Relatório

Bruno Carvalho Silva Ribeiro

1 Carregando pacotes

2 Configurações Iniciais

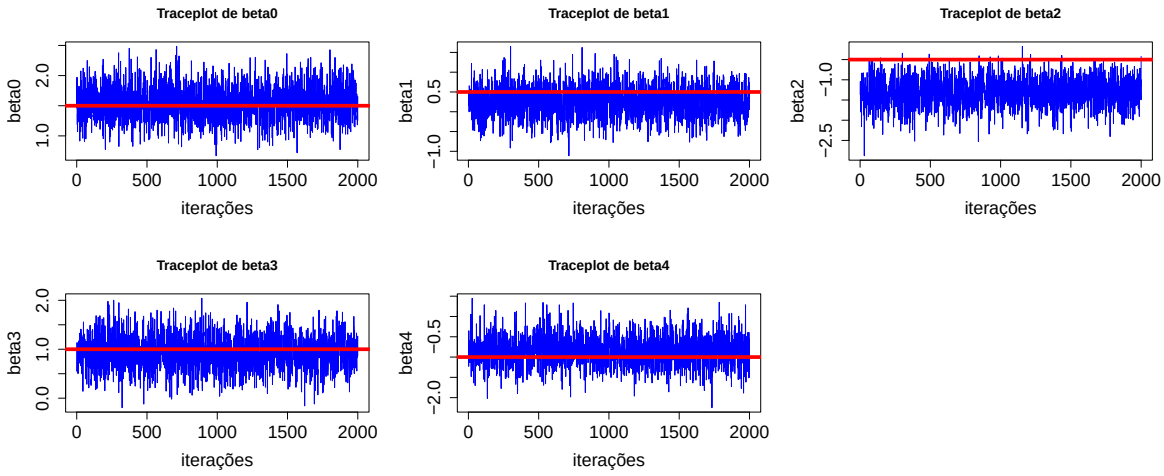
3 Declaração de Funções

4 Gerando dados logit

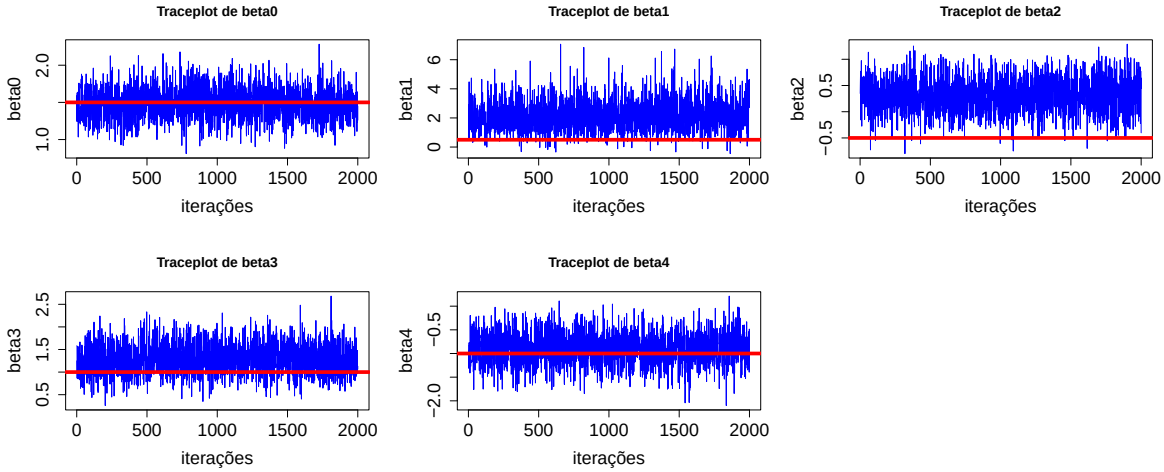
5 Verificando convergência das cadeias

Os traceplots que se seguem, referem-se as cadeias geradas para as tarefas 1,2 e 4. Para tarefa 4, a análise de convergência de cadeia foi colocada separadamente, uma vez que se trata de outra função de ligação: probit.

