"Modelagem Bayesiana do Índice de Performance Estudantil: Uma Análise de Fatores Acadêmicos e Comportamentais""Modelagem Bayesiana do Índice de Performance Estudantil: Uma Análise de Fatores Acadêmicos e Comportamentais"

Othavio Henrique de Jesus Ayres Arthur Sales, Sofia Marinho

Introdução

Tendo como objetivo estimar a performance dos alunos com base nas variáveis dadas : Hours.Studied, Previous.Scores, Extracurricular.activities, Sleep.hours, Question.Practiced.
Utilizamos do pacote **brms** no R para escrever o seguinte modelo:
Performance.Index ~ Hours.Studied + Previous.Scores + Extracurricular.Activities + Sleep.Hours + Sample.Question.
Neste modelo o erro segue uma distribuição Normal(0, σ^2).
Temos como verossimilhança:

$$f(\mathbf{y}|X,\boldsymbol{\beta},\sigma_\epsilon^2) = \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right).$$

Visto que:

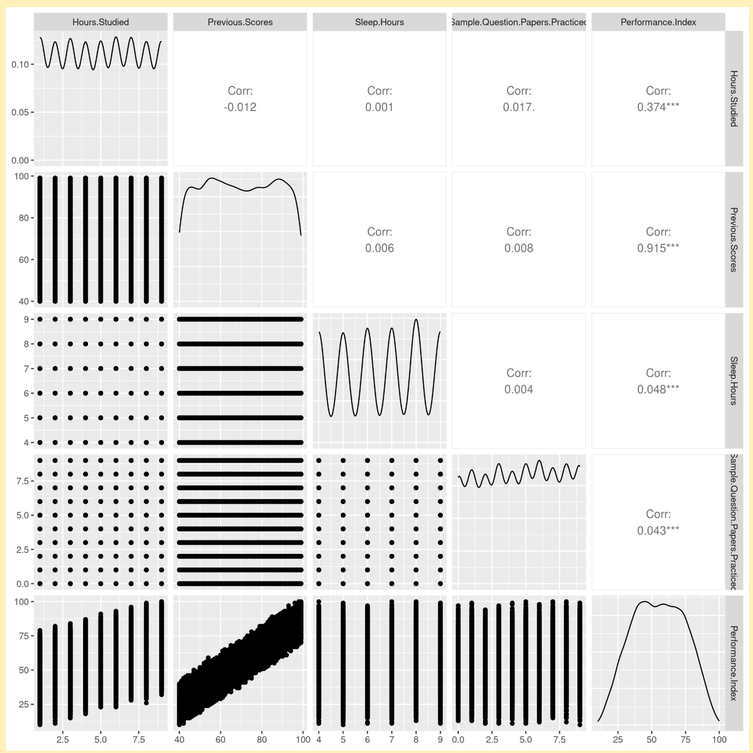
$$Y \sim \mathcal{N}(X\beta, \sigma^2 I)$$

Com as seguintes prioris ,para β e σ^2 :

$$\beta_j \sim \mathcal{N}(0, 10^6)$$

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha = 0.1,\ \beta = 0.001)$$

Abaixo temos a tabela de correlação dos dados utilizados



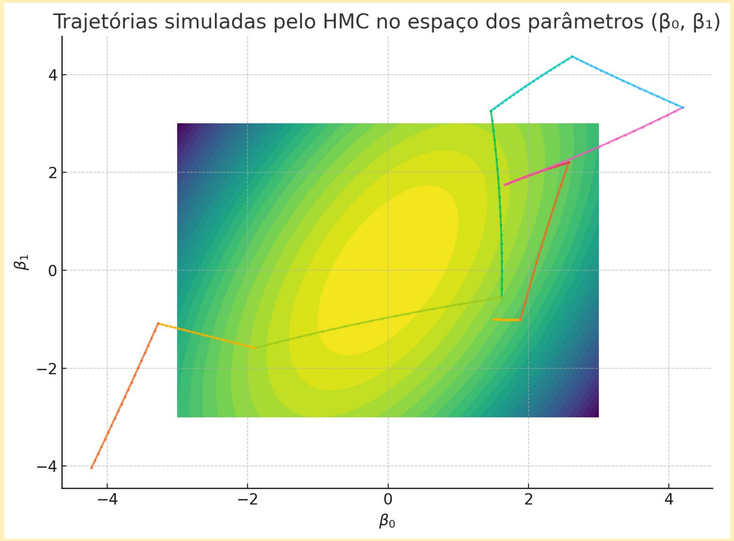
Metodologia

A inferência dos parâmetros foi realizada via MCMC (Markov Chain Monte Carlo), utilizando o pacote **brms** em R, com 2 cadeias, 2000 iterações por cadeia e 500 iterações de aquecimento (burnin) com um total de 10.000 amostras:
O **brms** utiliza do algoritmo *Hamiltonian Monte Carlo* (**HMC**) para gerar amotras dos parametros. O **HMC** simula o caminhar de uma particula pelo espaço. Este amostrador usa derivadas (gradientes) da log-posterior (Energia potência) para guiar o caminho de amostragem.
O **HMC** possui os seguintes parametros:

