ANÁLISE BAYESIANA DO DESEMPENHO ACADÊMICO DOS ALUNOS

Brenda Luiza Correa - RA216037; Vitória Linda da Silva Oliveira - RA212826;
Paula Liserre Calabrez - RA242782

INTRODUÇÃO

Com o objetivo de avaliar a influência de diversos fatores no desempenho dos alunos, ajustou-se um modelo de Regressão Linear Múltipla utilizando o Amostrador de Gibbs, um método de MCMC usado para estimar distribuições posteriores em modelos bayesianos. Ele funciona amostrando sequencialmente a partir das distribuições condicionais completas de cada parâmetro, visando garantir que as cadeias MCMC atinjam uma distribuição estacionária.

OBJETIVO

O objetivo deste projeto é analisar os fatores que influenciam o desempenho acadêmico dos alunos, utilizando técnicas de inferência Bayesiana para estimar os parâmetros do modelo. Buscamos: identificar a influência de variáveis preditoras no desempenho acadêmico; utilizar o Amostrador de Gibbs para estimar os parâmetros do modelo; avaliar a convergência das cadeias de Markov para garantir a precisão das estimativas.

BANCO DE DADOS

A base de dados utilizada contém 10.000 registros de alunos.

Variável Resposta: Desempenho (pontuação de desempenho do aluno);

Variáveis Preditoras: Horas Estudadas, Pontos Anteriores, Atividades Extras (1: Sim, 0: Não), Horas de Sono, Número de Questionários.

MÉTODOS

MODELO ESTATÍSTICO

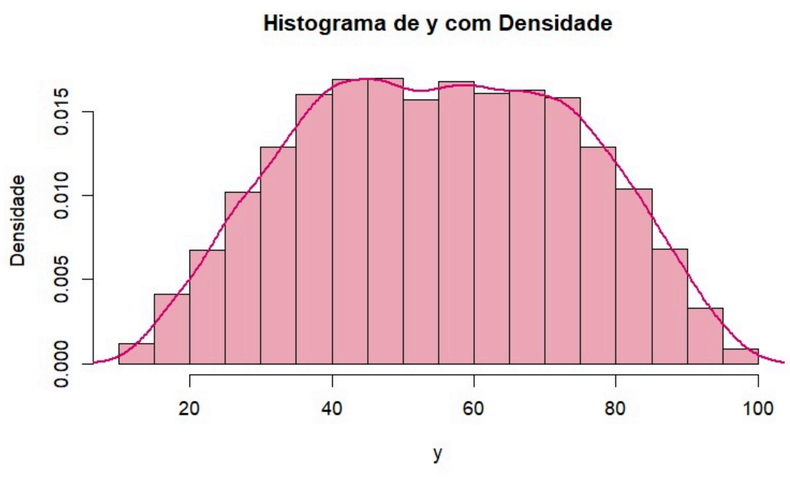
Foi utilizado um modelo de regressão linear Bayesiano para analisar a relação entre as variáveis preditoras e o desempenho acadêmico. O modelo inclui uma distribuição normal para os coeficientes de regressão e uma distribuição inversa-gama para a variância dos erros.

FÓRMULA DO MODELO: $Y = X\beta + \varepsilon$

X: Matriz dos valores das variáveis preditoras;

β: Matriz contendo os betas, de zero a cinco;

ε: erro aleatório.



Analisando o histograma da variável resposta nota-se uma distribuição simétrica, próxima de uma distribuição normal, o que condiz com estudo abordado, uma vez que na Regressão Linear Múltipla, Y segue uma distribuição normal.