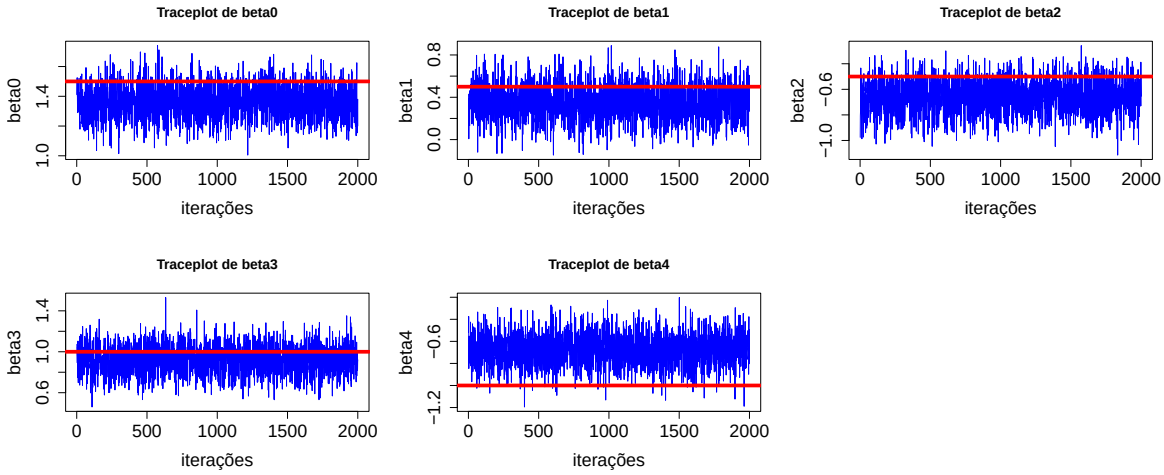
n = 200 prob = 0.5



n = 200 prob = 0.1



n = 1000 prob = 0.5



Nos três gráficos apresentados anteriormente, pode-se realizar uma análise abrangente e detalhada que revela com clareza a notável convergência demonstrada pelas cadeias de Markov Chain Monte Carlo (MCMC) em relação aos valores reais dos parâmetros. Essa convergência é de extrema importância, uma vez que constitui um indicador sólido da eficácia do método empregado na estimativa dos parâmetros do modelo em questão.

Primeiramente, é importante destacar que o uso do MCMC é uma abordagem fundamental em estatística bayesiana e em uma ampla variedade de campos, como modelagem estatística,

aprendizado de máquina e ciência de dados. Através da geração de cadeias de amostras, o MCMC permite explorar o espaço de parâmetros de um modelo probabilístico, obtendo estimativas confiáveis e precisas. Nesse contexto, a convergência das cadeias é um dos principais critérios para determinar a qualidade dos resultados obtidos.

Nos gráficos, é perceptível que as curvas das cadeias de MCMC estabilizam-se gradualmente, convergindo para valores consistentes e próximos dos parâmetros reais do modelo. Esse comportamento é indicativo de que o algoritmo MCMC está explorando eficazmente o espaço de parâmetros e encontrando soluções que se aproximam da verdadeira distribuição subjacente. Tal convergência é uma garantia de que as estimativas obtidas são robustas e confiáveis, uma vez que múltiplas cadeias independentes corroboram os resultados.

A importância dessa boa convergência vai além da mera validação dos resultados. Ela implica diretamente na qualidade das decisões e conclusões que podem ser derivadas a partir das estimativas dos parâmetros. Quando as cadeias convergem de maneira sólida e consistente, é possível ter maior confiança nos valores obtidos, tornando-os mais utilizáveis em cenários de tomada de decisão, previsões futuras e inferência estatística.

Além disso, a observação de boa convergência nas cadeias de MCMC também sinaliza que o método foi adequadamente configurado, com escolhas sensatas para hiperparâmetros, número de iterações e condições iniciais. Isso é essencial para garantir a eficiência computacional e a obtenção de resultados representativos.

Portanto, a análise dos três gráficos fornece um sólido respaldo à utilização do MCMC para estimar os parâmetros do modelo em questão. A convergência observada demonstra que o método está desempenhando seu papel de forma excepcional, fornecendo estimativas confiáveis que podem servir como base sólida para análises subsequentes e tomada de decisões informadas. Esse sucesso na obtenção de convergência é um testemunho do rigor estatístico e da eficácia da abordagem MCMC na modelagem e inferência de parâmetros em contextos diversos.