$$\begin{aligned}
\boldsymbol{\theta} &= [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma]^\top \quad \text{(posição)} \\
\mathbf{p} &= [p_0, p_1, p_2, p_3, p_4, p_5, p_\sigma]^\top \quad \text{(momento)} \\
E(\mathbf{p}) &= \frac{1}{2}\mathbf{p}^\top M^{-1}\mathbf{p} \quad \text{(energia cinética)} \\
U(\boldsymbol{\theta}) &= -\log p(\text{dados}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \quad \text{(energia potencial)} \\
H(\boldsymbol{\theta}, \mathbf{p}) &= U(\boldsymbol{\theta}) + E(\mathbf{p}) \quad \text{(Hamiltoniano)}
\end{aligned}$$

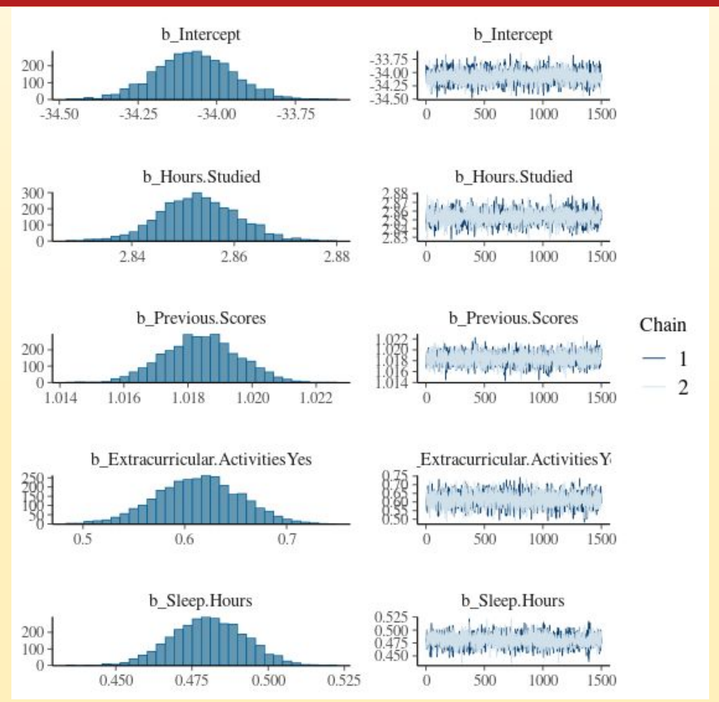
Nota: M é a matriz de massas (que no **brms** por padrão é matriz identidade)

O grafico abaixo simula a trajetoria (processo de atualização dos parametros via HMC) dentro do espaço parametrico. O grafico ilustra a densidade da verossimilhança como refêrencia.



As trajetórias dos paramêtros seguem as curvas de maior densidade (contornos mais claros), evitando movimentos aleatórios e aproveitando a informação dos gradientes.

Resultados e Conclusão



Distribuição e cadeias obtidas para cada paramêtro.

Parametros	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
Intercept	-34.07	0.12	-34.31	-33.83	1
Hours.Studied	2.85	0.01	2.84	2.87	1
Previous.Scores	1.02	0.00	1.02	1.02	1
Extracurricular	0.61	0.04	0.53	0.69	1
Sleep.Hours	0.48	0.01	0.46	0.50	1
Question.Practiced	0.19	0.01	0.18	0.21	1

O **Intercept** é aproximadamente -34.07, ou seja, quando todas as variáveis explicativas são zero, o índice de performance esperado é -34.07. Para **Hours.Studied**, o índice aumenta cerca de 2.85 pontos, indicando um forte efeito positivo do estudo no desempenho. O valor de **Previous.Scores** tem um coeficiente de aproximadamente 1.02, mostrando que cada ponto a mais nas notas anteriores contribui para um aumento similar no índice atual.
A participação em atividades extracurriculares (**Extracurricular**) adiciona cerca de 0.61 pontos no desempenho, evidenciando um benefício moderado. As **Sleep.Hours** aumentam o índice em torno de 0.48 por hora, indicando que dormir mais também ajuda. Por fim, praticar questões e provas (**Question.Practiced**) contribui com cerca de 0.19 pontos para o índice de performance, mostrando uma influência positiva, embora menor que as outras variáveis. Todos esses coeficientes possuem erro padrão pequeno e valores de Rhat igual a 1, indicando boa convergência do modelo.