INFERÊNCIA MCMC

A inferência foi conduzida utilizando o algoritmo MCMC, com duas cadeias iniciadas em pontos diferentes para avaliar a convergência. Foram realizadas 10.000 iterações, aplicando um burn-in de 5.000 iterações e subsequente amostragem a cada 5 iterações. Procedimentos:

- Inicialização dos parâmetros β e σ².
- Atualização de β e σ² iterativamente usando as distribuições condicionais completas a seguir:

$$\beta \mid \sigma^2, Y \,\alpha\, N\left(\left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} X^T X\right)^{-1}\left(\Sigma_\beta^{-1}\mu_\beta + \frac{1}{\sigma^2} X^T Y\right), \left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} X^T X\right)^{-1}\right)$$

$$\Sigma_\beta = diag\left(\sigma^2_{\beta_i}\right),\ i = 0,\ldots,5 \qquad \mu_\beta = \left[\mu_{\beta_0} \ \ldots \ \mu_{\beta_5}\right]^T$$

$$\sigma^2 \mid \beta, Y \,\alpha\, IG\left(\left(a + \frac{n}{2}\right), \left(b + \frac{1}{2}\left((Y - X\beta)^T (Y - X\beta)\right)\right)\right)$$

DIAGNÓSTICO DE CONVERGÊNCIA

Após realizar o burn-in e descartar as primeiras 5000 amostras e subamostrar as observações a cada 10 unidades, para reduzir as dependências entre as variáveis, foi verificada a independência das amostras e a convergência das cadeias, para isso foram realizadas as seguintes análises:

- **Análise de Autocorrelação:** os resultados abaixo de 0,02 indicaram baixa autocorrelação entre as amostras, sugerindo independência.
- **Estatística R^:** Todos os parâmetros apresentaram valores próximos de 1, sugerindo que as cadeias de Markov convergiram para a distribuição estacionária e indicando boa convergência.

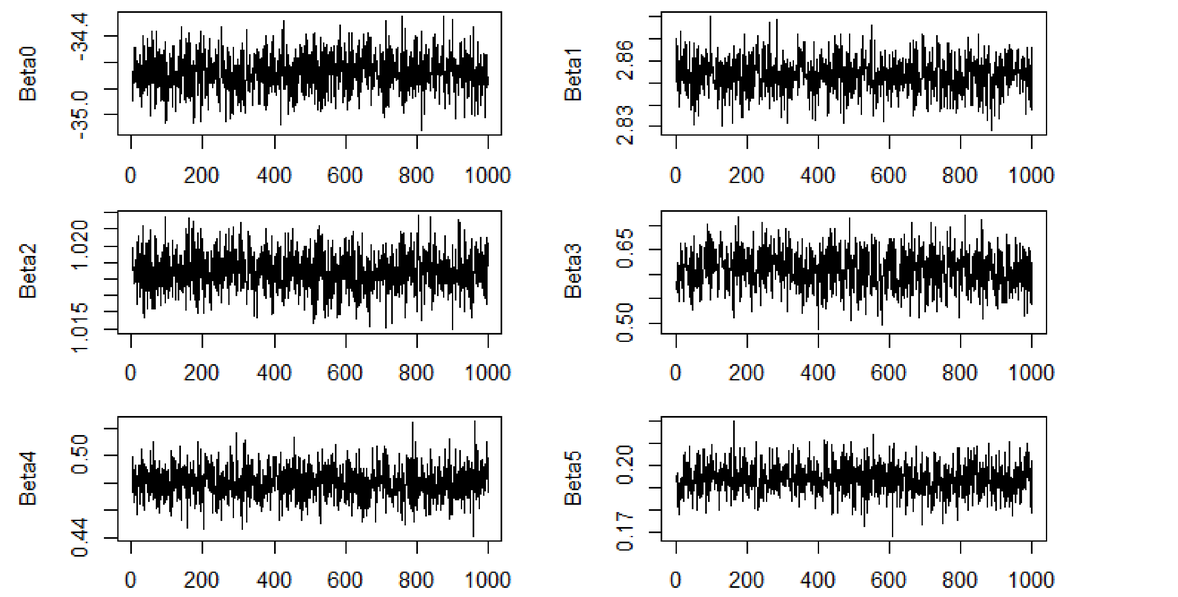
Caso a convergência não fosse inicialmente alcançada, seria aumentado o número de iterações ou ajustados os parâmetros iniciais para ajudar o algoritmo a convergir mais rapidamente.