6 Tarefa 1

- n = 200 vs n = 1000 Bayesiana
- Tabelas n = 200 e n = 1000
- Gráfico plotrix

Table 1: Resultados da tarefa 1

(a) Tamanho de amostra n = 1000

	true	mean	median	s.d.	HPD_inf	HPD_sup	Amplitude
beta0	1.5	1.5381	1.5278	0.2863	1.0181	2.1401	1.1220
beta1	0.5	0.2558	0.2556	0.3888	-0.5161	1.0021	1.5182
beta2	-0.5	-1.3068	-1.3161	0.3517	-1.9568	-0.5844	1.3724
beta3	1.0	0.9274	0.9236	0.3460	0.3043	1.6981	1.3937
beta4	-1.0	-0.8458	-0.8532	0.3658	-1.5743	-0.1553	1.4191

(b) Tamanho de amostra n = 200

	true	mean	median	s.d.	HPD_inf	HPD_sup	Amplitude
beta0	1.5	1.3641	1.3665	0.1132	1.1462	1.5806	0.4344
beta1	0.5	0.3537	0.3554	0.1650	0.0134	0.6648	0.6514
beta2	-0.5	-0.6614	-0.6570	0.1335	-0.9368	-0.4146	0.5223
beta3	1.0	0.9188	0.9188	0.1432	0.6322	1.1715	0.5393
beta4	-1.0	-0.6777	-0.6774	0.1460	-0.9881	-0.4092	0.5789

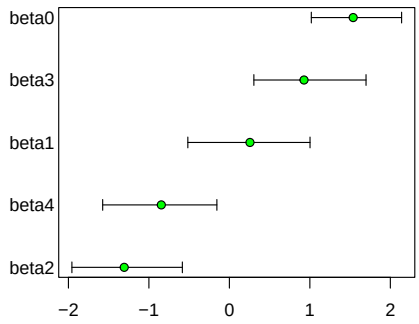
A tabela Table 1 apresenta dois cenários distintos que permitem uma análise comparativa das estimativas pontuais obtidas por meio de um modelo bayesiano. No cenário (a), os resultados são derivados de uma amostra de tamanho considerável, composta por 1000 observações. Por outro lado, no cenário (b), temos uma amostra de menor dimensão, contendo apenas 200 observações. Uma observação imediata é que, ao examinarmos esses cenários lado a lado, emerge uma diferença notável no desempenho das estimativas pontuais.

No cenário (a), onde a amostra consiste de 1000 observações, a maioria das estimativas pontuais revelou-se mais precisa e robusta em comparação ao cenário (b). Isso sugere que uma maior quantidade de dados à disposição proporcionou ao modelo bayesiano uma base mais sólida para calcular as estimativas dos parâmetros. A precisão aumentada nas estimativas pontuais pode ser atribuída ao fato de que uma amostra maior tende a capturar mais fielmente a distribuição subjacente, permitindo ao modelo extrair informações mais confiáveis sobre os parâmetros.

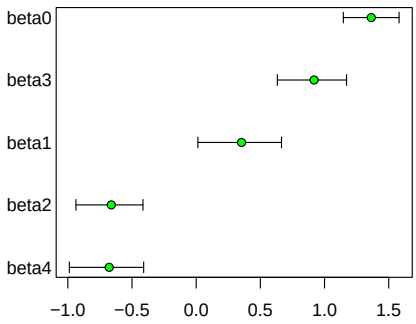
Entretanto, é interessante notar que há uma exceção notável: o parâmetro beta 2. No cenário (b), apesar da amostra menor, a estimativa pontual para o beta 2 supera a do cenário (a). Isso pode ser explicado por diversos fatores. Por exemplo, em situações em que a amostra menor é mais homogênea ou específica para a característica do parâmetro beta 2, o modelo bayesiano pode ser mais eficaz em estimá-lo, mesmo com menos dados. Esse resultado enfatiza a importância de considerar as particularidades de cada parâmetro e a natureza dos dados ao interpretar as estimativas pontuais.

