou-se que a curva da **Posteriori** tornou-se extremamente estreita em comparação à curva larga da Priori.
- **Interpretação:** Isso demonstra que os dados do ENEM forneceram uma quantidade massiva de informação (verossimilhança), reduzindo a incerteza inicial de centenas de pontos para uma margem de erro mínima.

3.3 Intervalos de Credibilidade (HDI 95%)

As estimativas pontuais foram analisadas conjuntamente com os Intervalos de Alta Densidade (HDI) de 95%.

	Disciplina	Diferença Média	Mínimo (95%)	Máximo (95%)
0	Natureza	62.72	60.50	65.18
1	Humanas	64.05	61.77	66.69
2	Linguagens	55.38	53.25	57.62
3	Matemática	68.35	65.54	70.76
4	Redação	73.89	71.13	77.12

Figure 4: Output - Intervalos de Credibilidade

Em Matemática, por exemplo, o intervalo situou-se entre **65.54 e 70.76 pontos**. Como este intervalo não inclui o valor zero e encontra-se distante dele, rejeita-se a hipótese nula de inexistência de efeito. Há 95% de probabilidade de que o verdadeiro impacto da escola privada esteja contido nesta faixa.

4 Interpretar os resultados do modelo e responder o problema de pesquisa

Problema de Pesquisa: "O fato do aluno estudar em escola pública ou privada implica numa diferença de desempenho nas notas do ENEM?"

Resposta: Sim. A modelagem Bayesiana confirmou, com alto grau de certeza estatística, que estudar em escola privada implica em um aumento substancial no desempenho em todas as áreas do conhecimento avaliadas pelo exame.

Conclusão: O modelo permite concluir que a dependência administrativa da escola é um preditor determinante do sucesso no ENEM. A "escola privada" atua não apenas como um local de ensino, mas como um fator de estratificação que desloca a curva de desempenho dos alunos para um patamar estatisticamente distinto da realidade pública, evidenciando uma profunda desigualdade estrutural no sistema educacional brasileiro.