RESULTADOS

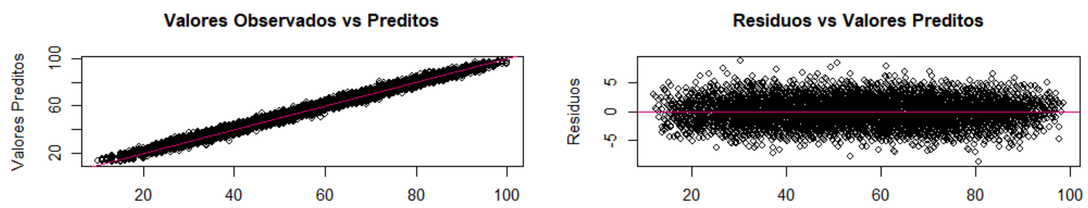
A inferência foi conduzida utilizando o algoritmo MCMC, com duas cadeias iniciadas em pontos diferentes para avaliar a convergência.

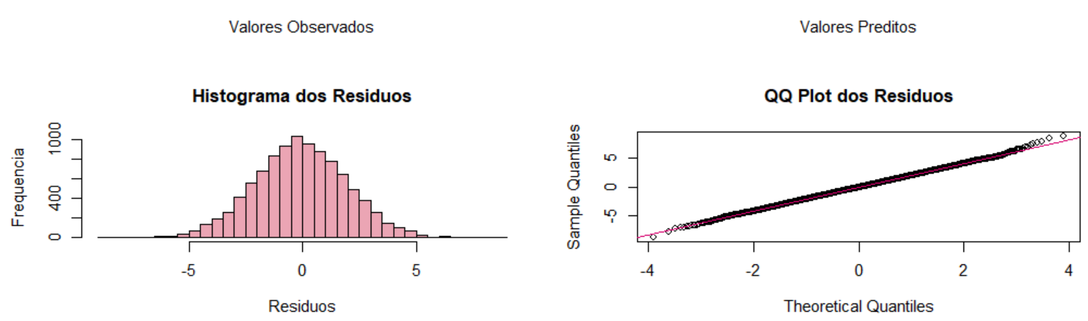
Parâmetros	Inferior	Superior	Estimativa
Intercepto	-34.97	-34.42	-34.68
Horas estudadas	2.84	2.87	2.85
Pontos Anteriores	1.016	1.021	1.02
Atividades Extras	0.53	0.69	0.61
Horas de Sono	0.45	0.50	0.48
Num. Questionários	0.18	0.21	0.19

Os intervalos de credibilidade são bastante estreitos para todos os coeficientes β, o que indica que as estimativas dos coeficientes são precisas. Além disso, não incluem o zero para nenhum parâmetro, assim, é razoável concluir que as variáveis preditoras possuem um efeito significativo na variável resposta.



Ao analisar o gráfico de cada beta, após o burn-in, nota-se que eles aparentam ter média e variância constantes, ou seja, são estacionários, indicando que as cadeias MCMC atingiram a distribuição estacionária.





Na Regressão Linear Múltipla é suposto que o erro segue uma distribuição normal com média zero e variância σ² e são homocedasticos. A normalidade é validada no modelo ao analisar o histograma, que apresenta uma distribuição simétrica e o qq-plot, no qual os pontos seguem bem a linha desenhada. No gráfico de Resíduos vs Valores Preditos os pontos são dispersos de forma aleatória, indicando homocedasticidade. Além disso, obteve-se um coeficiente de determinação múltipla de 0.988, ou seja, 98.8% da variabilidade dos dados observados é explicada pelo modelo utilizado. Portanto, o modelo ajustado é válido.