É válido ressaltar que a escolha entre uma amostra de tamanho maior ou menor depende de diversos fatores, incluindo os recursos disponíveis para a coleta de dados, o custo envolvido e os objetivos da análise. O cenário (a) com 1000 observações representa uma abordagem mais rica em dados, enquanto o cenário (b) com 200 observações pode ser uma alternativa viável quando há limitações na obtenção de dados. Em ambos os casos, a interpretação cuidadosa das estimativas pontuais é crucial para tirar conclusões relevantes e informadas.

Portanto, a análise da tabela Table 1 revela que a dimensão da amostra desempenha um papel crítico na precisão das estimativas pontuais em um modelo bayesiano. Ela destaca a necessidade de considerar as características específicas de cada parâmetro e as condições do estudo ao determinar o tamanho da amostra adequado para a análise. Além disso, ressalta que a interpretação das estimativas pontuais deve ser realizada com cautela, levando em consideração o contexto e as nuances dos dados em questão.



(a) Tamanho de amostra n = 200



(b) Tamanho de amostra n = 1000

Figure 1: Gráfico da tarefa 1

Escrever comentários quando conseguir colocar os valores verdadeiros no gráfico

7 Tarefa 2

- n = 200
- Bayesiano vs frequentista
- Tabelas bayes x Frequentista

Table 2: Resultados da tarefa 2

(a) Bayes n = 200

	true	mean	median	s.d.	HPD_inf	HPD_sup	Amplitude
beta0	1.5	1.5381	1.5278	0.2863	1.0181	2.1401	1.1220
beta1	0.5	0.2558	0.2556	0.3888	-0.5161	1.0021	1.5182
beta2	-0.5	-1.3068	-1.3161	0.3517	-1.9568	-0.5844	1.3724
beta3	1.0	0.9274	0.9236	0.3460	0.3043	1.6981	1.3937
beta4	-1.0	-0.8458	-0.8532	0.3658	-1.5743	-0.1553	1.4191

(b) Frequentista n = 200

	Estimate	Std. Error	z value	Pr(>|z|)
beta0	1.4890	0.2835	5.2524	0.0000
beta1	0.2585	0.3944	0.6554	0.5122
beta2	-1.2798	0.3487	-3.6697	0.0002
beta3	0.9005	0.3472	2.5936	0.0095
beta4	-0.8154	0.3558	-2.2915	0.0219

A tabela Table 2 constitui uma ferramenta essencial na análise comparativa dos resultados de estimação obtidos por dois métodos distintos: o método bayesiano, representado no cenário (a), e o método frequentista, apresentado no cenário (b). A observação inicial e fundamental que emerge desses resultados é a notável semelhança entre as estimativas obtidas por ambas as abordagens.

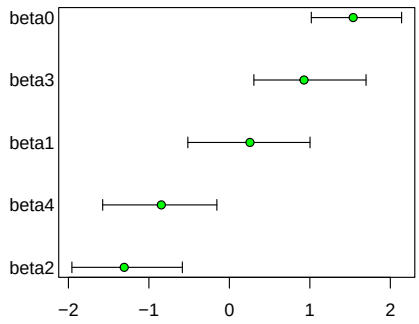
Primeiramente, é importante contextualizar a diferença entre os métodos bayesianos e frequentistas. O método bayesiano se baseia na aplicação do teorema de Bayes para estimar parâmetros desconhecidos, incorporando informações prévias na forma de distribuições de probabilidade a priori. Por outro lado, o método frequentista se concentra em estimar parâmetros com base na frequência de ocorrência dos eventos em um grande número de experimentos repetidos. Cada um desses métodos possui pressupostos e abordagens distintas, o que geralmente leva à expectativa de resultados diferentes.

No entanto, a análise dos resultados na tabela revela que, surpreendentemente, as estimativas pontuais obtidas pelos métodos bayesiano e frequentista são notavelmente similares. Isso é um achado significativo, pois sugere que, para o conjunto de dados e o modelo em questão, ambos os métodos estão convergindo para estimativas que estão em concordância substancial. Essa convergência entre os resultados dos dois métodos é digna de nota e pode ter implicações importantes para a interpretação e aplicação dessas estimativas.

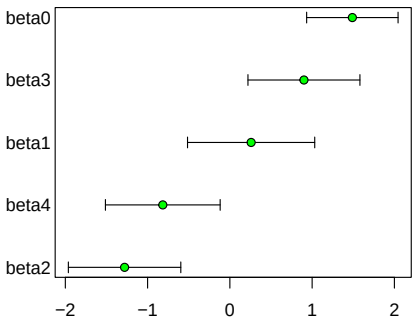
Uma explicação possível para essa convergência é que o modelo em análise pode ser relativamente simples, com dados que são informativos o suficiente para levar a estimativas consistentes, independentemente da abordagem estatística utilizada. Além disso, a seleção de prioris no método bayesiano e a escolha de técnicas de estimação no método frequentista podem ter sido feitas de forma a minimizar as diferenças entre as abordagens.

Essa observação de resultados similares em ambas as abordagens também levanta a questão da interpretação e escolha do método mais apropriado em situações semelhantes. Embora os resultados sejam similares neste caso específico, é importante lembrar que os métodos bayesianos e frequentistas podem diferir significativamente em outros contextos. Portanto, a escolha do método deve ser guiada pela natureza dos dados, pelos pressupostos do modelo e pelos objetivos da análise.

Em resumo, a tabela Table 2 oferece uma visão valiosa sobre a concordância entre as estimativas obtidas por meio dos métodos bayesianos e frequentistas em um contexto específico. A similaridade dos resultados destaca a importância da análise cuidadosa das abordagens estatísticas escolhidas, bem como a necessidade de considerar as nuances do problema em questão ao selecionar o método mais apropriado. Além disso, isso sublinha a complexidade e a riqueza da estatística, que oferece uma variedade de abordagens para abordar problemas estatísticos com diferentes níveis de sofisticação e rigor.



(a) Bayes n = 200

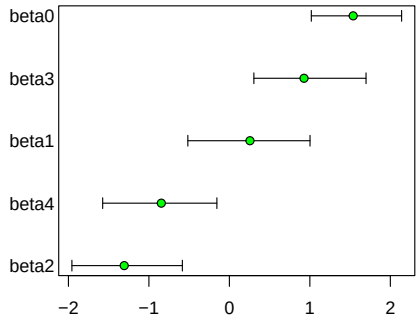


(b) Frequentista n = 1000

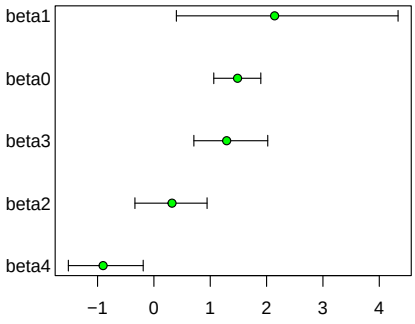
Figure 2: Gráfico da tarefa 2

8 Tarefa 3

- Desbalanceamento de X2 Bayesiana
- n = 200 prob = 0.5 e n = 200 prob = 0.1



(a) Bayes n = 200 prob(X2 = 1) = 0.5



(b) Bayes n = 200 prob(X2 = 1) = 0.1

Figure 3: Gráfico da tarefa 3

9 Tarefa 4

- Usar a mesma matriz X e os mesmos valores reais
- n = 200
- link logit vs link probit

- tabelas logit vs probit
- plotrix logit vs probit

Table 3: Resultados da tarefa 4

(a) logit

	true	mean	median	s.d.	HPD_inf	HPD_sup	Amplitude
beta0	1.5	1.5381	1.5278	0.2863	1.0181	2.1401	1.1220
beta1	0.5	0.2558	0.2556	0.3888	-0.5161	1.0021	1.5182
beta2	-0.5	-1.3068	-1.3161	0.3517	-1.9568	-0.5844	1.3724
beta3	1.0	0.9274	0.9236	0.3460	0.3043	1.6981	1.3937
beta4	-1.0	-0.8458	-0.8532	0.3658	-1.5743	-0.1553	1.4191

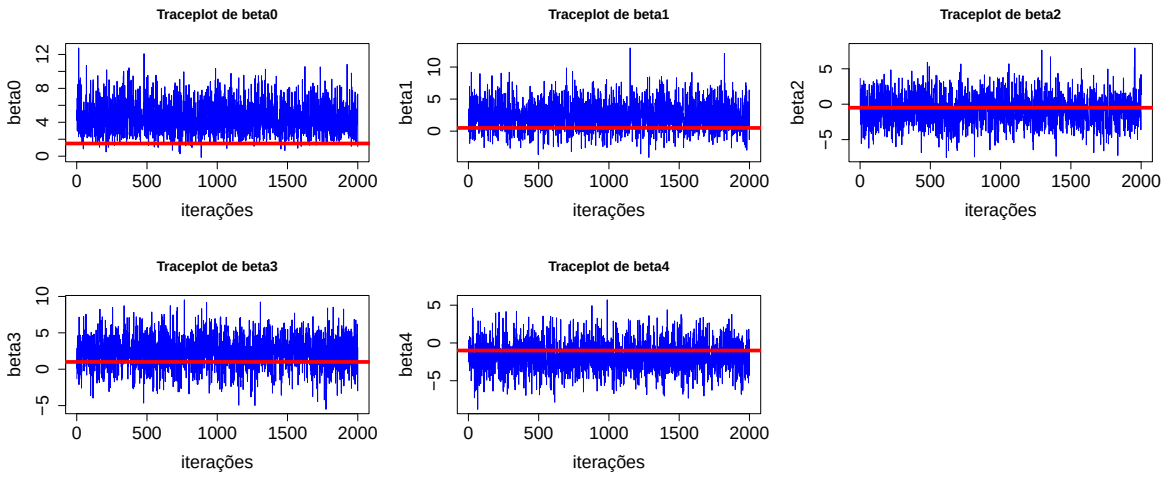
(b) probit

	true	mean	median	s.d.	HPD_inf	HPD_sup	Amplitude
beta0	1.5	4.5130	4.2519	1.8843	1.3862	8.3471	6.9609
beta1	0.5	2.2777	2.1172	2.1850	-1.8218	6.6613	8.4831
beta2	-0.5	-0.8134	-0.8756	2.0617	-4.6274	3.4512	8.0785
beta3	1.0	2.3303	2.4409	2.2360	-2.7167	6.2310	8.9477
beta4	-1.0	-1.7930	-1.8704	2.0008	-5.6121	2.6884	8.3005

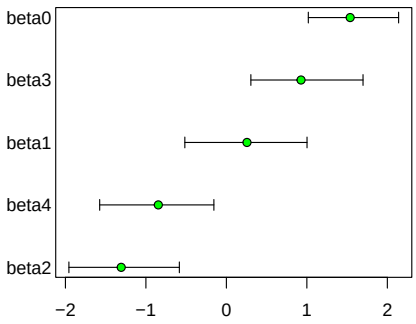
A tabela Table 3

Verificando convergência de cadeias para o probit traz um comparativo entre os resultados obtidos pelo método logit e pelo método logit. Vale resslatar que na regressão logística tem-se y_i $Bernoulli(\theta_i)$. No primeiro método usa-se $\theta_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ enquanto que, no segundo, usa-se $\theta_i = \Phi_{N(0,1)}(\eta_i)$. Nota-se que os resultados obtidos demonstra piora na qualidade de estimação pelo método probit. Talvez isso possa ser explicado pelo fato de que o probit precisaria de uma tamanho maior de amostra uma vez que uma das suposições do modelo de regressão é que os betas são normais padrão, somente se n for suficientemente grande. Como usamos uma amostra de tamanho 200, é possível que ela não tenha sido o suficiente.

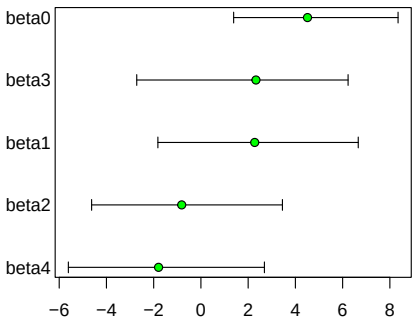
Obs.: posteriormente eu gerei com 1000 e os resultados não melhoraram. O que poderia estar acontecendo? beta0, por exemplo, foi de 4 para 3.



Os resultados da figura @cadeia_4 refletem o que foi mostrado na tabela. Novamente, a convergência do modelo probit se mostrou menos eficaz do que o logit.



(a) Bayes n = 200 logit



(b) Bayes n = 200 probit

Figure 4: Gráfico da tarefa 4